Table des matières

Lexique 1

Résumé automatique

Traitement de corpus 2

Modélisation probabiliste de l'interface syntaxe sémantique à l'aide de grammaires hors contexte probabilistes
Expériences avec FrameNet

Olivier Michalon
LIF, AMU, CNRS, 163 avenue de Luminy, 13288 Marseille Cedex 9
olivier.michalon@lif.univ-mrs.fr

Résumé. Cet article présente une méthode générative de prédiction de la structure sémantique en cadres d'une phrase à partir de sa structure syntaxique et décrit les grammaires utilisées ainsi que leurs performances. Ce système permet de prédire, pour un mot dans le contexte syntaxique d'une phrase, le cadre le plus probable. Le système génératif permet d'attribuer à ce mot un cadre et à l'ensemble de chemins des rôles sémantiques. Bien que les résultats ne soient pas encore satisfaisants, cet analyseur permet de regrouper les tâches d'analyse sémantique (sélection du cadre, sélection des actants, attribution des rôles), contrairement aux travaux précédemment publiés. De plus, il offre une nouvelle approche de l'analyse sémantique en cadres, dans la mesure où elle repose plus sur la structure syntaxique que sur les mots de la phrase.

Abstract. This paper presents a generative method for predicting the frame semantic structure of a sentence from its syntactic structure and describes the grammars used with their performances. This system allows to predict, for a word in the syntactic context of a sentence, the most probable frame. The generative system allows to give a frame to a word and semantic roles to a set of pathes. Although results are not yet satisfying, this parser allows to group semantic parsing tasks (frame selection, role fillers selection, role assignment) unlike previously published works. In addition, it offers a new approach to parse semantic frames insofar as it is based more on syntactic structure rather than words of the sentence.

Mots-clés : Analyse sémantique automatique, interface syntaxe sémantique, FrameNet.

Keywords: Automatic Semantic Parsing, syntax semantic parsing, FrameNet.

1 Introduction

Plusieurs théories de représentation de la structure sémantique coexistent actuellement, et parmi celles-ci, plusieurs projets ont vu le jour, parmi lesquels on trouve WordNet (Miller & Fellbaum, 1998), PropBank (Palmer *et al.*, 2005) ou encore FrameNet (Fillmore & Baker, 2001) . Alors que WordNet s'attache à représenter la hiérarchie entre différents noms et verbes (une voiture est une sous-classe de véhicule à moteur), PropBank associe une étiquette à certains actants syntaxiques des verbes. Le projet FrameNet quant à lui se base sur une analyse des situations nommées cadres sémantiques. Ces cadres sémantiques représentent chacun une action ou un concept, en incluant les éléments qui y jouent un rôle. Par exemple pour l'action *manger*, le cadre sémantique identifie la *personne qui mange*, la *nourriture*, ainsi que les *outils utilisés pour manger*. Ces différents rôles ne sont pas sans rapport avec les arguments de ProbBank mais l'inventaire est beaucoup plus riche et FrameNet recense des relations entre les rôles de différents cadres.
Chacun de ces projets a permis de créer des données annotées qui permettent d'estimer les paramètres d'outils d'analyse automatique basés sur les données.

L'approche que nous présentons dans cet article consiste à utiliser les données FrameNet de l'anglais (le corpus Français est en cours de construction) pour comprendre à quel point la structure syntaxique d'une phrase peut jouer un rôle dans la prédiction de sa structure sémantique. En effet la réalisation syntaxique d'actants sémantiques présente des régularités et ce sont précisément ces régularités que nous cherchons à modéliser ici. Pour l'action *manger*, la personne qui mange sera la plupart du temps le sujet du verbe, la nourriture sera l'objet et les outils seront des compléments. Pour automatiser cette tâche, nous utilisons les données annotées du corpus FrameNet pour estimer les probabilités d'une grammaire probabiliste hors contexte. Cette grammaire permet d'analyser des réalisations syntaxiques en leur associant la représentation sémantique la plus probable. Contrairement aux travaux réalisés par Modi *et al.* (2012), ou encore Das *et al.* (2010), dans lesquels la tâche d'analyse sémantique est scindée en trois voire en quatre parties distinctes (identification des mots dé-

clencheurs de cadre, sélection du cadre, identifications des mots acteurs du cadre, attribution des rôles à ces mots), notre grammaire permet de réaliser ces opérations simultanément, ce qui permet d'envisager l'analyse sémantique comme une tâche globale.
Nous ne nous attacherons dans cette étude qu'à trois parties de discours auxquelles on peut attribuer des cadres sémantiques : les adjectifs (*JJ*), les noms (*NN*), et les verbes (*VS*). Nous avons fait ce choix car ces trois parties de discours sont celles présentant la plus grande variété sémantique et syntaxique.

Après avoir présenté la théorie et les données de FrameNet, nous détaillerons le modèle que nous avons utilisé pour réaliser l'analyse sémantique. Nous continuerons avec les résultats de ce modèle et conclurons en proposant quelques pistes pour la suite.

2 FrameNet

2.1 Présentation du projet FrameNet

Le projet FrameNet regroupe à la fois une théorie de modélisation de la sémantique des phrases et un ensemble de données annotées manuellement. L'unité structurelle utilisée pour représenter la structure sémantique d'une phrase est appelée *Semantic Frame* (cadre sémantique en français), et elle est fondée sur les travaux de Charles J. Fillmore (Fillmore, 1976, 1977, 1982, 1985; Fillmore & Baker, 2010).

Par exemple, le verbe *continuer* (*continue*) peut évoquer deux situations différentes : *Une activité qui continue* (modélisée par le cadre *Process_continue*) ou *la poursuite d'une activité* (*Activity_ongoing*). Ces situations, bien que proches, ne représentent pas la même chose : la première concerne une tâche qui se poursuit, alors que la seconde concerne des participants qui continuent une activité.
Le cadre sémantique de *Process_continue* met en jeu essentiellement un *événement* (*Event*), mais peut aussi contenir des *circonstances* (*Circumstances*), une *manière de continuer* (*Manner*), ou encore *l'événement suivant* (*Next_subevent*), comme dans l'exemple suivant :

$$\text{The meeting}_{Event}\ \text{continued}_{PROCESS\_CONTINUE}\ \text{with a discussion}_{Next\_subevent}.$$
(La réunion s'est poursuivie par un débat.)

Le cadre sémantique de *Activity_ongoing* nécessite absolument de définir *l'activité* (*Activity*) et un *agent* (*Agent*) effectuant cette activité. Bien entendu ce cadre sémantique peut aussi contenir des indicateurs de *circonstances* (*Circumstances*), d'une *manière de continuer* (*Manner*), ou encore de *l'événement suivant* (*Next_subevent*), comme dans l'exemple suivant :

$$\text{We}_{Agent}\ \text{continued}_{ACTIVITY\_ONGOING}\ \text{the meeting}_{Event}\ \text{with a discussion}_{Next\_subevent}.$$
(Nous avons poursuivi la réunion par un débat.)

Les deux situations évoquées précédemment (*Process_continue* et *Activity_ongoing*) sont ce que nous appellerons des *cadres* (*frame* en anglais) et les intervenants (*Event*, *Agent*,...) sont appelés *rôles* (*frame elements*). Les mots qui pourront évoquer ces cadres (comme *poursuivre*) sont appelés *ancres* (*lexical units*) du cadre. Les rôles sont spécifiques à un cadre. Le mot ou l'expression qui joue un rôle dans un cadre est appelé *acteur*. Le nombre de rôles peut varier selon les cadres. Certains rôles peuvent ne pas être réalisés pour une occurrence de cadre donnée, le jeu de rôles instanciés peutdonc varier d'une occurrence de cadre à une autre. Lorsqu'une ancre est associée à un cadre, on dit que l'ancre déclenche le cadre, qui est alors instancié.
Notons également qu'un mot ne peut déclencher qu'un seul cadre alors qu'un mot peut occuper un rôle dans plusieurs cadres. De plus, un mot peut être à la fois ancre et acteur.

Le projet FrameNet a permis de créer des ressources lexicales et un corpus annoté pour l'anglais. En ce qui concerne la version 1.5 du projet, les données sont composées :

- d'un lexique associant une ancre à des cadres ;
- d'un inventaire de cadres et de leurs rôles ;
- d'un corpus annoté (*corpus FullText*) en cadres et étiqueté en parties de discours ;
- d'un ensemble de données d'exemple annotées pour des cadres spécifiques.

Le *lexique* comprend un peu moins de 12000 entrées. Chacune des entrées associe un cadre à une unité lexicale (un lemme avec une partie de discours). Une unité lexicale peut être associée à plusieurs cadres, comme les exemples précédents nous l'ont montré. On trouve en fin de compte 894 cadres différents et 9394 ancres différentes. Le tableau 1 regroupe trois informations importantes pour chaque partie de discours que nous allons traiter : le nombre d'ancres, le nombre de cadres

	Nombre d'entrées	Nombre de cadres	Ambiguïté moyenne
Adjectifs	1850	307	1.15
Verbes	3060	604	1.5
Noms	4166	605	1.14

TABLE 1 – Variété des ancres selon leur partie de discours dans la version 1.5 de FrameNet

pouvant être déclenchés, et l'ambiguïté moyenne, qui est le nombre moyen de cadres pouvant être déclenchés par ancre.
L'*inventaire* de cadres sémantique recense chaque cadre, liste et définit les rôles lui appartenant ainsi que les liens existants avec d'autres cadres.
Le *corpus* est composé d'un ensemble de phrases dont la structure sémantique est annotée manuellement. On y trouve aussi une annotation en parties de discours des phrases. Ce corpus est composé de 4037 phrases. On y trouve 23871 occurrences de cadres (706 cadres différents), et 46929 occurrences d'ancres (3345 ancres différentes). Le fait que le nombre d'occurrences d'ancres soit bien supérieur au nombre d'instances de cadres provient du fait que les ancres peuvent parfois véhiculer un sens qui n'est pas modélisé dans la théorie FrameNet.
Le *corpus d'exemples* est un corpus annoté à la main pour des cadres spécifiques. Chaque phrase de ce corpus est destinée à illustrer l'utilisation d'un seul cadre, par conséquent seul un cadre sera annoté par phrase. Ce corpus dispose lui aussi d'annotations en parties de discours.

2.2 Prétraitements effectués sur le corpus

Pour nos travaux nous utilisons le corpus de phrases annotées complètement en lui ajoutant une analyse syntaxique et les lemmes, puis nous le scinderons en un corpus d'entraînement et un corpus de test.

Du corpus FrameNet, nous utiliserons les mots, parties de discours, ainsi que les cadres sémantiques. Pour ce qui est des parties de discours, nous les normalisons pour qu'elles correspondent toutes à celles du *PennTreeBank*.
Pour savoir si un mot peut être une ancre ou non, nous avons besoin de connaître sa forme lemmatisée, nous ajoutons donc les lemmes des mots aux données extraites du corpus FrameNet. Notre approche s'appuyant sur la syntaxe, nous effectuons également une analyse syntaxique en dépendances de notre corpus. La lemmatisation et l'analyse en dépendances ont été réalisées avec le logiciel *Mate-tools* (Bohnet *et al.*, 2013; Bohnet, 2010), dont la performance en analyse syntaxique avec étiquettes de dépendances est à $90,33\%$ sur le corpus *Conll Shared Task 2009*. Nous devrons donc composer avec les erreurs syntaxiques générées par l'analyseur. Notons également que les lemmes de ce corpus sont automatiquement prédits.

Dans les annotations FrameNet, un acteur est souvent associé à un segment de phrase, généralement un syntagme. Par souci de simplicité, les acteurs composés de plusieurs mots ne seront plus que représentés par leur tête syntaxique. Bien que cette réduction puisse être vue comme une perte d'information, dans la plupart des cas la tête syntaxique correspond bien à la tête sémantique de l'expression.

Le corpus est ensuite séparé en deux : (*corpus d'entraînement* : 3055 phrases et *corpus de test* : 982 phrases) selon la séparation historique (Das & Smith, 2011) établie pour la tâche SemEval. On extrait ensuite du corpus d'entraînement l'ensemble des chemins syntaxiques (*corpus des chemins*) permettant d'aller d'une ancre aux acteurs correspondants.

Un *chemin syntaxique* correspond au chemin emprunté pour aller d'un noeud à un autre dans l'arbre de dépendances syntaxiques de la phrase. Il est possible d'aller dans le sens d'une dépendance syntaxique ou d'aller à contre courant (dépendant → gouverneur). Chaque branche de l'arbre peut être empruntée 0 ou 1 fois, ce qui garantit l'unicité du chemin entre deux noeuds donnés. Si une dépendance est empruntée à contre courant, on l'indique en faisant précéder l'étiquette de dépendance syntaxique du signe « - ». Un chemin syntaxique est donc défini comme une séquence *[(+/-), lex, lemme, PoS, fct]**, où le symbole de signe indique le sens de la dépendance, *lex* (et *lemme*) le mot (et son lemme) traversé par ce chemin, *PoS* sa partie de discours et *fct* sa fonction syntaxique. Pour nos expériences nous souhaitons faire des statistiques sur les chemins syntaxiques. Le corpus étant de taille réduite et afin d'obtenir des statistiques fiables, nous ne pouvons nous permettre d'avoir des chemins d'une grande variété mais avec une fréquence faible. Pour cela nous allons les simplifier afin de faire des regroupements plus conséquents. Nous réduirons donc les chemins syntaxiques à des suites d'étiquettes de dépendances syntaxiques orientées.

Par exemple, dans la phrase suivante :

Maurice$_{Cook}$ bakes$_{APPLY_HEAT}$ an apple pie$_{Food}$ in the oven$_{Heating_instrument}$.
(Maurice cuit une tarte aux pommes dans le four.)

Le mot *bakes* est une ancre pour le cadre *Apply_heat*, dont les rôles instanciés ici sont *Cook*, *Food* et *Heating_instrument*, dont les acteurs sont respectivement *Maurice*, *apple pie* (dont la tête syntaxique est *pie*) et *oven*.
L'arbre syntaxique de la phrase est représenté en figure 1, et les chemins associés au cadre *Apply_heat* sont décrits dans la table 2.

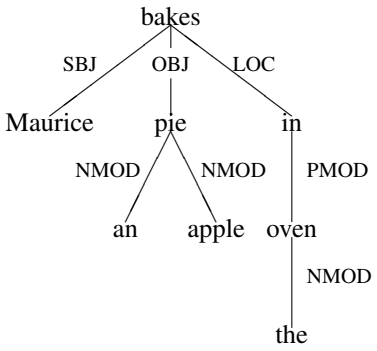


FIGURE 1 – Arbre syntaxique de la phrase *Maurice bakes an apple pie in the oven*

Rôle	**Début du chemin**	→	**Fin du chemin (Acteur)**	**Représentation du chemin**
Cook	bakes	→	Maurice	(+SBJ)
Food	bakes	→	pie	(+OBJ)
Heating_instrument	bakes	→	oven	(+LOC,+PMOD)

TABLE 2 – Chemins syntaxiques du cadre *Apply_heat* pour la phrase *Maurice bakes an apple pie in the oven.*

La table 3 présente les cinq chemins les plus fréquents pour les ancres des parties de discours que nous traitons (*Noms*, *Adjectifs* et *Verbes*). On remarque tout d'abord que certains chemins syntaxiques sont effectivement prédominants par rapport aux autres. Pour les noms et les adjectifs, on remarque également que le chemin vide (représenté *()*) est très représenté : les ancres ayant ces parties de discours sont donc souvent des acteurs des cadres qu'ils déclenchent. On remarque aussi que l'épithète du nom (*NMOD*) lorsque l'ancre est un nom joue souvent un rôle. De même lorsque l'ancre est un adjectif, le nom que qualifie l'adjectif est souvent acteur (*-NMOD*). En ce qui concerne les verbes, les traditionnels sujet (*SBJ*) et objet (*OBJ*) sont les éléments qui jouent le plus souvent un rôle.

Noms		Adjectifs		Verbes	
(NMOD)	5702/13381	(-NMOD)	1418/3366	(OBJ)	2696/12679
()	5213/13381	()	921/3366	(SBJ)	1802/12679
(LOC)	170/13381	(AMOD)	206/3366	(-VC,SBJ)	968/12679
(-NMOD,NMOD)	161/13381	(-NMOD,NMOD)	122/3366	(ADV)	858/12679
(-OBJ,SBJ)	113/13381	(-PRD,SBJ)	119/3366	(OPRD)	555/12679

TABLE 3 – Les cinq chemins syntaxiques les plus fréquents dans le corpus d'entraînement en fonction de la partie de discours de l'ancre

Si, pour une instanciation de cadre sémantique, on regroupe tous les chemins syntaxiques des rôles de ce cadre, on obtient une représentation linéaire de la structure syntaxique de ce cadre (à condition de connaître l'origine des chemins syntaxiques de ce cadre, à savoir l'ancre). Nous appelons cette représentation **signature syntaxique** du cadre, en voici un exemple pour la phrase *« Maurice bakes an apple pie in the oven. »*.

[bakes, (+SBJ), (+OBJ), (+LOC,+PMOD)]

Remarquons que l'ordre d'apparition des différents chemins syntaxiques n'a pas d'importance ici. On dit que la signature est **centrée sur l'ancre**.

3 Modèle

Notre approche consiste à modéliser les régularités syntaxiques des structures sémantiques du corpus d'entraînement afin de prédire les structures sémantiques du corpus de test. Pour cela, nous allons dans un premier temps proposer un système de prédiction de cadres sémantiques qui attache un cadre à un mot et des rôles à certains mots de la phrase. Ce système utilisera les données syntaxiques pour faire ses choix. Dans un second temps, nous évaluerons les performances de ce système. Pour rappel, nous ne cherchons à déterminer que les cadres dont les ancres ont pour parties de discours *Adjectif*, *Nom* ou *Verbe*.

Pour chaque ancre a apparaissant dans le corpus de test nous évaluons un certain nombre de possibilités d'étiquetages sémantiques et sélectionnons celle à laquelle le modèle attribue le meilleur score. L'étiquetage sémantique consiste à associer un cadre à a et des chemins syntaxiques aux rôles de ce cadre. En termes plus mathématiques, nous allons déterminer la réalisation de cadre $\widehat{F}$ la plus probable compte tenu d'une signature syntaxique S centrée sur l'ancre a :

$$\widehat{F} = \underset{F}{argmax} P(F|S)$$

$\widehat{F}$ est calculé en énumérant puis comparant toutes les représentations sémantiques compatibles avec la signature S. Pour énumérer toutes les représentations sémantiques possibles, nous créons une grammaire hors contexte qui reconnaît toutes les signatures possibles (symboles terminaux) en leur associant une structure sémantique (symboles non terminaux). La figure 2 illustre une dérivation possible de cette grammaire. Cette approche générative a deux intérêts principaux : elle produira toutes les analyses possibles et sera facile à modifier. La facilité de modification permettra de tester plusieurs hypothèses d'indépendance et d'améliorer l'analyse, à l'instar de Collins et de ses modèles génératifs d'analyse syntaxique (Collins, 1997).

Notre grammaire peut générer toutes les signatures composées d'une ancre et d'une suite de chemins syntaxiques (de la forme *[ancre, chemin$_1$, chemin$_2$, ..., chemin$_n$]*). Cette grammaire a pour axiome le symbole $\mathcal{F}$ pouvant se réécrire en chaque cadre. Chaque cadre peut être réécrit en une ancre (parmi les ancres possibles du cadre) et en un ensemble de rôles spécifiques à ce cadre. Chaque rôle peut être réécrit en chemin syntaxique. Dans cette grammaire, les cadres et les rôles sont donc des symboles non terminaux. Une signature reconnue par cette grammaire peut être reconnue de plusieurs façons différentes, chacune correspondant à un étiquetage sémantique particulier.

Afin de comparer les différents étiquetages sémantiques d'un même mot, nous leur attribuons une probabilité. Nous assignons alors à chaque règle de la grammaire une probabilité. Notre grammaire devient alors une grammaire hors contexte probabiliste. Les probabilités assignées à chaque règle sont estimées à partir du corpus d'entraînement. Pour sélectionner le meilleur étiquetage sémantique, nous utilisons l'algorithme d'Earley (Earley, 1970), qui calcule simultanément toutes les analyses et les représente sous forme de forêt partagée en un temps polynomial ($O(n^3)$ avec n le nombre de symboles non terminaux composant la signature). La représentation en forêt partagée est utile à la fois pour trouver la meilleure analyse et pour connaître le score d'une analyse partielle (notamment le score d'un cadre toutes sous-catégorisations confondues). La figure 2 représente un des arbres de dérivation pour la phrase *Maurice bakes an apple pie in the oven*. On remarque que l'axiome se réécrit en *Apply_heat*, qui lui même se réécrit en *Apply_heat(Verbe)*, introduisant la partie de discours de l'ancre avant de l'avoir générée.

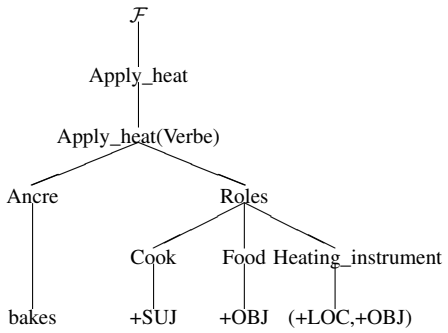


FIGURE 2 – Un arbre de dérivation de la phrase *Maurice bakes an apple pie in the oven* avec notre grammaire, en sélectionnant *bakes* comme ancre.

Notre grammaire permet maintenant de reconnaître des signatures syntaxiques en leur attribuant une structure sémantique. Cette grammaire permet d'assigner un cadre à l'ancre et un rôle à chaque chemin de la signature. Les rôles attribués sont soit des rôles appartenant au cadre, soit des rôles que nous qualifierons de vides si le chemin syntaxique correspondant ne participe pas à la structure sémantique.

Deux points importants restent à éclaircir, à savoir comment limiter le nombre de chemins et comment représenter le cas d'une ancre ne déclenchant pas de cadre.

Le cas de l'ancre ne déclenchant pas de cadre Rappelons que les ancres sont des mots répertoriés dans le lexique FrameNet et pouvant entraîner l'apparition d'un cadre sémantique. Certains occurrences de ces mots se font dans des situations dont la sémantique ne correspond pas à des cas définis sémantiquement. Ce cas est très présent (2070 occurrences pour 5504 ancres) et il faut donc modéliser dans nos grammaires ces cas. Notre approche est simple : nous allons créer un nouveau cadre, que nous nommerons le *cadre nul*. Ce cadre sera légèrement différent des autres cadres puisqu'il ne possèdera aucun rôle. Il ne possédera donc aucune réalisation syntaxique. Pour le prendre en compte, nous allons tout simplement estimer la probabilité qu'une ancre donnée déclenche le cadre nul à partir du corpus d'entraînement.

Sélection des chemins utilisables Pour sélectionner les chemins utilisables, nous ne prenons que les chemins liant une ancre à un rôle apparus au moins deux fois dans le corpus. Nous avons fait ce choix pour éliminer les chemins aberrants dûs à des erreurs d'analyse (l'analyseur en dépendance se trompe dans presque 1 cas sur 10 sur le corpus *Conll Shared Task 2009*) sans pour autant nous priver de données précieuses. En effet, le corpus d'entraînement comprend 2401 chemins ancre-rôle différents, et comme le montre le tableau 4, la variété des chemins décroît fortement lorsque leur nombre d'occurrences augmente. Ce choix permet aussi de limiter la taille des signatures que nous soumettons à l'algorithme d'Earley ainsi que la taille des grammaires qui sont utilisées.

occurrences	Nombre de chemins différents
> 0	2401
> 1	648
> 2	405
> 3	312
> 4	263
> 5	227

TABLE 4 – Nombre de chemins différents en fonction de leur nombre d'occurrences minimal

Exemple Prenons la phrase :

> We continued the meeting with a discussion

Si on se concentre sur l'ancre *continued*, il existe 7 chemins ayant cette ancre pour origine dans cette phrase. Admettons que parmi ces chemins seuls les suivants aient été observés plus d'une fois dans le corpus d'entraînement :

- $continued \rightarrow We$: +SBJ
- $continued \rightarrow meeting$: +OBJ
- $continued \rightarrow with$: +ADV
- $continued \rightarrow discussion$: +ADV,+PMOD

La signature que l'on soumettrait à la grammaire serait donc :

> continued,(+SBJ),(+OBJ),(+ADV),(+ADV,+PMOD)

L'algorithme d'Earley construit une forêt composée de tous les arbres possibles pour générer cette signature, la figure 3 présente deux des arbres de cette forêt d'analyses.

Les probabilités associées à chacun de ces arbres sont comparées afin de choisir la meilleure. Idéalement l'algorithme devrait choisir le cadre *activity_ongoing*, grâce à la présence d'un complément d'objet direct dans la phrase.

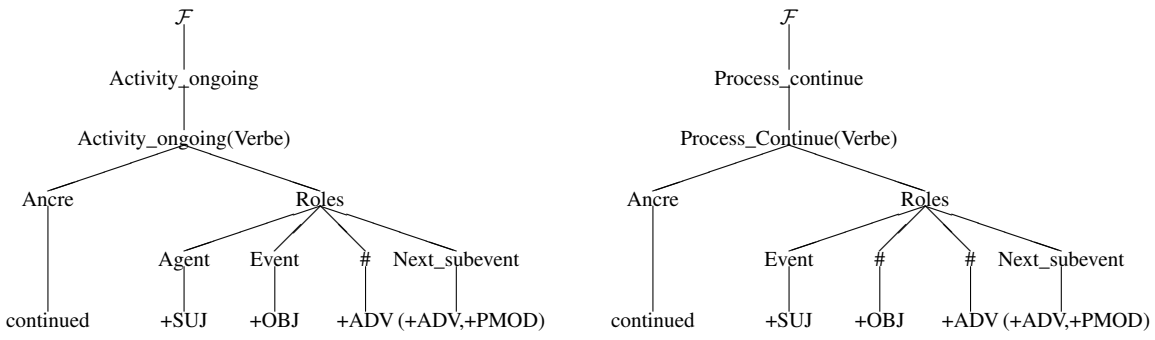


FIGURE 3 – Deux des arbres de la forêt d'analyses de la phrase *We continued the meeting with a discussion*. Le symbole # représente le rôle vide, c'est à dire les chemins qui ne correspondent à aucun rôle.

3.1 Grammaires utilisées

Pour nos expériences, nous avons créé trois grammaires différentes. La principale distinction entre ces grammaires tient dans la nature des interactions entre les rôles d'un même cadre. Nous appellerons *sous-catégorisation* l'ensemble des rôles observés pour une occurrence de cadre. La première grammaire va donner une importance capitale aux sous-catégorisations observées dans le corpus d'entraînement, alors que la seconde n'en tiendra absolument pas compte. La troisième grammaire sera plus nuancée, et plutôt que de s'attacher à la composition des sous-catégorisations, elle s'attachera au nombre de rôles réalisés dans chacune d'elles. Pour faire simple, nous avons deux grammaires extrêmes : la première qui suit scrupuleusement les sous-catégorisations observées lors de l'entraînement et la deuxième qui n'en tient aucunement compte. La troisième grammaire est en quelque sorte un compromis entre les deux. L'hypothèse commune à ces trois grammaires est que la réalisation syntaxique d'un rôle est indépendante de la réalisation des autres rôles instanciés pour le même cadre. Lorsqu'un rôle est attribué à une réalisation syntaxique, on ne regarde pas quels sont les autres réalisations syntaxiques sélectionnées. Une limite de cette approche provient du fait que seuls les cadres ayant été observés dans le corpus d'entraînement peuvent être générés par nos grammaires.

L'axiome de la grammaire est noté $\mathcal{F}$. Parmi les symboles non terminaux, les cadres seront notés F, les ancres A, les sous-catégorisations S, et les rôles R. Les symboles terminaux sont les mots du lexique et les chemins dans le format défini plus tôt, que nous désignerons respectivement par mot et $chemin$

Bien entendu les grammaires complètes ne sont pas présentées ici car trop vastes, les cadres sont alors numérotés de 1 à m et les chemins de 1 à p. Pour chaque cadre i, les ancres seront notées de 1 à a_i ,les rôles de 1 à r_i et les sous-catégorisations de 1 à s_i. Notons que la partie de discours de l'ancre influe sur les règles de réécriture des rôles en chemins.

3.1.1 G1 : Grammaire proche des sous-catégorisations

Cette grammaire cherche avant tout à reproduire les sous-catégorisations observées dans le corpus. Elle totalise 35468 règles.

	Probabilité	**Règle**	
1	$P(F_i)$	$\mathcal{F} \rightarrow F_i$	$\forall i \in 1,...,m$
2	$P(F_{nul})$	$\mathcal{F} \rightarrow F_{nul}$	
3	$P(F_{i,p}\|F_i)$	$F_i \rightarrow F_{i,p}$	$\forall i \in 1,...,m,\ p \in \{NN, JJ, VB\}$
4	$P(A_j, S_k\|F_{i,p}) = 1$	$F_{i,p} \rightarrow A_j S_k$	$\forall i \in 1,...,m,\ j \in 1,...,a_i,\ k \in 1,...,s_i$
5	$P(A_j\|F_{nul}) = 1$	$F_{nul} \rightarrow A_j$	$\forall j \in 1,...,a_{nul}$
6	$P(mot\|A_j)$	$A_j \rightarrow mot$	$\forall i \in 1,...,m,\ j \in 1,...,a_i$
7	$P(R_x, R_y, ...\|S_k)$	$S_k \rightarrow R_x R_y ...$	$\forall k \in 1,...,s_i,\ i \in 1,...,m;\ x, y \in [1, r_i], x \neq y$
8	$P(chemin\|R_r, p)$	$R_r \rightarrow chemin$	$\forall r \in 1,...,r_i,\ i \in 1,...,m$

Cette grammaire se base sur l'hypothèse que seules les sous-catégorisations observées lors de l'entraînement sont possibles. Pour que la grammaire accepte une signature, il faut que les chemins proposés correspondent chacun à un rôle

d'une sous-catégorisation observée dans le corpus d'entraînement, et que cette sous-catégorisation voie tous ses rôles instanciés.

Les probabilités des deux premières règles correspondent simplement à la probabilité d'un cadre d'apparaître. La règle numéro 3 permet de fixer la partie de discours de l'ancre de ce cadre pour la suite des règles, de façon à ce que les chemins soient cohérents avec cette partie de discours. La règle 4 permet de réécrire le cadre enrichi d'une partie de discours en deux symboles : l'un représentant les ancres possibles pour cette catégorie et pour ce cadre, l'autre représentant une sous-catégorisation possible pour ce cadre. L'ancre est réécrite sous la forme d'un lemme (symbole terminal), et la sous-catégorisation se réécrit en un ensemble de rôles. Chaque rôle se réécrit en un chemin qui tient compte de la partie de discours de l'ancre. Notons que les règles 4 et 5 ne sont que des réécritures, leurs probabilités sont donc égales à 1.

La principale limite de cette hypothèse réside dans le fait que les sous-catégorisations n'ayant jamais été observées ne peuvent pas être proposées par l'algorithme. Le corpus d'entraînement étant restreint, il ne serait pas surprenant que d'autres combinaisons puissent exister. L'hypothèse d'indépendance des réalisations syntaxiques de chaque rôle est matérialisée par la règle 8.

3.1.2 G2 : Grammaire ne tenant pas compte des sous-catégorisations

Cette grammaire se distingue de la précédente principalement par le fait qu'elle permet de générer des sous-catégorisations jamais observées dans le corpus d'apprentissage. Il n'y aura donc plus qu'un symbole non terminal par cadre pour représenter sa sous-catégorisation. Les r rôles possibles d'un cadre ne dépendent que de la probabilité d'avoir observé ce rôle avec ce cadre indépendamment des autres rôles qui seraient instanciés. Comme cette grammaire est moins restrictive que la précédente, elle ne comporte "que" 19124 règles de réécriture.

	Probabilité	**Règle**	
1	$P(F_i)$	$\mathcal{F} \rightarrow F_i$	$\forall i \in 1, ..., m$
2	$P(F_{nul})$	$\mathcal{F} \rightarrow F_{nul}$	
3	$P(F_{i,p}\|F_i)$	$F_i \rightarrow F_{i,p}$	$\forall i \in 1, ..., m,\ p \in \{NN, JJ, VB\}$
4	$P(A_j, S_k\|F_{i,p}) = 1$	$F_{i,p} \rightarrow A_j S_i$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
5	$P(A_j\|F_{nul}) = 1$	$F_{nul} \rightarrow A_j$	$\forall j \in 1, ..., a_{nul}$
6	$P(mot\|A_j)$	$A_j \rightarrow mot$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
7	$P(R_x\|S_i)$	$S_i \rightarrow R_x S_i$	$\forall i \in 1, ..., m\ x \in 1, ..., r_i$
8	$P(\varepsilon\|S_i)$	$S_i \rightarrow \varepsilon$	$\forall i \in 1, ..., m$
9	$P(chemin\|R_x, p)$	$R_r \rightarrow chemin$	$\forall x \in 1, ..., r_i,\ i \in 1, ..., m$

Cette fois l'hypothèse est que les rôles sont complètement indépendants les uns des autres. Le nombre de rôles peut donc varier librement.

Par rapport à la grammaire G1, la règle 7 est modifiée, et la 8 est ajoutée. La modification de la règle 7 permet une récursivité dans le choix des rôles, avec la règle 8 mettant un terme à la récursion en réécrivant la sous-catégorisation en ε. Le calcul de la probabilité de la règle 8 a nécessité quelques ajustements, car cette probabilité n'est pas estimable sur le corpus d'entraînement. En effet, elle représente le fait qu'un cadre possède un nombre fini de rôles. Sa probabilité devrait être égale à 1. Sa valeur a été déterminée empiriquement à 0.005. Cette masse de probabilités a été prélevée sur les règles du même type que la règle 7.

La principale limite de cette hypothèse réside dans le fait qu'un même rôle peut être présent plusieurs fois dans une même occurrence de cadre. De même, deux rôles s'excluant l'un l'autre pourraient ainsi cohabiter. Cependant, comme nous l'avons déjà énoncé, cette grammaire a l'avantage de pouvoir créer des sous-catégorisations nouvelles, bien que limitées par leur taille maximale.

3.1.3 G3 : Grammaire tenant compte de la taille des sous-catégorisations

Pour cette grammaire nous allons aussi faire l'hypothèse de l'indépendance des rôles, mais nous allons aussi donner de l'importance à la taille de la sous-catégorisation proposée. Pour cela, on introduit un nouveau type de symbole non terminal : $S_i(t = 1)$. Ce symbole représente les sous-catégorisations de taille 1 pour le cadre i. De même T_i sera la taille maximale des sous-catégorisations d'un cadre i. Cette grammaire, un peu plus coercitive que la précédente mais toujours bien moins que la première, compte 20366 règles de réécriture.

	Probabilité	**Règle**	
1	$P(F_i)$	$\mathcal{F} \rightarrow F_i$	$\forall i \in 1, ..., m$
2	$P(F_{nul})$	$\mathcal{F} \rightarrow F_{nul}$	
3	$P(F_{i,p}\|F_i)$	$F_i \rightarrow F_{i,p}$	$\forall i \in 1, ..., m,\ p \in \{NN, JJ, VB\}$
4	$P(A_j, S_i(t >= 0)\|F_{i,p}) = 1$	$F_{i,p} \rightarrow A_j S_i(t >= 0)$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
5	$P(A_j\|F_{nul}) = 1$	$F_{nul} \rightarrow A_j$	$\forall j \in 1, ..., a_{nul}$
6	$P(mot\|A_j)$	$A_j \rightarrow mot$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
7	$P(R_x, S_i(t >= 1)\|S_i(t >= 0))$	$S_i(t >= 0) \rightarrow R_x S_i(t >= 1)$	$\forall i \in 1, ..., m\ x \in 1, ..., r_i$
8	$P(\varepsilon\|S_i(t >= 0))$	$S_i(t >= 0) \rightarrow \varepsilon$	$\forall i \in 1, ..., m;$
9	$P(R_x, S_i(t >= 2)\|S_i(t >= 1))$	$S_i(t >= 1) \rightarrow R_x S_i(t >= 2)$	$\forall i \in 1, ..., m,\ x \in 1, ..., r_i$
10	$P(\varepsilon\|S_i(t >= 1))$	$S_i(t >= 1) \rightarrow \varepsilon$	$\forall i \in 1, ..., m;$
	$\vdots$		
11	$P(R_x\|S_i(t >= T_i - 1))$	$S_i(t >= T_i - 1) \rightarrow R_x$	$\forall i \in 1, ..., m\ x \in 1, ..., r_i$
12	$P(\varepsilon\|S_i(t >= T_i - 1))$	$S_i(t >= T_i - 1) \rightarrow \varepsilon$	$\forall i \in 1, ..., m;$
13	$P(chemin\|R_x, p)$	$R_x \rightarrow chemin$	$\forall x \in 1, ..., r_i,\ i \in 1, ..., m$

Par rapport à la grammaire précédente, cette grammaire introduit quantité de nouveaux symboles, correspondant aux tailles des sous-catégories en cours de création. Les règles 7 et 8 de G2 deviennent ici une série de règles tenant compte de la taille de la sous-catégorisation actuelle. Nous perdons donc le caractère récursif de la précédente. Les probabilités associées à chaque étape de la grammaire permettent de forcer les sous-catégorisations générées à contenir un nombre de rôles observé dans le corpus d'entraînement.

L'hypothèse principale ici est proche de la précédente : la réalisation d'un rôle est indépendante de celles des autres. On ajoute ici l'hypothèse que la taille des sous-catégorisations est importante. De ce fait, seules les tailles de sous-catégorisations observées dans le corpus d'entraînement seront valables. Cette grammaire étant basée sur la même hypothèse que la précédente, elle souffre des même biais. Ici aussi un même rôle peut être répété plusieurs fois pour une seule occurrence de cadre, et des rôles antagonistes peuvent apparaître simultanément.

3.2 Résultats

Avant de donner les résultats obtenus pour nos différentes grammaires, il nous faut définir quelles sont les mesures que nous allons utiliser pour les comparer et quelle référence nous allons utiliser.

Commençons par la référence. Nous avons créé une référence triviale qui à chaque ancre potentielle repérée assigne le cadre le plus observé pour cette ancre, cadre nul compris. Dans le cas où l'ancre n'a jamais été observée dans le corpus d'entraînement, on attribue à l'ancre le cadre compatible (selon le lexique) le plus observé avec des ancres de cette catégorie. Si jamais à ce stade toujours aucune solution n'a été trouvée, on attribue à l'ancre le cadre nul. Cette référence, contrairement à notre tâche, ne se charge que d'attribuer un cadre à une partie de discours, et ne se préoccupe pas des rôles de ce cadre.

La grammaire va nous permettre de calculer deux probabilités : la probabilité qu'un mot soit associé à un cadre et la probabilité d'une instanciation complète (cadre et sous-catégorisation). La première probabilité est comparable à la référence alors que la seconde ne l'est pas.

Deux types de prédiction sémantique sont faites :
- $\widehat{F} = \underset{F_{SC}}{argmax} P(F_{SC}|S)$ pour la sélection de cadres avec une sous-catégorisation ;
- $\tilde{F} = \underset{F}{argmax} \sum_{SC} P(F_{SC}|S)$ pour la sélection de cadres toutes sous-catégorisations confondues.

A l'issue des traitements on prédit pour un mot :

1. Son cadre le plus probable indépendamment des rôles instanciés ($\tilde{F}$) ;
2. L'instanciation la plus probable (sélection du cadre et des rôles $\widehat{F}$).

Pour mesurer les performances en sélection de cadres indépendamment, nous allons utiliser un *taux de réussite* (exprimé en pourcent). On compte le nombre d'ancres dans le corpus analysé, et le nombre d'ancres dont le cadre a été correctement assigné (cadre nul compris).

Les performances en étiquetage des rôles seront mesurées uniquement parmi les cadres correctement attribués. Nous utiliserons ici trois scores pour mesurer la performance : précision, rappel et F-mesure (tous les trois exprimés en pourcent). La précision mesure, parmi tous les mots qui ont été étiquetés rôles ceux qui l'ont été correctement. Le rappel mesure, parmi tous les mots qui sont des rôles dans la référence combien ont été correctement éiquetés. La F-mesure est une moyenne harmonique des deux scores précédents.

Pour mieux comprendre les forces et les faiblesses de notre automatisation nous allons donner chacune de ces quatre mesures selon la classe syntaxique de l'ancre.

La table 5 permet de comparer les résultats des trois grammaires et de la référence sans tenir compte des rôles (correspond au calcul de $\tilde{F}$).

Partie de discours	Nombre d'occurrences	Référence	G1	G2	G3
Verbes	775	58,32	51,48	53,16	53,55
Adjectifs	660	55,15	48,48	49,09	49,09
Noms	1687	62,89	55,6	58,92	58,86
global	3122	60,12	53,07	55,41	55,48

TABLE 5 – Taux de réussite exprimés en pourcentages, pour la référence et les trois grammaires dans la tâche de sélection de cadres sans tenir compte des rôles

On observe dans la table 5 que la référence est supérieure en tous points à nos grammaires sur la tâche d'assignation de cadres sémantiques (sans sélection des rôles). Les grammaires G2 et G3 donnent les meilleurs résultats. La grammaire G1 ne permet pas de prédire les sous-catégorisations qui n'ont jamais été observées dans le corpus d'entraînement (622 occurrences parmi 4427 dans le corpus de test), ce qui explique probablement ses moindres résultats. La grammaire G2, malgré sa rusticité permet de dépasser les performances de G1, certainement car elle ne se restreint pas aux sous-catégorisations observées dans le corpus d'entraînement. G3 obtient les meilleurs résultats, car à l'instar de G2, elle permet de prédire des sous-catégorisations jamais observées. En revanche, elle introduit un peu plus de contraintes en disqualifiant les sous-catégorisations ayant un nombre de rôles improbable. Cependant la différence entre G2 et G3 n'est pas significative, ce qui laisse penser que l'ajustement de la règle 8 de G2 est suffisante pour contraindre le nombre de sous-catégorisations.
Il est intéressant de noter que les performances dans cette tâche ne sont pas directement liés à l'ambiguïté moyenne de ces parties de discours (cf. table 1).

3.2.1 G1 : Grammaire proche de la sous-catégorisation

La table 5(a) représente les résultats de la grammaire G1 pour la tâche d'analyse sémantique complète (cadre et rôles, équivalent à $\widehat{F}$). On peut remarquer que la précision sur les verbes est assez bonne, les prédictions de rôles faites sur cette catégorie sont donc raisonnables (63,82), dès lors que le cadre a été bien sélectionné. Par contre les résultats sur les adjectifs et noms ont une précision globalement plus faible, probablement du fait que les moyens de réalisation syntaxique des actants sémantiques sont plus limités, d'où plus d'ambiguïté.

3.2.2 G2 : Grammaire ne tenant pas compte des sous-catégorisations

La table 5(b) présente les résultats de la grammaire G2 pour la tâche de sélection des cadres avec rôles. Les performances en sélection de cadres sont légèrement supérieures à celles de la grammaire G1, surtout en ce qui concerne les noms. La précision en sélection de rôles est cependant meilleure pour les verbes et pour les noms. On remarque également que les scores de rappel sont inférieurs à la grammaire G1 : contrairement à ce que nous pensions, cette grammaire sélectionne moins de rôles (rappel bas, notamment pour les verbes : 26,30) que la précédente. Les scores de F-mesure montrent cependant que la grammaire 1 est plus performante pour sélection de rôles.

(a) G1

PoS	Taux de réussite sélection cadre	rôles concernés	Precision rôles	Rappel rôles	F-mesure rôles
Verbes	51,23	723	63,82	43,87	52,00
Adjectifs	47,27	556	59,33	71,58	64,88
Noms	54,71	1768	49,08	57,54	52,97
global	52,27	3047	53,96	56,24	55,08

(b) G2

PoS	Taux de réussite sélection cadre	rôles concernés	Precision rôles	Rappel rôles	F-mesure rôles
Verbes	52,26	736	67,72	26,30	37,89
Adjectifs	47,34	542	59,17	59,64	59,40
Noms	58,49	1690	56,04	51,16	53,49
global	54,59	2968	58,44	45,37	51,08

(c) G3

PoS	Taux de réussite sélection cadre	rôles concernés	Precision rôles	Rappel rôles	F-mesure rôles
Verbes	52,26	735	67,73	25,99	37,57
Adjectifs	47,58	546	58,82	59,74	59,28
Noms	58,51	1689	56,53	51,03	53,64
global	54,64	2970	58,66	45,25	51,09

TABLE 6 – Résultats des grammaires présentées en 3.1 pour la sélection de cadres et de rôles ($\widehat{F}$), en fonction de la partie de discours de l'ancre
La colonne *Taux de réussite sélection cadre* renseigne sur la quantité de cadres correctement attribués, alors que les suivantes sont des statistiques sur les rôles des cadres correctement attribués.

3.2.3 G3 : Grammaire tenant compte de la taille des sous-catégorisations

La table 5(c) représente les résultats pour la grammaire G3 en sélection de cadres complets (cadre et rôles). Bien que meilleure que la grammaire G2, les résultats de ces deux grammaires sont quasiment identiques. Peut-être que la contrainte sur la taille des sous-catégorisations n'est pas suffisante pour creuser l'écart.

4 Conclusion

Les expériences que nous venons de présenter nous montrent que la syntaxe d'une phrase ne permet d'induire que partiellement sa structure sémantique. Nous avons cependant pu remarquer des résultats significativement meilleurs lorsque les ancres des cadres sont des verbes, malgré une ambiguïté sémantique en moyenne plus grande. Ceci s'explique par la richesse des réalisations syntaxiques associées à cette partie de discours : une réalisation syntaxique est toujours associée à un même rôle. Quant aux noms et aux adjectifs, les réalisations syntaxiques qui leurs sont associées sont peu variées : une même réalisation syntaxique peut être associée à plusieurs rôles.
Les grammaires que nous avons présentées ici ne sont pas encore au niveau de notre référence, et bien des améliorations peuvent encore y être apportées. La modification la plus importante à ajouter serait de prendre en compte la nature des mots jouant un rôle dans les cadres. Par exemple pour l'action *manger*, seuls les *êtres vivants* sont capables d'endosser le rôle de *mangeur*. Pour réaliser cette modification, nous pourrions soit faire appel à la sémantique distributionnelle à partir de grands corpus, soit utiliser les ressources réalisées dans des projets comme *WordNet*.
Concernant les améliorations au niveau de la syntaxe, comme nous traitons les cadres indépendamment les uns des autres, nous pourrions utiliser le corpus des phrases d'exemple, annoté partiellement pour la sémantique. Cet ajout permettrait d'enrichir notre corpus d'entraînement et ainsi d'affiner les probabilités de nos grammaires.
Au niveau des données FrameNet, nous pourrions tirer un meilleur parti des informations représentées dans les données. En effet, il existe des relations de plusieurs types entre cadres sémantiques, comme des relations d'héritage ou d'utilisation. En ce qui concerne la première, elle permettrait de pallier le manque de données dans notre corpus de phrases :

les structures syntaxiques d'un cadre père sont proches de celles de ses fils, de plus il existe une table de correspondance entre leurs rôles. Par exemple le cadre *Scrutiny* (prêter attention) est un cadre père de *Research* (rechercher). Grâce à cette méthode, il est possible de regrouper les données des cadres fils au sein d'un même père afin d'augmenter la précision, du moins pour les rôles qui correspondent entre père et fils. La relation d'utilisation est une relation observable dans le corpus, elle consiste à savoir quels cadres sont observés en *symbiose*, c'est à dire qu'un acteur de l'un est un acteur ou une ancre de l'autre. Cette relation, en plus d'être observable, est définie dans les données FrameNet. Elle nous permettrait de créer un modèle identifiant un cadre en fonction des autres cadres de la phrase courante.

Remerciements

Ces travaux sont financés par le projet ANR Asfalda ANR-12-CORD-0023. Nous tenons à remercier Emmanuel Prunet qui nous a fourni une implémentation efficace de l'analyseur d'Earley, ainsi que Marie Candito, Alexis Nasr et Benoit Favre pour leurs relectures et leur encadrement.

Références

BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 89–97 : Association for Computational Linguistics.

BOHNET B., NIVRE J., BOGUSLAVSKY I., GINTER R. F. F. & HAJIC J. (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Associtation for Computational Linguistics*, **1**.

COLLINS M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, p. 16–23 : Association for Computational Linguistics.

DAS D., SCHNEIDER N., CHEN D. & SMITH N. A. (2010). Probabilistic frame-semantic parsing. In *Human language technologies : The 2010 annual conference of the North American chapter of the association for computational linguistics*, p. 948–956 : Association for Computational Linguistics.

DAS D. & SMITH N. A. (2011). Semi-supervised frame-semantic parsing for unknown predicates : Supplementary material. In *Proc. of ACL*.

EARLEY J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, **13**(2), 94–102.

FILLMORE C. (1982). Frame semantics. *Linguistics in the morning calm*, p. 111–137.

FILLMORE C. J. (1976). Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, **280**(1), 20–32.

FILLMORE C. J. (1977). The case for case reopened. *Syntax and semantics*, **8**(1977), 59–82.

FILLMORE C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, **6**(2), 222–254.

FILLMORE C. J. & BAKER C. (2010). A frames approach to semantic analysis. *The Oxford handbook of linguistic analysis*, p. 313–339.

FILLMORE C. J. & BAKER C. F. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.

MILLER G. & FELLBAUM C. (1998). Wordnet : An electronic lexical database.

MODI A., TITOV I. & KLEMENTIEV A. (2012). Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, p. 1–7 : Association for Computational Linguistics.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank : An annotated corpus of semantic roles. *Computational Linguistics*, **31**(1), 71–106.

Traitement FrameNet des constructions à attribut de l'objet

Marianne DJEMAA
INRIA, UMR-I 001, Alpage, Univ Paris Diderot, Sorbonne Paris Cité, F–75013 Paris, France
marianne.djemaa@inria.fr

Résumé. Dans le cadre du projet ASFALDA, qui comporte une phase d'annotation sémantique d'un FrameNet français, nous cherchons à fournir un traitement linguistiquement motivé des constructions à attribut de l'objet, un exemple typique de divergence syntaxe-sémantique.
Pour ce faire, nous commençons par dresser un panorama des propriétés syntaxiques et sémantiques des constructions à attribut de l'objet. Nous étudions ensuite le traitement FrameNet des verbes anglais typiques de cette construction, avant de nous positionner pour un traitement homogénéisé dans le cas du FrameNet français.

Abstract. Within the ASFALDA project, which includes the production of a French FrameNet, we try to provide a linguistically motivated treatment for a typical example of syntax-semantics mismatch : object complement construction. In order to do so, we first give an overview of syntactic and semantic properties of object complement constructions. Next, we study the way FrameNet deals with English verbs taking part in those constructions, and finally take a stance for a homogenized treatment of the construction within the French FrameNet.

Mots-clés : FrameNet, français, construction à attribut de l'objet, divergence syntaxe-sémantique.

Keywords: FrameNet, French, object complement construction, syntax-semantics mismatch.

1 Introduction

Ce travail s'inscrit dans le cadre du projet ASFALDA (Candito *et al.*, 2014), qui a pour but de développer un corpus du français annoté sémantiquement et un analyseur sémantique prédisant le même type d'annotations. Ce projet s'appuie sur le modèle FrameNet (Baker *et al.*, 1998) d'annotation en cadres et rôles sémantiques et sur des corpus préexistants annotés pour la morphologie et la syntaxe : le French TreeBank, (Abeillé *et al.*, 2003) (ci-après FTB) et le Sequoia Treebank (Candito & Seddah, 2012), qui suit le schéma d'annotation du FTB. Le projet FrameNet initial propose un ensemble structuré de situations prototypiques, appelées *frames*, associées à des caractérisations sémantiques des participants impliqués (ci-après les *rôles*), ainsi qu'un lexique de lexèmes évoquant ces frames et un ensemble d'annotations en frames pour l'anglais. Le projet ASFALDA propose de tirer parti de cette structure, qui s'est déjà avérée largement portable à d'autres langues (Boas, 2009), pour construire un FrameNet français, i.e. (i) un lexique indiquant les lexèmes français pouvant évoquer les frames (ci-après les *déclencheurs* de frames), et (ii) l'annotation des occurrences de frames et des rôles associés sur les corpus arborés cités supra.

Pour réaliser la ressource anglaise FrameNet, le parti-pris semble avoir été de conserver au maximum une orientation sémantique : les critères mis en avant pour le "découpage" en frames ou pour le typage des rôles sont avant tout sémantiques, au contraire par exemple des classes de Levin (1993) pour les verbes anglais, qui sont construites principalement via les alternances syntaxiques, ou des tables du Lexique-Grammaire (Gross, 1975; Leclère, 2002) pour le français. Cependant des critères syntaxiques sont tout de même utilisés dans FrameNet : sauf dans des cas dûment recensés, les rôles sémantiques essentiels d'un frame doivent pouvoir être réalisés localement pour tous les lexèmes évoquant ce frame [1]. Il est donc attendu que les cas typiques de "divergences" entre syntaxe et sémantique posent problème.

Dans cet article nous nous penchons sur la construction à attribut de l'objet, qui est l'un de ces cas. Il s'agit de constructions de type N_0 V N_1 X où N_0 est le sujet de V, et N_1 l'objet direct de V et le sujet du prédicat tête de X. Les V de ces constructions présentent la particularité de sous-catégoriser trois arguments syntaxiques (N_0, N_1 et X) et seulement deux

1. Les cas de non-localité autorisés sont recensés, et relèvent de phénomènes syntaxiques comme le contrôle ou les constructions à verbe support (Ruppenhofer *et al.*, 2006, pp. 27-31). En pratique cela dit, les annotations anglaises FrameNet ne suivent pas toujours ces directives.

arguments sémantiques (N_0 et X). X est quant à lui un prédicat se rapportant à N_1.

Pour pouvoir fournir un traitement linguistiquement motivé de ces constructions, nous dressons d'abord un panorama de leurs propriétés syntaxiques et sémantiques en passant en revue les différentes sous-catégories identifiables. Nous nous intéressons ensuite aux arguments qui permettent de considérer que le verbe de ces constructions sous-catégorise trois arguments syntaxiques pour deux arguments sémantiques, et examinons de plus près la potentielle équivalence sémantique entre complétive et construction à attribut de l'objet. Ruppenhofer *et al.* (2006) ne mentionnant pas de choix général pour le traitement de l'alternance CAO/complétive, nous considérons pour différents déclencheurs repérés les rôles prévus par les frames qu'ils évoquent et la réalité des annotations effectuées. Nous nous positionnons enfin pour un traitement homogénéisé de la construction dans le FrameNet français.

2 Typologie

Cette section répertorie quelques critères de classement des constructions à attribut de l'objet, qui constituent un ensemble assez vaste et très étudié, afin de circonscrire les cas sur lesquels nous nous pencherons dans cet article.

2.1 Complément vs. modifieur

On reconnaît généralement deux grands types d'attribut de l'objet, que Guimier (1999) appelle les "compléments attributifs" et "modifieurs attributifs" (respectivement (1a) et (1b)).

(1) a. *Sam croit son café chaud.*
b. *Sam boit son café chaud.*

Cette dichotomie s'appuie sur le caractère sous-catégorisé (pour les compléments attributifs) vs. optionnel (pour les modifieurs attributifs) de l'attribut de l'objet dans ces constructions et le fait qu'il constitue un argument sémantique du verbe uniquement dans les cas de construction à complément attributif.

Les deux catégories ont été abondamment décrites dans la littérature : Riegel (1981) parle de verbes essentiellement attributifs et verbes occasionnellement attributifs, puis dans (Riegel, 1991) de construction à double-complémentation et d'attribut amalgamé. Rémi-Giraud (1991) parle d'attributs obligatoires vs. facultatifs et Muller (2000) de verbes opérateurs et verbes à concomitance. D'autres travaux, comme (Blanche-Benveniste, 1991) et (Tobback, 2005) identifient la dichotomie sans en nommer les catégories.

Nous recensons ici les arguments principaux (syntaxiques et sémantiques) cités par Guimier (1999, chap. 1 section 2.1) qui justifient cette distinction.

Plusieurs tests syntaxiques mettent en évidence le caractère sous-catégorisé des compléments attributifs et celui d'ajout des modifieurs attributifs. Le critère le plus clivant est le caractère obligatoire des compléments attributifs vs. optionnel pour les modifieurs attributifs. Ces derniers peuvent en effet être supprimés sans incidence sur la grammaticalité de la phrase et avec stabilité du sens du verbe (exemple (1d)), ce qui n'est pas le cas des compléments attributifs de l'objet. Par exemple en (1c) la suppression de l'attribut donne une phrase grammaticale mais inacceptable. Plus précisément, on peut obtenir une phrase acceptable avec la suppression d'un complément attributif, comme en (5b), mais la suppression modifie le rôle sémantique joué par l'objet direct (nous y revenons infra). Autre test syntaxique permettant de distinguer compléments et modifieurs attributifs : seuls les dépendants des compléments attributifs peuvent être extraits ou cliticisés (exemples (2) et (3), repris des exemples (44) et (45) de Guimier (1999)).

(1) a. *Sam croit son café chaud.*
b. *Sam boit son café chaud.*
c. *#Sam croit son café.*
d. *Sam boit son café.*

(2) a. *Les aubergines, Paul les croyait farcies de chair à saucisse*
b. *De quel type de chair Paul les croyait-il farcies ?*
c. *La chair à saucisse, dont Paul les croyait farcies, était avariée.*
d. *Paul en croyait les aubergines farcies.*

(3) a. *Les aubergines, Paul les a mangées farcies de chair à saucisse*
b. **De quel type de chair Paul les a-t-il mangées farcies ?*
c. **La chair à saucisse, dont Paul les a mangées farcies, était avariée.*
d. **Paul en a mangé les aubergines farcies.*

En ce qui concerne la sémantique, Guimier (1999) pointe que le modifieur attributif n'appartient pas à la structure argumentale du verbe principal, tandis que le complément attributif est un argument sémantique du verbe avec lequel il apparaît. Ceci peut être vérifié simplement avec le test de suppression, cette fois utilisé pour dégager des propriétés sémantiques : lorsqu'on supprime un modifieur attributif, la phrase reste non seulement toujours acceptable (1d), mais on a bien une relation d'implication entre la version avec le modifieur attributif et la version sans celui-ci.

(4) a. *Sam boit son café chaud.*
b. $\Rightarrow$ *Sam boit son café.*

Lorsqu'on supprime un complément attributif, on peut obtenir soit une inacceptabilité (1c), soit une phrase certes acceptable mais sémantiquement non impliquée par la version avec attribut (5b), ce qui montre que l'on passe à une version où l'objet direct joue un rôle sémantique différent.

(5) a. *Sam croit Paul honnête.*
b. $\not\Rightarrow$ *Sam croit Paul.*

Pour la suite de ce travail, nous nous intéresserons uniquement aux constructions à *complément* attributif de l'objet, qui posent le problème le plus crucial de divergence syntaxe-sémantique [2]. Sauf mention explicite, c'est dorénavant à celles-ci que nous ferons référence quand nous parlerons de *constructions à attribut de l'objet* (ou *CAO*).

2.2 Différentes catégories d'attributs

La fonction d'attribut de l'objet peut être occupée par des constituants de catégorie variée.

Nous donnons ci-dessous un exemple de chaque catégorie ainsi qu'un nombre d'occurrences comptées sur nos corpus cibles. Pour le FTB, nous utilisons la version sortie à l'occasion de la shared task SPMRL 2013 (Seddah *et al.*, 2013), comprenant 18 535 phrases avec annotations fonctionnelles, et pour le Sequoia Treebank, la version 4.0 [3] qui en comprend 3 204.

Ces nombres d'occurrences ne sont donnés qu'à titre indicatif et pour avoir une idée des proportions de constructions à attributs de l'objet présentes dans les corpus. En effet, n'ayant compté que les constituants annotés *ATO*, l'étiquette FTB dénotant un attribut de l'objet, nous ne relevons pas les cas de CAO avec changement de diathèse (les CAO dont le verbe principal est au passif sont surfaciquement annotés comme attributs du sujet). Nous sommes par ailleurs tributaires d'éventuelles erreurs d'annotation dans l'un ou l'autre des corpus. Parmi les exemples effectivement annotés ATO dans ces corpus, on trouve :

— des syntagmes adjectivaux (166 occurrences) :
(6) *Les analystes trouvent [$_{NP\text{-}OBJ}$les chiffres] [$_{AP\text{-}ATO}$"raisonnablement bons"].*

— des infinitives (88 occurrences) :
(7) *Les cégétistes auront néanmoins vu [$_{NP\text{-}OBJ}$leur "part de marché"] [$_{VPinf\text{-}ATO}$passer de 42% à 33%].*

— des syntagmes prépositionnels (55 occurrences) :
(8) *L'ethnologue [...] considère [...] [$_{NP\text{-}OBJ}$l'entreprise africaine] [$_{PP\text{-}ATO}$comme un phénomène culturel].*

— des participiales (29 occurrences) :
(9) *un montant [$_{NP\text{-}OBJ}$que] les experts étrangers jugent [$_{VPpart\text{-}ATO}$sous-estimé]*

— des syntagmes nominaux (27 occurrences) :
(10) *[les] services des finances qui jugeaient [$_{NP\text{-}ATO}$pur gaspillage] [$_{NP\text{-}OBJ}$les 15 milliards de francs requis pour la liaison Rhône-Rhin]*

2. Pour les modifieurs, dont nous ne parlons pas plus avant, notons rapidement que l'annotation FrameNet se fera par deux frames distincts pour les deux relations de prédication, et que la relation de concomitance (Muller, 2000) qui lie ces deux prédications ne sera pas capturée.
3. https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia

On répertorie aussi des cas où l'attribut de l'objet est un syntagme adverbial (exemple (12)) ou une proposition relative (exemple (11)) mais on n'en trouve aucune occurrence dans le FTB.

(11) *Je vois [$_{NP\text{-}OBJ}$Sam] [$_{Srel\text{-}ATO}$qui arrive].*

(12) *Alex trouve [$_{NP\text{-}OBJ}$ça] [$_{Ad\text{-}ATO}$bien].*

2.3 Typologie sémantique

On peut regrouper les verbes à compléments attributifs en plusieurs catégories. Cette typologie est très variable selon les auteurs, c'est pourquoi nous n'en tracerons qu'une grossière esquisse.

2.3.1 Verbes pour lesquels une alternance avec une phrase à complétive est possible

Il s'agit de verbes pour lesquels la construction à attribut de l'objet de forme N_0 *V* N_1 *A* peut être paraphrasée par une phrase où le verbe *V* sera complémenté par une proposition complétive *que* N_1 *copule A*.

(13) *a. Sam trouve Alex sympathique.*
b. Sam trouve qu'Alex est sympathique.

On verra en 3.3 que cette relation de paraphrase peut être questionnée, mais elle fonctionne en première approximation.

Parmi les verbes autorisant cette alternance, Guimier (1999) distingue les verbes comme *croire* et les verbes comme *vouloir*, qui se différencient des premiers parce qu'ils n'autorisent ni la cliticisation des compléments de l'attribut, ni l'insertion de la copule *être* devant l'attribut en cas de relativisation de l'objet direct. De plus, la complétive des verbes comme *vouloir* est généralement au subjonctif.

(14) *a. Son film, que Sam croit (être) fidèle à la réalité historique.*
*b. Son film, que Sam veut *(être) fidèle à la réalité historique.*

(15) *a. Sam croit que son film est fidèle à la réalité historique.*
b. Sam veut que son film soit fidèle à la réalité historique.

2.3.2 Verbes de perception

Les verbes tels que *voir*, *entendre* et *sentir*, permettent eux-aussi une construction avec complétive *que* N_1 *copule A* comme en (16b), mais le sens de cette dernière ne correspond pas à celui de la version CAO (exemple (16a)) qui serait mieux paraphrasée par une autre construction (exemple (16c)).

(16) *a. Sam voit Alex ivre.*
b. Sam voit qu'Alex est ivre.
c. Sam voit Alex et/au moment où Alex est ivre

Guimier (1998) rattache les constructions à attribut de l'objet *adjectival* de ces verbes de perception aux constructions à attribut accessoire de l'objet de Le Goffic (1994). Ces dernières correspondent aux constructions à modifieur attributif vues plus haut – en particulier parce que, comme évoqué en 2.1, la suppression de l'adjectif attribut n'entraîne ni agrammaticalité, ni changement de sens de l'énoncé restant. En outre, la version CAO implique sémantiquement la version sans attribut, comme on peut le voir en (17).

(17) *a. Sam entend Alex ivre.*
b. ⇒ Sam entend Alex.

À première vue, cela vaudrait aussi pour les constructions attributives infinitives et relatives :

(18) *a. Sam entend Alex arriver.*
b. ⇒ Sam entend Alex.

(19) a. *Sam entend Alex qui arrive.*
b. ⇒ *Sam entend Alex.*

Néanmoins, des exemples clairs tels que (20) et (21) (Guimier, 1998), où la relation d'implication ne tient pas, nous montrent qu'on a bien affaire à des constructions à *complément* attributif de l'objet.

(20) a. *Sam entend le prof se faire chahuter.*
b. ⇏ *Sam entend le prof.*

(21) a. *Sam entend le prof qui se fait chahuter.*
b. ⇏ *Sam entend le prof.*

2.3.3 Verbes causatifs

Toujours selon Guimier (1999), ces verbes se distinguent de ceux de la catégorie précédente en ce qu'ils ne peuvent alterner avec une construction complétive.

(22) a. *Sam rend Alex malade.*
b. **Sam rend que Alex est malade.*

De plus, leur objet direct n'est pas nécessairement réalisé (exemple (23)) et certains d'entre eux n'autorisent que les attributs de type adjectival (exemples (24)).

(23) *L'absinthe rend fou*

(24) a. **Sam rend Alex en colère.*
b. **Sam rend Alex debout.*

Notre travail se préoccupant beaucoup des problèmes d'alternance CAO/complétive, nous ne nous intéresserons plus par la suite aux constructions à attribut de l'objet à verbes causatifs.

3 Divergence syntaxe-sémantique

Nous avons vu en section 2.3 qu'au moins pour certains types de CAO, il existait, en première approximation, une relation de paraphrase entre les deux versions de l'alternance CAO vs. complétive (cf. exemple (13)). L'existence d'une telle équivalence paraît indiquer que la séquence N_1 *X* d'une phrase à CAO correspond à un seul argument sémantique de *V*, comme c'est le cas pour le constituant *que* N_1 *copule X*. Or, cette séquence N_1 *X* se décompose en deux constituants syntaxiques. C'est cette opposition apparente entre le nombre d'arguments syntaxiques et sémantiques du verbe d'une CAO qui fait que ces dernières sont souvent pointées comme un cas de divergence syntaxe-sémantique.

Dans cette section, nous nous intéresserons à cette divergence, en étudiant les indices favorisant une analyse syntaxique en trois arguments, ainsi que ce qui motive une analyse sémantique en deux arguments. Nous nous pencherons ensuite de manière plus détaillée sur la relation de paraphrase évoquée supra.

3.1 Nombre d'arguments syntaxiques

La partie annotation du projet ASFALDA consistant en fait à ajouter une couche d'annotations sémantiques sur un schéma d'annotation existant, les choix quant à la représentation syntaxique des CAO ne nous reviennent pas.

Le guide d'annotation du FTB annonce l'objectif de *"contribuer à l'émergence d'un standard de découpage en constituants, sans doute un peu grossier mais suffisamment consensuel, et traductible dans différents cadres théoriques"* (Abeillé *et al.*, 2004, p. 6). Pour les CAO, de la forme N_0 *V* N_1 *X*, le FTB analyse la séquence N_1 *X* en deux constituants, au contraire de nombreuses analyses de type génerativiste qui en considèrent un seul. Cette analyse en "petite proposition" a particulièrement cours en anglais (e.g. Williams (1975), Stowell (1995), Rothstein (1995) et Hoekstra (1992)).

Nous redonnons rapidement dans cette section les arguments qui justifient le choix du FTB d'analyser la phrase (25a) en (25b) plutôt que (25c).

(25) a. *Sam trouve cette personne sympathique.*
b. *Sam trouve [cette personne] [sympathique].*
c. *Sam trouve [cette personne sympathique].*

Guimier (1999) puis Tobback (2005) pointent plusieurs tests classiques de constituance :

— La relativisation de la séquence *[N_1 X]* est agrammaticale alors que l'objet peut parfaitement être relativisé :

(26) a. *Cette personne que Sam trouve sympathique*
b. **Cette personne sympathique que Sam trouve*

— Il est tout à fait possible de faire porter une question sur l'objet ou sur son attribut, alors qu'il est impossible de questionner la séquence *[N_1 X]* dans son intégralité :

(27) a. *Qui est-ce que Sam trouve sympathique ? Cette personne.*
b. *Comment est-ce que Sam trouve cette personne ? Sympathique.*
c. **Qu'est-ce que Sam trouve ? Cette personne sympathique.*

— De même, il est tout à fait possible de cliver l'objet ou son attribut, alors que le clivage de la séquence *[N_1 X]* est agrammatical :

(28) a. *C'est cette personne que Sam trouve sympathique.*
b. *C'est sympathique que Sam trouve cette personne.*
c. **C'est cette personne sympathique que Sam trouve.*

— Il est possible de cliticiser l'objet mais pas vraiment la séquence *[N_1 X]* (Guimier, 1999, exemple (144a))

(29) a. *Sam la trouve sympathique.*
b. *? ?Sam trouve cette personne sympathique mais Alex ne le trouve pas.*

Certains de ces tests sont nuancés par Tobback (2005), qui pointe que si N_1 et X forment chacun un constituant distinct, cela ne les empêche pas de former ensemble un constituant. L'auteur note aussi qu'il n'y a pas unanimité sur l'inacceptabilité de la pronominalisation de la séquence *[N_1 X]*. Pour ce qui est du clivage, elle rappelle l'argument de Guimier (1999) que certains verbes anglais comme *fear* acceptent le clivage de *[N_1 X]*, et constate qu'à l'inverse le clivage n'est pas toujours très grammatical même pour des séquences dont le statut de constituant est peu controversé. Il nous semble cependant clair que d'après les critères appliqués au sein du FTB pour dégager les constituants, les tests donnent de manière évidente une analyse en deux constituants séparés.

3.2 Nombre d'arguments sémantiques

Il s'agit dans cette section de voir dans quelle mesure il y a divergence entre syntaxe et sémantique pour les CAO, i.e. si les trois compléments sous-catégorisés sont sémantiquement sélectionnés par le verbe.

Guimier (1999) montre facilement que N_1 est sélectionné par X : la différence d'acceptabilité entre (30a) et (30b) indique que l'attribut impose des restrictions de sélection à l'objet, qui est donc bien son sujet sémantique.

(30) a. *Sam croit la route praticable.*
b. *#Sam croit le café praticable.*

Elle entend ensuite montrer qu'en revanche, N_1 n'est pas sélectionné par le verbe (dans le cas où on a affaire à un complément attributif et non pas à un modifieur). Son argumentation repose sur des paires comme (31).

(31) a. *Sam croit la route praticable.*
b. *#Sam croit la route.*

Nous avons déjà relayé en 2.1 sa preuve selon laquelle dans le cas d'un complément attributif, la version sans l'attribut met en jeu des propriétés sémantiques différentes : la phrase devient agrammaticale ou prend un sens différent, parce que l'objet ne remplit pas le même rôle dans une construction transitive stricte (sans CAO) et dans une CAO. L'inacceptabilité de (31b) ne peut donc pas nous donner de détails sur la structure argumentale de *croire* dans un cas de construction à attribut de l'objet.

Pour tenter de montrer que dans une construction à attribut de l'objet, l'objet direct du verbe principal n'est pas un argument sémantique de ce dernier, nous pointerons donc simplement que nous n'avons pu ni trouver, ni construire d'exemple

où la construction à attribut de l'objet serait inacceptable alors même que l'objet et son attribut sont sémantiquement compatibles.

Pour le cas des verbes de perception, comme nous l'avons déjà mentionné, l'implication observable en (17) est souvent présente, puisque l'objet direct est bien l'entité à l'origine de ce qui est perçu. On pourrait donc penser que l'objet direct est bien un argument sémantique du verbe de perception, contrairement à ce qu'on vient de voir pour les verbes comme *croire*. Encore une fois, des exemples comme (20), où le son perçu n'est pas émis par l'objet *le professeur* mais bien par ses élèves le chahutant, nous montrent que dans les CAO, l'objet direct n'est pas forcément cette entité et n'est donc pas un argument sémantique du verbe de perception [4].

3.3 Alternance avec construction à complétive

Cela a été évoqué supra : pour une certaine classe de verbes à CAO, on considère traditionnellement (voir par exemple Riegel *et al.* (2009)) qu'il existe une relation de paraphrase entre la phrase avec attribut de l'objet de la forme N_0 *V* N_1 *A* et la phrase avec complétive telle que N_0 *V que* N_1 *copule A*. Ainsi (13a) et (13b) apparaissent de sens similaire malgré leur différence de construction.

Avant de nous poser la question de la réalité d'une équivalence sémantique entre phrase avec CAO et phrase avec complétive, il semble pertinent de mentionner que parmi les verbes autorisant cette alternance, on rencontre des cas où la réalisation d'une version CAO d'une phrase avec complétive est tout bonnement impossible. Cela s'explique par le fait qu'un cas avec complétive est sémantiquement plus générique qu'un cas avec CAO, qui, elle, impose forcément une prédication sur l'actant réalisé par N_1. Dès lors, on peut construire des exemples comme (32) où un verbe qui autorise généralement la construction attributive est complémenté par une complétive, laquelle exprime une prédication sans actant et ne donne donc lieu à aucune version attributive.

(32) Sam trouve qu'il pleut beaucoup à Paris.

Lorsque l'alternance complétive/attribut de l'objet est réalisable, des phrases telles que (13a) et (13b) présentent à première vue une relation de paraphrase. Nous récapitulons ici les exceptions à cette hypothèse trouvées dans la littérature.

Pour Ruwet (1982), l'exemple (33a) ne révèle aucun jugement de l'énonciateur sur la validité de l'opinion mentionnée, alors que (33b) suppose un jugement négatif de l'énonciateur.

(33) a. Paul croit qu'il est malade.
b. Paul se croit malade.

De la même façon, (Olsson, 1976) note que contrairement à (34a), dans (34b) le verbe est performatif : "il décrit toujours un acte de communication" mais "implique en outre que la personne accomplissant l'acte de communication est également responsable de l'état des choses, déterminé conventionnellement, qu'il 'rend public' " (Borkin, 1974, p. 89)

(34) a. Le juge a déclaré que l'accusé était innocent.
b. Le juge a déclaré l'accusé innocent.

Nous avons vu en 2.3.2 que les constructions à attribut de l'objet *adjectival* des verbes de perception étaient des constructions à *modifieur* attributif de l'objet. Les arguments qui suivent ont trait la potentielle relation de paraphrase entre constructions attributives *infinitives* de l'objet et constructions complétives correspondantes.

Guimier (1998) pointe un changement de sens entre construction à attribut de l'objet (sens purement perceptif) et avec complétive (sens plus constatif) pour les verbes de perception, et Willems & Defrancq (2000) examinent les différents sens possibles du verbe *voir* suivi d'une complétive ou d'une construction à attribut de l'objet et le fait que les différentes constructions donnent lieu à différentes interprétations.

(35) a. Je vois Jean travailler consciencieusement son latin.
b. Je vois que Jean travaille consciencieusement son latin.

Il apparaît donc que la relation de paraphrase entre complétive et construction à attribut de l'objet est possible mais pas systématique. Les différents contre-exemples exposés supra concernent des cas spécifiques (verbes performatifs et verbes de perception) ou reposent sur des lectures n'étant pas forcément universellement partagées ; nous considèrerons donc que la paraphrase vaut dans le cas général, mais ne tient qu'en première approximation dans certains cas.

4. Guimier (1998) dégage des propriétés sémantiques différentes pour les deux cas, le contenu de la relative étant asserté ce qui n'est pas le cas pour l'infinitive.

4 Exemples de traitement FrameNet

Notre objectif est de fournir un traitement linguistiquement motivé des CAO dans le cadre du projet ASFALDA. La méthodologie du projet est de partir de la modélisation en frames et du lexique existant pour l'anglais, et de proposer un équivalent pour le français (pour certains domaines notionnels ciblés) où les modifications sont justifiées soit par des différences anglais/français, soit par une "correction" d'incohérences manifestes.

Nous devons donc commencer par étudier ce qui a été fait dans FrameNet, pour des cas typiques de CAO en anglais (à noter donc que nous passons pour cette section à des données anglaises, ne cadrant pas forcément entièrement avec tout ce que nous avons vu supra sur les données du français). Nous avons cherché les frames associés à des verbes "typiques" de l'attribut de l'objet en anglais pour observer le traitement qui y serait fait de ces différents cas. Par "traitement", nous entendons ici le choix du nombre de rôles "core" définis pour les frames regroupant des déclencheurs qui permettent les deux constructions. En effet, FrameNet distingue les rôles "core" des rôles "non-core" (ci-après *essentiels* vs. *non-essentiels*), les premiers étant des arguments sémantiques des prédicats englobés sous un frame. *En général*, cette distinction relève grosso modo de la distinction syntaxique entre argument sous-catégorisé et modifieur. D'après la caractérisation sémantique des CAO faite ci-dessus, on s'attend plutôt à une modélisation par des frames avec deux rôles essentiels. Cependant, les critères syntaxiques sont parfois utilisés pour définir et typer les rôles, et Ruppenhofer *et al.* (2006, p. 20) indiquent en particulier qu'un objet direct est censé être representé comme un role essentiel. Il y a là une contradiction évidente, et cela se retrouve dans les données FrameNet.

Nous avons pu recenser 3 types de traitement, dont nous parlerons ici plus avant.

— les frames avec trois rôles essentiels (un pour le sujet, un pour l'objet direct et un pour l'attribut)
— les frames avec deux rôles essentiels (un pour le sujet et un pour le deuxième argument sémantique du verbe)
— les frames équipés de deux jeux de rôles essentiels utilisés distinctement pour annoter les deux types de constructions

On doit cependant préciser qu'étant donnée notre méthodologie, cette liste ne peut pas être exhaustive.

4.1 Frames avec trois rôles essentiels

Le verbe *consider*, qui peut s'employer avec une construction à attribut de l'objet ou avec une complétive [5], est un déclencheur de CATEGORIZATION. Dans ce frame, une entité sentiente (Cognizer) considère qu'un élément (Item) appartient à une certaine catégorie (Category). Les déclencheurs *consider*, *view*, *construe*, *understand* évoquent tous le frame CATEGORIZATION.

(36) *Whigs* *CONSIDERED* *his scruples over the oath* *ridiculous and inconsistent*
"Les Whigs considéraient ses scrupules à propos du serment comme ridicules et inconsistants."

Cette décision d'annotation pose le problème des verbes qui acceptent les constructions avec complétive aussi bien que les constructions à attribut de l'objet. Les données annotées FrameNet, annotations full-text comprises, ne comportent aucune occurrence du verbe *consider* avec complétive. Posons-nous donc hypothétiquement la question de l'annotation d'un exemple comme (37). Il apparaît qu'il faudrait choisir entre ne pas annoter ce type d'exemples comme évoquant le frame CATEGORIZATION et "descendre" dans la complétive pour y annoter les rôles Item et Category. Cela paraît difficile puisque FrameNet ne comporte absolument aucun exemple où les rôles seraient annotés au sein d'une complétive (lorsque le déclencheur n'appartient pas à cette complétive).

(37) *Whigs considered that his scruples over the oath were ridiculous and inconsistent.*
"Les Whigs considéraient que ses scrupules à propos du serment étaient ridicules et inconsistants."

Ce découpage en rôles pose deux autres problèmes. D'abord, il acte une relation sémantique entre déclencheur et Item qui, nous l'avons vu en section 3.2, n'existe pas : l'objet du verbe principal d'une construction à attribut de l'objet ne remplit aucune position argumentale de ce verbe. Ensuite, tout frame évoqué par le remplisseur du rôle Category aurait pour participant le remplisseur du rôle Item. Annoter ces rôles pour le frame Categorization apparaît comme redondant avec l'annotation de ce potentiel frame.

5. La version complétive de *consider* semble cependant bien moins usitée que sa traduction française.

4.2 Frames avec deux rôles essentiels

Nous avons étudié les frames liés aux verbes de connaissance, d'opinion et de croyance, tels que *know*, *think* et *believe*, qui s'emploient aussi bien avec une complétive qu'avec une construction à attribut de l'objet. Le frame AWARENESS présente des annotations des deux versions de cette alternance. Dans ce frame, une entité sentiente (Cognizer) a un certain contenu (Content) dans son modèle du monde.

Exemples d'annotation :

(38) *Some* *BELIEVED* *her husband was dead*
"Certains croyaient son mari mort."

(39) *He* *KNEW* *he had made a bad mistake*
"Il savait qu'il avait fait une grosse erreur"

(40) *I* *had THOUGHT* *that I was the only murderer in the family*
"Je pensais que j'étais le seul meurtrier de la famille."

Les occurrences de constructions à attribut de l'objet sont dans ce frame toutes annotées avec le même rôle : Content, qui englobe alors l'objet du verbe et son attribut.

(41) *You* *may sincerely BELIEVE* *yourself* *capable of running a nightclub*
"Tu te crois probablement sincèrement capable de gérer une boîte de nuit."

(42) *She* *KNEW* *herself* *to be mortally ill*
"Elle se savait mortellement malade."

(43) *Coffin* *THOUGHT* *her* *very clever*
"Coffin la pensait très intelligente."

Ce choix d'annoter de la même manière les deux versions de l'alternance semble cohérent puisqu'il fait correspondre un rôle à chaque argument sémantique et permet de rendre compte de la relation de paraphrase entre deux phrases comme les exemples (13a) et (13b).

En nous intéressant aux verbes de perception évoqués supra comme pouvant sous-catégoriser un attribut de l'objet, tels que *see*, *hear* et *feel*, nous avons découvert un cas problématique. Dans le frame PERCEPTION_EXPERIENCE, évoqué par les verbes susmentionnés, une entité sentiente (Perceiver_passive) perçoit sensoriellement un phénomène (Phenomenon). L'observation des annotations permet de trouver des exemples tels que (44) où, comme pour AWARENESS ci-dessus, un même rôle (Phenomenon) permet d'annoter l'objet et son attribut, mais aussi des exemples comme (45). Dans ces derniers, Depictive , un rôle non-essentiel du frame qui permet de spécifier l'état du phénomène, est utilisé pour annoter les occurrences d'attribut de l'objet.

(44) a. *She* *had never SEEN* *him* *so apoplectic* .
"Elle ne l'avait jamais vu si furieux"

b. *Three couples* *had already been SEEN* *leaving the hotel with baggage* .
"On avait déjà vu trois couples quitter l'hôtel avec leurs bagages."

(45) a. *She* *could SEE* *fly across his face* *a whole sequence of emotions and responses* .
"Elle voyait défiler sur son visage tout une séquence d'émotions et de réactions."

b. *I* *had SEEN* *him* *painfully sewing on a shirt-button* .
"Je l'avais vu cousant avec difficulté un bouton de chemise."

Nous n'avons pu identifier de constante qui permette de différencier les exemples annotés comme (44) de ceux annotés comme (45). Cette inconsistance apparaît donc comme un problème de cohérence dans la définition des rôles, à moins que l'on n'ait affaire, dans un cas ou dans l'autre, à des erreurs d'annotation.

4.3 Frames avec deux jeux complémentaires de rôles essentiels

Nous avons aussi examiné des verbes moins prototypiques de la construction à attribut de l'objet, mais qui fournissent un exemple de frame avec un troisième type de traitement FrameNet. Guimier (1999) ne cite pas les verbes de souvenir

se rappeler, *revoir* ni *se remémorer* comme verbes à attribut de l'objet, et on ne trouve dans le FTB aucun emploi de la construction avec ces verbes, mais FRANTEXT permet de trouver quelques exemples, comme en (46), et on peut aussi facilement en forger.

(46) a. *Il se rappelait la vieille grand'mère rentrant avec un filet à provisions.*
b. *Il revoyait son père lui montrant le pays et les grandes collines.*

En ce qui concerne les unités lexicales ayant trait à la mémoire, FrameNet fait le choix d'un découpage en frames selon l'axe *se souvenir d'un fait ou d'une information* (REMEMBERING_INFORMATION) vs. *se souvenir d'une expérience ancrée dans le temps* (REMEMBERING_EXPERIENCE). Seul ce deuxième frame permet une construction à attribut de l'objet. Cela est un peu compliqué par l'existence d'un frame MEMORY qui, au niveau de la définition, semble être un doublon de REMEMBERING_INFORMATION mais qui englobe les deux cas au niveau de l'annotation. Nous détaillons ici le découpage des rôles pour ces deux frames, bien que seul le premier présente des annotations.

Dans MEMORY, comme dans AWARENESS supra, les constructions à attribut de l'objet sont annotées avec un seul rôle Content :

(47) *He also REMEMBERED* *Dougal* *saying something about blackmail*.
"Il se rappelait aussi Dougal disant quelque chose à propos de chantage."

(48) *I REMEMBER* *dad* *running after me and having a long conversation trying to make me understand*.
"Je me rappelle papa me courant après et ayant une longue conversation pour essayer de me faire comprendre."

Le frame REMEMBERING_EXPERIENCE, quant à lui, permet d'annoter avec des rôles différents les versions avec complétive ou attribut de l'objet d'une même phrase. Dans ce frame, une entité sentiente (Cognizer) se remémore une expérience (Experience) passée, ou une propriété (Impression) passée d'une entité (Salient_entity), ou encore un état (State) dans lequel se trouvait cette même entité au cours d'une expérience passée implicite.

La définition du frame précise que la présence d'une entité exclut la mention explicite d'une expérience et appelle en général celle d'un état ou d'une impression caractérisant l'entité au moment remémoré par le Cognizer. Cela est répercuté au niveau des contraintes sur les rôles que FrameNet permet de définir, puisque le frame précise que Salient_entity , State et Impression sont en exclusion mutuelle avec Experience , et que l'annotation d'un State ou d'une Impression implique obligatoirement celle d'une Salient_entity .

Ce système de contraintes fournit donc deux jeux de rôles, ce qui devrait permettre d'annoter dans le même frame mais avec des rôles différents les constructions complétive et à attribut de l'objet de même sens. On ne trouve cependant aucune annotation de *remember* comme déclenchant REMEMBERING_EXPERIENCE dans les annotations FrameNet.

Exemples fabriqués :

(49) a. *I* *REMEMBER* *them* *taller*.
"Je me les rappelle plus grands."
b. *I* *REMEMBER* *that they were taller*.
"Je me rappelle qu'ils étaient plus grands."

(50) a. *I* *REMEMBER* *them* *singing on stage*.
"Je me les rappelle chantant sur scène."
b. *I* *REMEMBER* *that they were singing on stage*.
"Je me rappelle qu'ils chantaient sur scène."

5 Choix de traitement pour le FrameNet français

Après avoir observé les différents traitements appliqués par FrameNet au cas de discordance syntaxe-sémantique présenté par les constructions à attribut de l'objet, nous avons voulu prendre une décision que nous pourrions appliquer uniformément dans tous les frames du projet ASFALDA dont des déclencheurs autoriseraient cette construction. Cela évite que le choix ne revienne aux différents annotateurs à chaque occurrence du phénomène, ce qui permettra à la fois de leur faciliter la tâche et d'obtenir une meilleure cohérence des données.

Encoder qu'un même lexème déclenchera deux frames différents en fonction de la construction dans laquelle il s'insère entraînerait une polysémie systématique et, nous semble-t-il, injustifiée. Cela reviendrait à considérer que le verbe *trouver* possède un sens différent dans les exemples (13a) et (13b) que nous affichons à nouveau ici.

(13) *a. Sam trouve Alex sympathique.*
b. Sam trouve qu'Alex est sympathique.

De même, même si prévoir un jeu de rôles spécifique pour l'annotation de ce type de constructions nous permettrait d'encoder (13a) et (13b) sous le même frame, cette solution ne nous paraît pas non plus convenir. En effet, prévoir un rôle essentiel pour l'attribut de l'objet reviendrait à considérer ce dernier comme un argument sémantique du verbe déclencheur, ce qui n'est, nous l'avons vu plus haut, pas le cas.

Nous projetons donc d'annoter l'objet et son attribut au moyen d'un seul rôle, le même qui sera utilisé pour annoter une complétive de sens équivalent dans le même frame. Nous rendons ainsi compte de la similarité sémantique approximative entre les versions complétive et à attribut de l'objet d'une même phrase, et conservons un point de vue sémanticiste en prévoyant un rôle essentiel par rôle sémantique plutôt que de nous fonder sur la syntaxe. La relation de prédication entre l'objet et son attribut sera, le cas échéant, encodée par un frame déclenché par cet attribut.

6 Conclusion

Dans le cadre du projet ASFALDA (Candito *et al.*, 2014), qui a pour but de développer un FrameNet du français, nous nous sommes intéressés à la construction à attribut de l'objet comme exemple typique de divergence syntaxe-sémantique pour lequel nous souhaitions prévoir un traitement adapté et cohérent.

Pour ce faire, nous avons d'abord passé en revue les différentes catégories de constructions attributives de l'objet, ce qui nous a amené à préciser que notre étude se focaliserait sur les constructions à *complément* attributif de l'objet, et ce uniquement pour les verbes permettant une alternance entre cette construction et une construction complétive.

Nous nous sommes ensuite intéressés au statut de divergence syntaxe-sémantique des constructions à attribut de l'objet. Nous avons passé en revue les arguments qui peuvent amener à considérer que le verbe d'une CAO sous-catégorise trois compléments, puis ceux qui permettent de déterminer que seulement deux d'entre eux remplissent une position argumentale. Nous avons ensuite questionné l'apparente relation de paraphrase entre une CAO et la version à complétive d'une même phrase, qui apparaît possible mais pas systématique.

Après cet examen des propriétés syntaxiques et sémantiques des CAO, nous avons voulu étudier le traitement qui en était fait par FrameNet, en examinant les frames associés à des verbes "typiques" de l'attribut de l'objet en anglais. Nous avons mis au jour trois grands types de traitement, c'est-à-dire de choix des rôles essentiels définis pour les frames regroupant des déclencheurs qui permettent à la fois CAO et construction complétive. Il apparaît que selon les frames, on a à disposition deux ou trois rôles, ou encore deux jeux complémentaires de rôles pour annoter les arguments des déclencheurs, et que cette annotation n'est pas toujours consistante à l'intérieur d'un même frame.

Pour le FrameNet français, nous avons décidé d'homogénéiser la grande diversité de traitements des CAO rencontrée dans FrameNet en annotant au sein du même frame et au moyen d'un seul et même rôle l'objet et son attribut ou une complétive de même sens. Cela permettra de rendre compte de la similarité sémantique entre les deux constructions, et de conserver une équivalence entre le nombre de rôles essentiels et celui d'arguments sémantiques. Nous souhaitons ainsi privilégier un point de vue sémanticiste, dans l'esprit original de FrameNet.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a Treebank for French. In *Treebanks : Building and Using Parsed Corpora*, p. 165–188. Springer.

ABEILLÉ A., TOUSSENEL F. & CHÉDARAME M. (2004). *Corpus le monde, annotation en constituants, guide pour les correcteurs, version du 31 mars 2004*.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, p. 86–90, Stroudsburg, PA, USA : Association for Computational Linguistics.

BLANCHE-BENVENISTE C. (1991). Deux relations de solidarité utiles pour l'analyse de l'attribut. In M. GAULMYN, S. RÉMI-GIRAUD & L. BASSET, Eds., *À la recherche de l'attribut*, p. 83–97. Presses universitaires de Lyon.

H. C. BOAS, Ed. (2009). *Multilingual FrameNets in computational lexicography : methods and applications*. Trends in linguistics. Berlin, New York : Mouton de Gruyter.

BORKIN A. (1974). *Raising to Object Position*. PhD thesis, University of Michigan.

CANDITO M., AMSILI P., BARQUE L., BENAMARA F., DE CHALENDAR G., DJEMAA M., HAAS P., HUYGHE R., MATHIEU Y. Y., MULLER P., SAGOT B. & VIEU L. (2014). Developing a french framenet : Methodology and first results. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN*, Grenoble, France.

GROSS M. (1975). *Méthodes en syntaxe : régime des constructions complétives*. Actualités scientifiques et industrielles. Hermann.

GUIMIER É. (1998). Les constructions à prédicat de l'objet des verbes de perception. In *M. Forsgren, K. Jonasson & H. Kronning (éds), Prédication, assertion, information. Actes du colloque d'Uppsala, Uppsala : Acta Universitatis Upsaliensis*, p. 231–241.

GUIMIER É. (1999). *Les constructions à prédicat de l'objet en français : aspects syntaxiques, interprétatifs et formels*. PhD thesis, Université Paris VII-Denis Diderot.

HOEKSTRA T. (1992). Small clause theory. *Belgian Journal of Linguistics*, **7**(1), 125–151.

LE GOFFIC P. (1994). *Grammaire de la phrase française - Livre de l'élève - Edition 1994*. HU Langue française. Hachette Éducation.

LECLÈRE C. (2002). Organization of the lexicon-grammar of french verbs. *Lingvisticae Investigationes*, **25**(1), 29–48.

LEVIN B. (1993). *English Verb Classes and Alternations : A Preliminary Investigation*. University of Chicago Press.

MULLER C. (2000). Les constructions à adjectif attribut de l'objet, entre prédication seconde et complémentation verbale. *Langue française*, **127**(1), 21–35.

OLSSON K. (1976). *La Construction verbe + objet direct + complément prédicatif en français : aspects syntaxiques et sémantiques*. PhD thesis, Université de Stockholm.

RIEGEL M. (1981). Verbes essentiellement ou occasionnellement attributifs. *L'information grammaticale*, **10**(1), 23–27.

RIEGEL M. (1991). Pour ou contre la notion grammaticale d'attribut de l'objet : critères et arguments. In M. GAULMYN, S. RÉMI-GIRAUD & L. BASSET, Eds., *À la recherche de l'attribut*, p. 99–118. Presses universitaires de Lyon.

RIEGEL M., PELLAT J. & RIOUL R. (2009). *Grammaire méthodique du français*. Linguistique Nouvelle. Presses Universitaires de France.

ROTHSTEIN S. (1995). Small clauses and copular constructions. *Small clauses*, p. 27–48.

RUPPENHOFER J., ELLSWORTH M., PETRUCK M. R., JOHNSON C. R. & SCHEFFCZYK J. (2006). *FrameNet II : Extended Theory and Practice*. Berkeley, California : International Computer Science Institute. Distributed with the FrameNet data.

RUWET N. (1982). *Grammaire des insultes et autres études*. Travaux Linguistiques. Éditions du Seuil.

RÉMI-GIRAUD S. (1991). Adectif attribut et prédicat. approche notionnelle et morpho-syntaxique. In M. GAULMYN, S. RÉMI-GIRAUD & L. BASSET, Eds., *À la recherche de l'attribut*, p. 151–157. Presses universitaires de Lyon.

SEDDAH D., TSARFATY R., K'UBLER S., CANDITO M., CHOI J., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIORKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & VILLEMONTE DE LA CLÉRGERIE E. (2013). Overview of the spmrl 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proc. of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages : Shared Task*, Seattle, WA.

STOWELL T. (1995). Remarks on clause structure. *Syntax and semantics*, **28**, 271–286.

TOBBACK E. (2005). *Les constructions à attribut de l'objet et le marquage de la relation prédicative seconde*. PhD thesis, Université de Gand.

WILLEMS D. & DEFRANCQ B. (2000). L'attribut de l'objet et les verbes de perception. *Langue française*, p. 6–20.

WILLIAMS E. (1975). Small clauses in english. *Syntax and semantics*, **4**, 249–273.

Expressions différenciées des besoins informationnels en Langue Naturelle : construction de profils utilisateurs en fonction des tâches de recherche d'informations

Marilyne Latour[1, 2]
(1) Université Grenoble-Alpes, GRESEC, F-38040 Grenoble
(2) ReportLinker, 4 Rue Montrochet, 69002 Lyon
marilyne.latour@ac-grenoble.fr, marilyne.latour@reportlinker.com

Résumé. Devant des collections massives et hétérogènes de données, les systèmes de RI doivent désormais pouvoir appréhender des comportements d'utilisateurs aussi variés qu'imprévisibles. L'objectif de notre approche est d'évaluer la façon dont un utilisateur verbalise un besoin informationnel à travers un énoncé de type « expression libre » ; appelé langage naturel (LN). Pour cela, nous nous situons dans un contexte applicatif, à savoir des demandes de remboursement des utilisateurs d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli via ce moteur, les demandes en LN sur 5 années consécutives totalisant un corpus de 1398 demandes. Nous avons alors comparé l'expression en tant que tel du besoin informationnel en fonction de la tâche de recherche d'informations (RI) de l'utilisateur.

Abstract. With the massive and heterogeneous web document collections, IR system must analyze the behaviors of users which are unpredictable and varied. The approach described in this paper provides a description of the verbalizations of the information need in natural language. For this, we used data collected (*i.e.* users' complaints in natural language) through a search engine dedicated to economic reports in French over 5 consecutive years totaling a corpus of 1398 natural language requests. Then, we compared the expression as such of the information need according to the IR task.

Mots-clés : Recherche informations ; Besoin informationnel, Expression et interprétation des besoins ; Formulation question ; Langage naturel ; comportement utilisateur ; tâches de recherche d'informations.

Keywords: Information Retrieval ; Information Need, Query formulation and Query Expression ; Query Formulation ; Natural Language ; User Behavior ; IR task.

1 Introduction

Le 26 septembre 2013, *Google* annonce lors d'une conférence de presse pour fêter ses 15 ans, son nouvel algorithme baptisé *« Hummingbird »*. Ce nouvel algorithme s'éloigne de la logique de la recherche d'informations (RI) des requêtes en mots-clés pour s'ouvrir aux requêtes en langage naturel (LN). C'est un changement majeur pour le géant de la recherche dont l'objectif affiché est d'être capable de traiter des requêtes plus complexes et plus longues tout en prenant en compte le sens des mots dans leur contexte. Si l'on reprend l'exemple de Danny Sullivan [1] à la requête « Quel est l'endroit le plus près de chez moi pour acheter un *iPhone* 5S ? », *Google* devrait pouvoir prendre en compte et lier les notions « endroit près de chez moi », « acheter », « iPhone 5S » sous-entendu : l'objet désiré, le type d'action exercée sur l'objet (celui d'acheter) et la couverture géographique (à proximité de là où est localisé l'internaute « chez moi »). L'objectif est donc double : *(a)* comprendre la demande dans sa globalité pour lui donner un « sens » plus exact et *(b)* capitaliser d'autres informations que celles apparaissant sur les pages de recherche. Plus largement et au travers de la demande exprimée, c'est l'interprétation du besoin informationnel de l'utilisateur qui est visé pour le premier point. Ainsi, dans l'exemple cité, cela reviendrait à comprendre que la demande concerne le domaine de la téléphonie mais aussi que le besoin s'étend également à une volonté d'acheter un appareil. Le second point, lui, s'intéresse au lieu d'habitation si l'internaute a déjà renseigné cette information. Également des requêtes précédemment effectuées sur le moteur peuvent aider à « contextualiser » la demande notamment sur les centres d'intérêts de l'utilisateur. Cet algorithme rend compte d'une prise de conscience de la part des développeurs des systèmes de recherche d'informations (SRI) de traiter plus efficacement les requêtes avec une

1. « FAQ : All About The New Google « Hummingbird » Algorithm », Danny Sullivan, 27/09/2013, disponible sur : http ://searchengineland.com/google-hummingbird-172816 .

meilleure contextualisation du besoin informationnel. Dans ce sens, plusieurs pistes d'améliorations sont en cours : les premières ont pour objectif de mieux définir le besoin informationnel ; les buts des utilisateurs ainsi que les tâches sous-jacentes à la réalisation de celui-ci ; les secondes portent sur une meilleure contextualisation du contexte de l'utilisateur et de son environnement.

2 Le besoin informationnel contextualisé par la tâche de RI

Pour [Cabanac11], un utilisateur est un individu qui, dans un contexte donné -professionnel ou personnel- a besoin des résultats de sa requête médiatisé par un système informatisé -un logiciel quelconque, ou un système de recherche d'informations- pour réaliser une tâche avec un objectif spécifique. Dans ce cadre, le besoin informationnel est la prise de conscience d'un utilisateur lorsqu'il est confronté à l'exigence d'une information qui lui est à la fois déficiente et nécessaire. Ce besoin apparaît comme étant ancré, déterminé par la position qu'occupe un individu dans son environnement social [Lecoadic98] ou de travail. Or, de nombreuses études dont notamment celles de [Ingwersen05], [Ramirez06] expriment un décalage entre le besoin informationnel et son expression à travers un SRI. Ce décalage s'exprime généralement à travers des requêtes uniquement. Or ces requêtes sont généralement courtes (entre 2 et 3 termes) exprimant un but plus ou moins explicite [Strohmaier08]. Celles qui ont un but explicite sont des requêtes qui décrivent avec précision leur intention de recherche, *i.e.* pouvant être reliées à un but spécifique, de manière reconnaissable et non ambiguë. Exemple : *« acheter une voiture »*, *« réparer une voiture »*, *« aller à Miami »*, alors que celles qui ont un but implicitement exprimé sont des requêtes où il est difficile d'obtenir le but spécifique des intentions de recherche. L'exploitation du contexte de la tâche de recherche permettrait de mieux prédire le type de besoin des requêtes traduisant la nature de la tâche de recherche, en exploitant les caractéristiques morphologiques des requêtes ainsi que le profil et le contexte de la session.

Un des premiers constat est que pour faire face à ce décalage, il est nécessaire de travailler sur la compréhension de la tâche de RI. Pour ce faire, une variété d'approches en RI ont vu le jour se basant sur une taxonomie des buts de l'utilisateur ou de la tâche à effectuer. Le type de besoin inhérent à la requête est alors défini comme étant **informationnel** -lié à la recherche du contenu informationnel de documents-, **navigationnel** -lié à la recherche des sites d'accueil des personnes, organisations ou autres- ou **transactionnel** - lié à la recherche des services en ligne [Broder02], [Rose04], [Jansen08]. L'exploitation du contexte de la tâche de recherche permettrait de mieux prédire le type de besoin des requêtes traduisant la nature de la tâche de recherche, en exploitant les caractéristiques morphologiques des requêtes ainsi que le profil et le contexte de la session. Alors que la plupart des approches exploitent seulement des caractéristiques morphologiques de la requête [Kang03], [Kang05] ou des indicateurs de comportement de l'utilisateur (clics, données de consultation de la page, nombre de documents consultés...), l'objectif sous-jacent des moteurs qui proposent la LN comme moyen d'interrogation voire d'interaction : proposer un mode d'interrogation non contraignant pour l'utilisateur, afin qu'il puisse exprimer librement son besoin informationnel.

3 Expérimentation

Nous nous intéresserons ici à la phase de formulation du besoin informationnel d'un utilisateur de documents spécifiques à savoir des études de marchés économiques et ceci à travers sa formulation en LN. Plus précisément, nous voulons évaluer quelles sont les informations comme la zone géographique, la date, le type de données ou encore le prix, etc présentées dans la demande en LN en fonction de sa tâche de RI. L'objectif est, d'un point de vue linguistique, de mieux connaître les stratégies de l'utilisateur pendant sa tâche de RI ainsi que la formulation de son besoin informationnel. Le but intrinsèque est de pouvoir établir à long terme et en fonction des profils identifiés des recommandations pour les systèmes de recherche d'informations (SRI) sur les fonctionnalités à développer (aide à la recherche ou à la navigation, pré-enregistrement de filtres de recherche, environnement adapté en fonction du profil utilisateur, etc.). Nous nous situons dans un contexte applicatif spécifique, à savoir un moteur de recherche dédié à des études économiques en français : www.plusdetudes.comDétenu par la société Ubiquick, Lyon (France) avec un autre moteur de recherche en anglais : www.reportlinker.com. . Nous avons recueilli, les demandes en LN en français effectuées sur 5 années consécutives (de 2002 à 2007) de ce moteur de recherche, ainsi que les données utilisateurs (identité, fonction, domaine d'activité) totalisant un corpus de 1398 demandes en LN. Le fonds documentaire de ce moteur est d'environ 10 000 études de marché, couvrant 450 secteurs d'activités économiques et organisé autour de six axes principaux dans le thésaurus sectoriel : Agroalimentaire, Technologies de l'information et Médias, Biens et services de consommation, Sciences de la vie, Industrie, Services.

3.1 Recueil du corpus

Deux types de données ont été recueillis : *(i)* les données utilisateurs *via* le champs « Contacts » du formulaire SAV : ces données concernent l'identité de la personne, le nom de l'entreprise (ou université) de laquelle il ou elle dépend, la fonction exercée ainsi que les coordonnées téléphoniques, mail et adresse ; *(ii)* les demandes en LN *via* le champs « Expression de la demande » du formulaire SAV : ce formulaire est utilisé quand les recherches s'avèrent infructueuses ou tout simplement lorsque l' utilisateur veut obtenir un remboursement car juge être non satisfait des résultats obtenus. En analysant ces deux types de données, une typologie en fonction de trois types de tâche de RI se distingue nettement de notre corpus :

– [TACHE-CREA] : **le but de la tâche de RI est la création d'entreprise ou le lancement d'un nouveau produit ou encore d'une nouvelle marque** ; les objectifs opérationnels sont de se procurer une étude de faisabilité, d'identifier la concurrence éventuelle. Elle concerne 379 demandes de notre corpus (soit 27,11%).
– [TACHE-SCO] : **le but de la tâche de RI est la réalisation d'une tâche scolaire** ; les objectifs opérationnels sont de préparer un examen, d'écrire un mémoire, de travailler sur une étude de cas.... Elle concerne 513 demandes de notre corpus (soit 36,69%).
– [TACHE-PRO] : **le but de la tâche de RI est l'obtention d'informations dans un cadre professionnel** ; les objectifs opérationnels sont de mieux connaitre le marché, ses éléments chiffrés, d'identifier les tendances d'un marché ou d'un produit, de faire de la veille stratégique...Elle concerne 506 demandes de notre corpus (soit 36,19%).

Nous pouvons supposer que si ces caractéristiques sont fortement liées à un aspect spécifique des demandes et besoins informationnels et de manière relativement stable en fonction du contexte de la recherche, alors il serait possible de trouver des corrélations entre ces caractéristiques et les buts des utilisateurs dans le contexte considéré.

3.2 Chaîne d'analyse d'une demande en LN

Nous avons développé un environnement d'analyse chaîne de traitement en deux étapes des demandes en LN ainsi obtenus. La première concerne la segmentation des demandes en LN en blocs d'informations ; la seconde permet d'extraire des concepts associés (zone géographique, scope temporel, etc).

Première phase : Segmentation des demandes en LN en blocs d'informations

Nous nous concentrons dans cette phase de segmentation sur un scénario de recherche particulier *i.e.* la recherche de données économiques via un moteur de recherche en français. Basé sur l'analyse des formulaires SAV, on constate que la plupart de ces demandes comportent une structure sous-jacente. Par conséquent, un ensemble de règles ont été écrites manuellement pour décrire les différents scénario de la structure de la demande en LN. Cette dernière s'articule autour de plusieurs éléments, désignés comme autant de *blocs d'informations* :
– [SALUTATION-DEBUT] : formules de salutations (début). Exemple : « Bonjour »
– [FONCTION-CLIENT] : présentation de la fonction. Exemple : « je suis étudiante »
– [CONTEXTE] : annonce du contexte de la recherche. Exemple : « en vue de la création de ma future entreprise »
– [INTENTION-RECHERCHE] : introduction d'une intention de recherche. Exemple : « je souhaiterais obtenir des données »
– [TYPES-DONNEES] : indications des types de données recherchées. Exemple : « les parts de marchés de »
– [REFERENT] : définition du ou des référent(s) : *i.e.* la verbalisation de l'objet même de la recherche, ce sur quoi porte le but de la démarche de RI. Exemple : « revêtement de sol »
– [PRECISIONS] : ajouts de précisions. Exemple : « type Haagen Dazs »
– [SALUTATION-FIN] : formules de salutations (fin) : « avec mes remerciements »
Ces blocs d'informations ne sont pas tous remplis par les utilisateurs ; certains n'utilisant qu'un schéma simple de demande d'informations.

Dans un premier temps, notre méthodologie a consisté à extraire de manière empirique les caractéristiques et les structurations des blocs d'information d'un corpus d'apprentissage constitué d'environ 15 % de notre corpus total soit 200 demandes en LN. Il s'agit d'utiliser les régularités que manifeste notre corpus pour effectuer des découpages et des structurations. Pour cela, nous avons relevé manuellement :
– le **vocabulaire** et les formulations d'une demande pour les différents blocs d'informations en s'appuyant sur des marqueurs sémantiques (« par exemple », « aussi », « concrètement », « en effet »...) ;

– des **marqueurs typographiques** comme la ponctuation (virgule, le point virgule) et les majuscules).
– des **lexiques** spécifiques selon le différents blocs comme des lexiques de verbes (« vouloir », « désirer »…) ou des lexiques de noms (informations, données…).

Par exemple, sont répertoriés dans le même bloc : « je voudrais obtenir des informations sur […] », « je désire toutes les données concernant […] ».

Les règles de segmentation s'effectuent sur les phrases se basent sur le triptyque suivant : *(a)* une syntaxe et un ordonnancement de règles (écrites manuellement) *(b)* un analyseur morpho-syntaxique[2] pour faire appel à des étiquettes (nom commun, nom propre, adjectif, déterminant, etc.) *(c)* du lexique (principalement issues de thésaurus interne à la société possédant le moteur de recherche).

Une première évaluation manuelle de cet échantillon a montré que dans 75 % des cas le découpage en motifs obtenu était correcte (*i.e.* correspondait à un découpage qui aurait été fait humainement) ; 20% des séquences restaient non étiquetées (*i.e.* pas d'appartenance aux différents blocs d'informations identifiés) et 5 % des séquences étaient étiquetées de manière erronée. Sur les 20% de séquences non étiquetées, nous avons pu remarquer que *(i)* des erreurs de segmentation pouvaient être rectifiées en retravaillant sur les règles (12%), *(ii)* des erreurs étaient tout simplement dues à un mauvais ou absence d'étiquetage de l'analyseur morpho-syntaxique (8%). Nous avons travaillé sur ce premier point notamment en les ajustant par profils utilisateurs ce qui a permis de supprimer les 12% de séquences non étiquetées mais non de remédier aux erreurs d'étiquetage de l'analyseur morpho-syntaxique.

Une fois les règles de segmentation ajustées, nous avons ensuite réalisé ce découpage de façon automatique sur l'ensemble des demandes en LN soit 1398 demandes. Ceci nous permet d'avoir un usage de ces blocs différencié : le bloc [REFERENT] nous livre un bon nombre d'informations sur l'objet même de la recherche alors que les blocs [FONCTION-CLIENT] et [CONTEXTE] contiennent des informations sur le profil utilisateurs. Le bloc [INTENTION-RECHERCHE] pour savoir comment les utilisateurs verbalisent leurs besoins. Également les blocs [TYPES-DONNEES], [PRECISIONS] et [REFERENT] nous donnent accès à des concepts associés comme la zone géographique recherchée, le scope temporel demandé, les critères de prix s'ils sont présents, ou d'autres éléments de contexte de la recherche comme le délai/l'urgence ou non de la demande. Enfin les [SALUTATION-DEBUT] et [SALUTATION-FIN] ne sont pas étudiés dans ce présent travail ; notre objectif est donc uniquement de les identifier. Une comparaison plus fine peut alors s'effectuer sur le bloc [REFERENT] qui représente le thème de la recherche et la requête.

Deuxième phase : Extraction de concepts associés

En parallèle de la segmentation de la demande en blocs d'informations, sont extraits plusieurs concepts importants : la zone géographique, les noms de marque, les expressions de temps, les valeurs monétaires ou encore le caractère urgent (le délai). Ces concepts peuvent apparaître de façon disparates dans les différents blocs d'informations. Pour la zone géographique et les noms de marque, nous les repérons grâce à l'étiquette nom propre de l'analyseur morpho-syntaxique. Nous confrontons alors ces noms propres aux thésaurus internes (thésaurus géographique et thésaurus entités nommées). L'extraction des autres concepts se fait à base de règles que nous décrivons ci-dessous. Les concepts sont relevés à partir des règles suivantes :

– la zone géographique : étiquette nom propre et appel à un lexique issu du thésaurus géographique interne à la société. Ce thésaurus regroupe à la fois les noms propres des noms de pays (« France ») ainsi que leurs formes adjectivales (« français » dans « marché français »),
– les noms de marques et/ou d'entreprises : étiquette nom propre et appel à la fois à un lexique comportant le nom et les marques des plus grandes entreprises et également appel à des règles linguistiques pour détecter celles qui ne sont pas renseignées. Ces règles linguistiques se basent sur plusieurs éléments : l'apparition d'une majuscule qui ne soit pas au début d'une phrase, des mentions de forme juridique des entreprises comme SA, SARL, SCI pour les formes françaises et enfin sur la fonction exercée dans l'entreprise comme « gérant de », « propriétaire de », « secrétaire chez »,
– les expressions de temps : combinaison de règles et de lexiques permettant de recueillir toutes les formes de dates *e.g.* simple mention de l'année, ou mois + année, etc ainsi que des notions de temporalité comme le semestre par exemple,
– les valeurs monétaires : combinaison de règles et de lexiques permettant la reconnaissance d'entités numériques suivies ou précédées d'un signe monétaire,

2. L'analyseur morpho-syntaxique utilisé est *xelda* (développé par *xerox*). Les grandes étapes de cet analyseur sont : il *(i)* identifie tout d'abord la langue (à partir des premiers caractères), *(ii)* segmente en phrases, *(iii)* tokénine (*i.e.* scinde un texte en unités lexicales élémentaires), *(iv)* analyse morphologiquement (renvoie les catégories grammaticales potentielles pour tous les mots identifiés durant la tokénisation) et enfin *(v)* désambiguïse morpho-syntaxiquement en déterminant la catégorie grammaticale d'un mot en fonction de son contexte. Cet analyseur a été utilisé pour déterminer tous les traits morphologiques et syntaxiques.

– le caractère urgent : combinaison de règles basées sur du lexique.

Sur ce présent corpus, nous avons évalué manuellement les résultats obtenus sur 200 demandes en LN. Il s'avère que la reconnaissance géographique est assez pertinente : moins de 2 % de bruit (sur le corpus de test de 200, 68 demandes en LN contenaient une géographie). Les données numériques et les format dates sont également bien reconnus : moins d' un 1% de bruit (sur le corpus de test de 200, 15 demandes en LN contenaient une donnée numérique ou une date). La reconnaissance des noms de marque et/ou de noms d'entreprise s'est avérée plus fastidieuse puisque celle-ci s'élève à 7 % de bruit (sur le corpus de test de 200, 20 demandes en LN contenaient un nom de marque et/ou de nom d'entreprise).

4 Quelle expression des besoins informationnels selon la tâche de RI ?

La [TACHE-PRO] formule avec moins de termes ses besoins informationnels : en moyenne 17,66 termes contre 21,51 pour la [TACHE-CREA] et 22,27 pour la [TACHE-SCO]. Cette distribution est représentée dans la Figure 1. Les histogrammes de la figure 1 différencient la distribution de la [TACHE-PRO] (histogramme de gauche) ; les deux premières catégories représentent plus de 150 demandes en LN qui contiennent moins de 20 termes. Peu de demandes sont longues dans ce groupe de personnes.

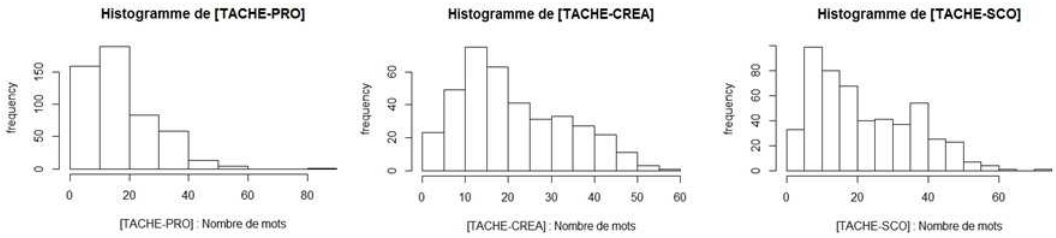


FIGURE 1 – Distribution du nombre de mots dans la demande en LN par type de tâche de RI, représentation en histogrammes

4.1 Traits morphologiques de la demande en LN différenciés par type de tâche en RI

En ce qui concerne les traits morphologiques, nous avons pu relever certaines différences entre la formulation des demandes en LN et la tâche de RI.

En ce qui concerne l'usage et la répartition des concepts à l'intérieur de la demande en LN, plusieurs caractéristiques se dégagent notamment à partir de la figure 2. Les deux concepts qui subissent le plus de variations entre les différents groupes d'utilisateurs sont la [FONCTION-CLIENT] et le [CONTEXTE]. En effet, le groupe [TACHE-SCO] utilise d'avantage le concept [FONCTION-CLIENT] pour indiquer qu'ils sont étudiants et éventuellement précise le degré d'étude (master, license) ou encore le nom de leur université. Le [CONTEXTE] est très peu utilisé par le groupe [TACHE-PRO] par rapport aux deux autres groupes d'utilisateurs : très peu d'informations supplémentaires sont données en plus du secteur recherché qui permettraient de mieux contextualiser leurs demandes avec des exemples ou en définissant plus précisément le cadre de leur recherche. Cette conclusion corrobore notamment le fait que le groupe tend à utiliser moins de termes pour formuler ses demandes en LN.

De façon moins prononcée, le concept [SALUTATION] est moins représenté par la [TACHE-PRO]. Cela suggère une démarche plus rapide et plus directe du formulaire SAV. Le groupe [TACHE-CREA] mentionne moins souvent que les deux autres groupes d'utilisateurs le type d'informations souhaité [TYPE-DONNÉES].

Nous relevons également d'autres écarts de comportements en fonction des types de tâche de RI sur les concepts de prix, de dates, de marques ou de géographies.

Ces concepts sont repris dans la Figure 3 à partir de laquelle nous observons que la géographie a une place plus importante pour le groupe d'utilisateurs ayant une [TACHE-CREA] ; ceci s'explique principalement par le fait que le développement d'une activité professionnelle est très liée à son emplacement géographique (*e.g.* « ouvrir un restaurant à Bordeaux »).

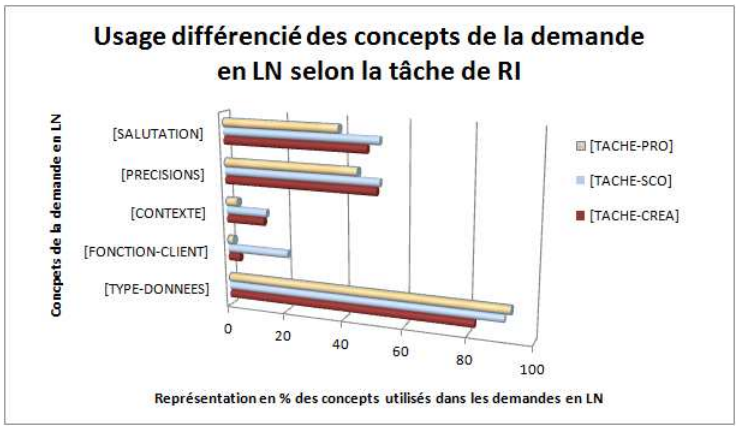


FIGURE 2 – Usage différencié des concepts de la demande en LN selon la tâche de RI

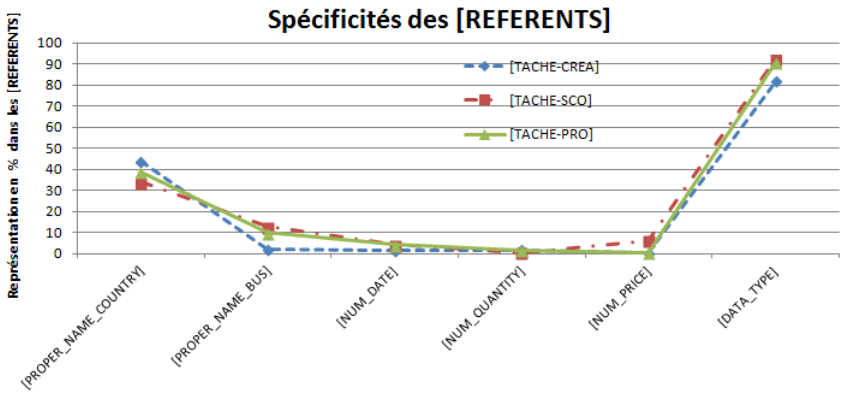


FIGURE 3 – Spécificités des (REFERENTS] dans les demandes en LN par type de tâche de RI

C'est le concept principalement développé dans les demandes en LN de ce groupe d'utilisateur : les concepts de dates, marques et prix apparaissent alors secondaires dans la façon de présenter leurs besoins. Le concept de prix est prédominant pour la [TACHE-SCO] qui demande une réduction voire la gratuité des études. Rappelons que la [TACHE-SCO] regroupe les utilisateurs qui ont une tâche scolaire à effectuer. Ce concept est moins présent pour la [TACHE-CREA] et inexistant pour la [TACHE-PRO]. Le concept de marque est également important pour la [TACHE-SCO] : les demandes comportent alors de nombreuses mentions de noms de marque ou de sociétés qui peuvent être l'objet même de leur travail (« réaliser une étude sur Coca-Cola ») ou sur un marché mais avec une demande particulière sur les principaux acteurs du secteur (*i.e* « les principaux leaders du marché des boissons énergétiques (CA de Isostar) »). Le concept de dates est utilisé de façon similaire par les groupes [TACHE-SCO] et [TACHE-PRO] ; ces utilisateurs ayant des contraintes de temps (délai) ou des demandes plus précises sur le scope temporel que doivent couvrir les études de marché. Les [TYPES-DONNEES] (non représentées dans le schéma) sont de 81,79% [TACHE-CREA], 92,39% pour la [TACHE-SCO] et 90,51% [TACHE-PRO].

4.2 Traits syntaxiques de la demande en LN différenciés par type de tâche en RI

Les principaux traits syntaxiques étudiés sont présentés dans la Figure 4. Notons tout d'abord que le groupe d'utilisateurs [TACHE-PRO] utilise davantage les tournures impersonnelles et moins de pronoms personnels (PP) que les deux autres groupes. Le groupe [TACHE-SCO] emploie plus souvent la première personne du pluriel, se situant dans une démarche de groupe (« nous effectuons un dossier sur »). Pour sa part, le groupe sur la [TACHE-CREA] utilise d'avantage la première personne du singulier, se situant dans une démarche plus personnelle (« je vais créer une entreprise »). Nous notons également que les utilisateurs ayant une [TACHE-SCO] s'expriment plus avec des phrases syntaxiquement correctes, notamment par rapport aux utilisateurs ayant une [TACHE-PRO] qui eux utilisent souvent une syntaxe incorrecte avec des

phrases partielles et/ou incomplètes (exemple : « étudier la croissance de Novartis »). Certaines de leurs demandes en LN peuvent s'apparenter d'ailleurs à des requêtes (exemple : « marché du parquet »).

Il semblerait que le groupe d'utilisateurs [TACHE-PRO] formule ses besoins en LN dans une structure moins formelle que les deux autres groupes ; leurs demandes en LN se rapprochent d'avantage à une formulation hybride entre une expression en LN et une requête.

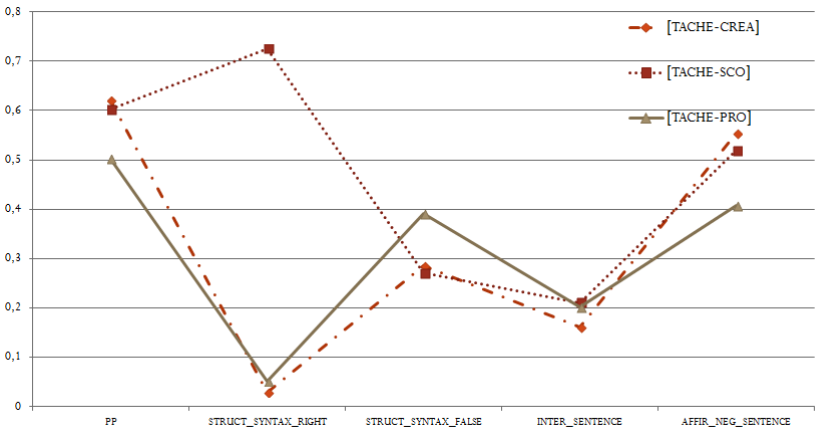


FIGURE 4 – Traits syntaxiques des demandes en LN différenciés par type de tâche de RI

4.3 Traits morpho-syntaxiques du [REFERENT] de la demande en LN selon les types d'utilisateurs

A partir du découpage en blocs d'informations, nous avons effectué un travail plus spécifique sur les référents. En effet, c'est dans ce bloc d'informations (qui peut contenir de 1 à 7 référents) que se retrouvent la plupart des éléments également formulées dans la requête du moteur de recherche. Le [REFERENT] est le concept porteur de l'information principale de la demande en LN ; il contient le thème de la recherche (secteur d'activité recherché). Afin de pouvoir être comparé plus finement avec la requête, ce bloc a fait l'objet d'une analyse qualitative avec une étude morpho-syntaxique de tous ses termes constitutifs.

Le nombre de [REFERENTS] dans les demandes en LN par type de tâche de RI : ce nombre est présenté dans la Figure 5. Cette figure représente la distribution du nombre de [REFERENTS] ; allant de $R = 0$ référent à $R = 7$ référents dans la demande en LN). Nous déduisons de cette figure que le groupe [TACHE-SCO], représenté en pointillé dans le schéma, formule davantage ses besoins informationnels avec un seul référent : 84,80% de leur demandes se verbalisent dans le [REFERENT-1] contre 67,28% pour le groupe [TACHE-CREA] et 65,61% pour le groupe [TACHE-PRO]. Les groupes d'utilisateurs [TACHE-CREA] et [TACHE-PRO] semblent avoir le même comportement par rapport au nombre de [REFERENTS] dans les demandes en LN.

La longueur des n-grammes des [REFERENTS] : cette longueur est présentée dans la Figure 6. La longueur des n-grammes dans les [REFERENTS] est de 1 terme ($n = 1$) dans 49,32% pour l'ensemble des [REFERENTS] du groupe [TACHE-SCO], 47,42% pour le groupe [TACHE-PRO] et de 38,92% pour la [TACHE-CREA]. Ce dernier groupe apparait en pointillé sur le graphique ; la distribution de la longueur des n-grammes est différente en donnant plus de poids aux formulations longues. Le $n = 0$ sur la Figure représente les 8 demandes en LN ne contenant pas de [REFERENT].

L'analyse morpho-syntaxique des [REFERENTS] : l'analyse morpho-syntaxique du bloc [REFERENT] de la demande en LN a été différenciée selon la tâche de RI des utilisateurs. Elle a relevé un usage différencié des catégories morpho-syntaxiques que nous présentons dans les Figures 7 et 8.

Il se dégage de ces graphiques certains traits caractéristiques récurrents par typologie d'utilisateurs :

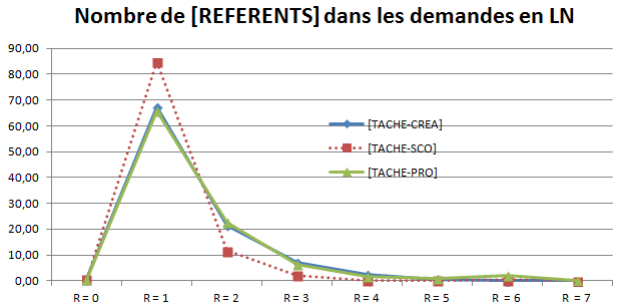


FIGURE 5 – Distribution du nombre de [REFERENTS] dans les demandes en LN selon le type de tâche de RI

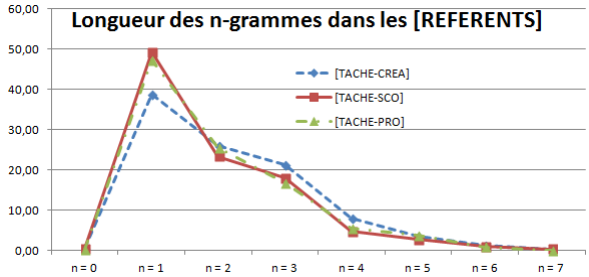


FIGURE 6 – Longueur des n-grammes dans les [REFERENTS] différenciée par type de tâche de RI

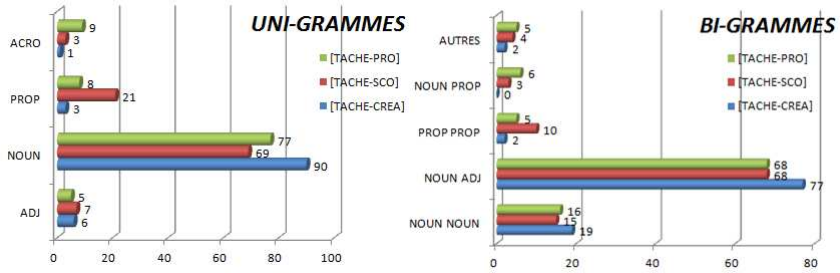


FIGURE 7 – Catégories morpho-syntaxiques pour les uni- et bi-grammes des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI (en pourcentages)

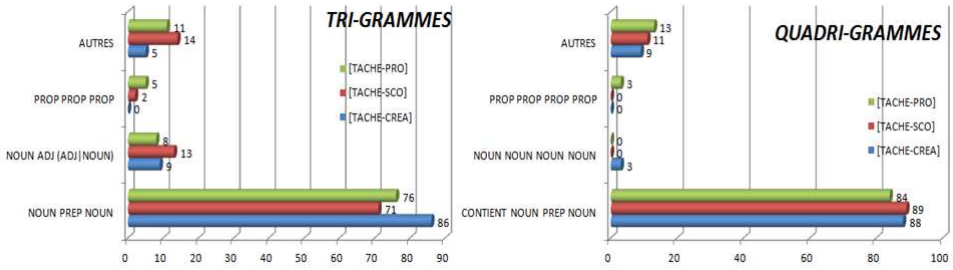


FIGURE 8 – Catégories morpho-syntaxiques pour les tri- et quadri-grammes des [REFERENTS] issus des demandes en LN et différenciés par type de tâche de RI

– la [TACHE-CREA] utilise des noms et principalement des noms au singulier (NOUN-SG) dans le cas d'uni-grammes ainsi que des noms propres (PROP) et très peu d'autres formes, des NOUN NOUN dans le cas de bi-grammes ainsi que des

NOUN ADJ, des NOUN PREP NOUN pour les tri- et quadri-grammes ainsi que des expressions composées exclusivement de NOUN,
– la [TACHE-SCO] : utilise davantage les noms propres (PROP) à la fois dans les uni- et bi-grammes, puis introduisent fréquemment une forme adjectivale pour formuler ses référents,
– la [TACHE-PRO] a davantage recours à des acronymes, ainsi qu'à des noms pour les uni-grammes, des NOUN PROP pour les bi-grammes, des PROP PROP PROP pour les tri-grammes ou encore PROP PROP PROP PROP pour les quadri-grammes.

4.4 Traits sémantiques de la demande en LN différenciés par type de tâche en RI

Nous avons mesuré plusieurs aspects dans les traits sémantiques de la demande en LN : *(a)* la valeur polysémique [POLYSEMY-VALUE] : nombre de fois que le terme ou les termes du [REFERENT] apparaissant dans les thésaurus, *(b)* la complexité linguistique [LINGUISTIC-COMPLEXITY] : profondeur des nœuds dans le thésaurus du [REFERENT], *(c)* l'évaluation de l'ambigüité de la tâche : calcul en fonction des traits sémantiques évaluées comme favorisant l'ambiguïté de la tâche, *(d)* les secteurs d'activité mentionnés dans les [REFERENTS] : nombre de correspondances avec les thésaurus internes de l'entreprise.

(a) La valeur polysémique [POLYSEMY-VALUE] : cette valeur est calculée en fonction du nombre de fois que les termes apparaissent sous leurs formes lemmatisées dans les thésaurus internes de la société. Plus un terme apparait dans les thésaurus, plus sa valeur polysémique est élevée car non spécifique à un secteur donné particulier. Ainsi, la *polysemy value* à 0 indique que le terme n'est mentionné qu'une seule fois dans les différents thésaurus ; il est donc peu polysémique. Les *polysemy values* 1 à 3 sont présentées dans le Tableau 1. Les termes dont la *polysemy value* vaut $k(k \geq 0)$ apparaissent $k+1$ fois sauf si $k=3$ où le terme apparait au moins $k+1$ fois. La lemmatisation rend possible les correspondances quelle que soit la flexion ; la correspondance se fait sur le lemme et non sur sa forme. Un inconvénient est que cela peut entrainer des rapprochements non satisfaisants, particulièrement dans les cas des uni-termes.

Les *[unknown]* sont les termes qui ne sont pas présents dans les thésaurus. Un *token* inconnu désigne une entité (ou unité) lexicale qui n'a pas été reconnue lors de l'analyse et les comparaisons avec les termes des thésaurus. Cette information peut soit indiquer que le terme a été mal orthographié faussant l'analyse avec la correspondance des termes issus des thésaurus, soit que le terme n'est pas renseigné dans les thésaurus par manque de précision ou de recouvrement d'un secteur d'activité. Notons que sigles et noms des plus grandes marques ou entreprises peuvent être reconnus si ceux-ci correspondent à des entrées dans le thésaurus. La valeur polysémique, représentée dans le Tableau 1, permet d'avoir un aperçu de la polysémie relative à chaque groupe d'utilisateurs selon la tâche de RI.

[POLYSEMY-VALUE]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
0	62,77%	66,60%	78,10%
1	6,82%	7,91%	6,07%
2	2,34%	2,57%	2,11%
3	5,65%	7,11%	3,43%
UNKNOWN	22,42%	15,81%	10,29%

TABLE 1 – Valeur polysémique du [REFERENT] de la demande en LN différenciée par type de tâche de RI

Le groupe dont les termes utilisés dans les [REFERENTS] en LN est le moins porteur de polysémie est celui de la [TACHE-PRO] avec un taux de 0 égal à 78,10%. Il a aussi peu de valeurs fortement polysémiques au niveau 3 (3,43%). La [TACHE-SCO] a une valeur polysémique assez élevée en ce qui concerne le niveau 3 avec 7,11%. Les valeurs UNKNOWN sont importantes pour le groupe de la [TACHE-CREA] puisqu'ils sont à 22,42%.

(b) La complexité linguistique [LINGUISTIC-COMPLEXITY] : la complexité linguistique correspond à la profondeur des nœuds dans le thésaurus du terme ou de l'expression du [REFERENT]. Le [LEVEL 1] correspond aux catégories supérieures généralistes dans la hiérarchie du thésaurus sectoriel : Agro-alimentaire, Biens et Services de Consommation, Industrie lourde, Technologies de l'information et Médias, Sciences de la Vie et Services. Les niveaux suivants de [LEVEL 2] [...] [LEVEL 5] sont des niveaux du thésaurus hiérarchiquement descendants. Les termes sont plus spécifiques. Un même terme peut apparaître dans plusieurs branches de la hiérarchie. La distribution des [REFERENTS] par niveaux et par

type de tâche de RI est présentée dans le Tableau 2 ; l'usage du thésaurus donne un aperçu de l'utilisation de la profondeur des nœuds et donc de la spécificité de la recherche.

[LINGUISTIC-COMPLEXITY]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
1	2,53%	1,98%	1,06%
2	10,53%	16,40%	9,76%
3	40,74%	35,18%	41,69%
4	19,30%	22,92%	28,76%
5	4,48%	7,71%	8,44%
UNKNOWN	22,42%	15,81%	10,29%

TABLE 2 – Complexité linguistique du [REFERENT] de la demande en LN différenciée par type de tâche de RI

D'après le Tableau 2, nous pouvons relever que le groupe d'utilisateurs [TACHE-PRO] se distingue particulièrement aux niveaux 4 et 5 ; ils utilisent des termes qui sont plus spécifiques et plus fins que les deux autres groupes. *A contrario*, le groupe [TACHE-SCO] a une demande plus importante au niveau 2 du thésaurus, représentant des domaines plus généraux. Les nombres UNKOWN sont bien sur identiques à la dernière ligne du Tableau 1 de la [POLYSEMY-VALUE] ; il est difficile de dégager des tendances générales pour le groupe [TACHE-CREA]. Ce chiffre élevé de UNKOWN pour la [TACHE-CREA] peut toutefois révéler que les termes employés notamment les entités nommées de noms de marques, d'entreprises ou géographiques sont trop spécifiques (noms de petites entreprises ou noms géographiques d'une petite commune) pour figurer dans les ressources utilisées pour faire les comparaisons.

(c) L'ambiguïté de la tâche : nous avons défini un modèle reprenant certaines valeurs comme la polysémie ou encore la profondeur des nœuds dans le thésaurus afin de rendre compte de l'ambiguïté de la tâche par groupe d'utilisateurs.

Cette ambiguïté se calcule par la somme de fonctions de chacun des traits sémantiques présentés dans la Figure 3 : dans la moitié supérieure du tableau, les catégories font croître la complexité sémantique de la demande en LN. Dans la moitié inférieure du tableau, les catégories font au contraire décroître cette même complexité : plus elles sont renseignées moins la demande en LN est ambiguë pour les SRI, c'est-à-dire plus sa compréhension est aisée.

$$ambiguité(tâche) = \sum_{i=1}^{n} f_i(tâche), \tag{1}$$

où n est le nombre de traits, f_i est une valeur liée à la tâche et qui pondérera positivement ou négativement un de ses traits. La forme précise de chaque f_i est donnée dans le Tableau 4.

trait$_i$	f_i
[ACRONYM]	nombre d'[ACRONYM]
[UNKNOWN]	nombre de [UNKNOWN]
[POLYSEMY VALUE]	valeur de la [POLYSEMY VALUE]
[FOREIGN]	nombre de termes en langue étrangère
[STRUCT-SYNTAX-FALSE]	nombre de demandes avec une structure syntaxique fausse
[LINGUISTIC-COMPLEXITY]	valeur de [LINGUISTIC-COMPLEXITY]
[CONTEXT-PRECISIONS]	nombre de [CONTEXT-PRECISIONS]
[SYNT-DEPTH]	nombre de n-grammes [SYNT-DEPTH]
[TYPE-DONNEES]	nombre de [TYPE-DONNEES]
[PROPER-NAME-COUNTRY]	valeur de la [PROPER-NAME-COUNTRY]
[NUM-PRICE]	nombre de [NUM-PRICE]
[NUM-DATE]	nombre de [NUM-DATE]

TABLE 3 – Ambiguité de la tâche de RI en fonction des traits sémantiques

Il en ressort que les groupes [TACHE-SCO] et [TACHE-PRO] ont une valeur quasiment équivalente à 5.7 alors que le groupe [TACHE-CREA] a une valeur a 4.6. Il semblerait donc que l'ambiguïté de la tâche est moins importante pour le groupe

[TACHE-CREA]. Ce groupe doit en effet gagner en précision puisque la tâche est souvent assez claire : développer une activité, ouvrir un magasin, etc. La variance est un peu plus importante pour le groupe [TACHE-SCO] ; certains utilisateurs au sein de ce groupe ont une tâche plus ambiguë que d'autres.

[AMBIGUITY-TACHE]	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
mean (moyenne)	4.604222	5.758285	5.693676
var (variance)	11.16041	12.41412	11.97133
SD (Ecart-Type)	3.340719	3.523367	3.45996

TABLE 4 – Ambiguité de la demande en LN en fonction de la tâche de RI des groupes utilisateurs

A partir de l'estimation de l'ambiguïté de la tâche, nous avons voulu tester si la longueur ([LENGTH] décrite à la page 5) de la demande en LN était révélatrice de cette ambiguïté. On peut effet émettre l'hypothèse que si une tâche est ambiguë (dans le sens difficile à expliciter) l'utilisateur aura tendance à utiliser plus de termes pour l'exprimer. Nous avons réalisé deux mesures pour tester cette hypothèse. Pour la première, nous avons utilisé le coefficient de corrélation de Pearson, indice statistique qui exprime l'intensité et le sens (positif ou négatif) de la relation linéaire entre deux variables quantitatives. Ce coefficient de corrélation a déjà été utilisé dans d'autres études équivalentes notamment par [Mothe2005]. Ces mesures ont été obtenues à l'aide du logiciel R. L'importance de la valeur de corrélation est exprimée par la valeur p associée . $P - valeur$ est une estimation de la probabilité que les résultats aussi extrême ou plus extrême se produisent par hasard. Un $p - valeur$ proche de 0 indique une grande confiance dans la corrélation , tandis qu'une $p - valeur$ proche de 1 indique une forte chance pour l'indépendance entre les variables . Nous retiendrons comme significatives les corrélations dont la $p - valeur$ est inférieure à $0,05$. Les résultats sont donnés dans le Tableau 5. Ils indiquent que les [TACHE-SCO] et [TACHE-PRO] ont une $p - valeur$ inférieure à 0,05 et ont donc des résultats significatifs : la longueur de la demande en LN est liée à l'ambiguïté de la tâche. Cette hypothèse est moindre pour la [TACHE-CREA] puisque sa $p - valeur$ est 0,0501.

	[TACHE-CREA]	[TACHE-SCO]	[TACHE-PRO]
Corrélation de Pearson	-0,2067269	-0,2121073	-0,1681577
P-value	0,0501	0,01249	0,0001445

TABLE 5 – Ambiguité de la tâche de RI - Coéfficient de corrélation de Pearson

(d) Les secteurs d'activité mentionnées dans les [REFERENTS] : certains secteurs mentionnés dans les [REFERENTS] des demandes en LN sont prédominants selon le type de tâche de RI. Ainsi les utilisateurs du groupe [TACHE-SCO] recherchent davantage d'informations sur l'alimentation et le textile ; le groupe [TACHE-PRO] sur la construction et BTP ainsi que sur la chimie et les produits chimiques tandis que les utilisateurs de la [TACHE-CREA] recherchent davantage sur les secteurs liés aux loisirs et à la restauration. Ces résultats sont présentés dans la Figure 9.

5 Conclusions et Perspectives

Nous avons recueilli et analysé les besoins informationnels exprimés en LN, ceci dans un contexte bien particulier ; celui d'une demande de remboursement effectuée par des utilisateurs d'un moteur de recherche après des recherches a priori infructueuses. Nous nous avons ensuite établi des règles linguistiques et une analyse morpho-syntaxique qui nous a permis de schématiser l'énoncé en LN en fonction de la tâche de RI. Grâce à notre corpus très spécifique, nous avons observé des régularités sur la construction des demandes en LN en fonction de la tâche de RI et nous pouvons conclure qu'en fonction du profil et du type de tâche à réaliser, l'interface de navigation ainsi que les résultats à proposer à l'utilisateur doivent être différenciés par les SRI. Ainsi la [TACHE-SCO] mentionne davantage des critères de prix dans leurs demandes, la [TACHE-CREA] pour leur part indique davantage la zone géographique de leur recherche. La [TACHE-PRO] emploie plus facilement des tournures impersonnelles pour formaliser leurs demandes, etc. Ce sont autant d'indices (morphologiques, syntaxiques, sémantiques) qui peuvent aider à construire le profil utilisateur et à proposer des résultats plus spécifiques en fonction de la tâche de RI et des profils utilisateurs ainsi identifiés. Des améliorations sont envisagées notamment de reconnaissance d'entités nommées dans les demandes en LN ; un outil de reconnaissance et d'extraction d'entités nommées pourrait éventuellement améliorer les résultats ainsi obtenus.

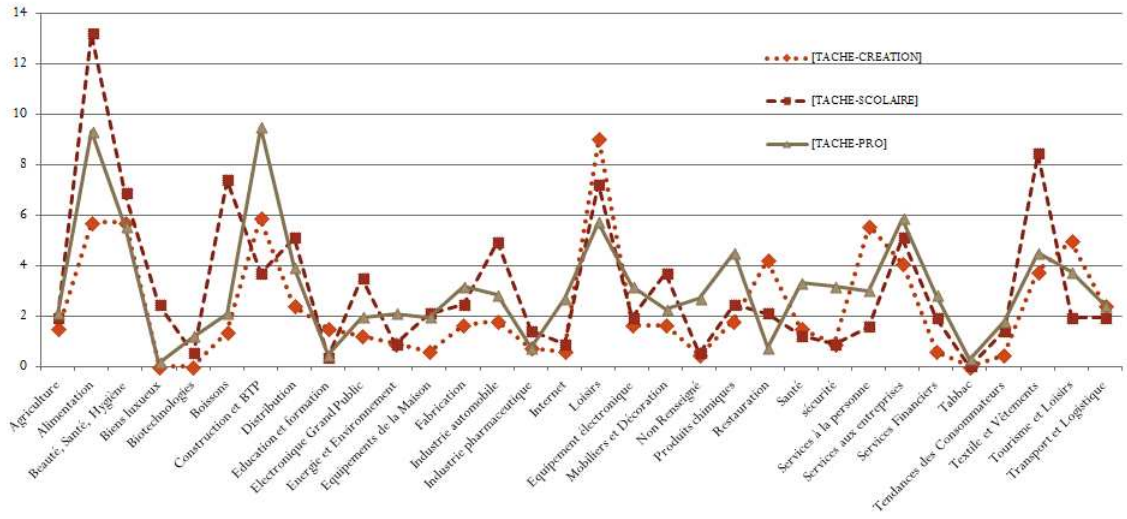


FIGURE 9 – Traits sémantiques des demandes en LN différenciés par type de tâche de RI

Références

[Broder02] BRODER Andrei. *A taxonomy of web search*. In : SIGIR FORUM, 2002, vol.36, n.2, pp.3-10.

[Cabanac11] CABANAC Guillaume, CHEVALIER Max, CIACCIA A. et al. *Recherche d'information et modélisation usagers*. In P. Bellot (Ed.) Recherche d'information contextuelle, assistée et personnalisée. Paris : Hermès, 2011.

[Ingwersen05] INGWERSEN Peter, JARVELIN Kalervo. *The Turn. Integration of information seeking and retrieval in context*. Dordrecht : Springer, 2005, 448 p.

[Jansen08] JANSEN Bernard J.,BOOTH L. Danielle, SPINK Amanda. *Determining the informational, navigational and transactional intent of Web queries*, Information Processing and Management, 2008, vol. 44, pp. 1251-1266.

[Kang05] KANG In-Ho. *Transactional query identification in web search*. In AIRS'05 : Proceedings Information Retrieval Technology, Second Asia a Information Retrieval Symposium, Jeju Island, Korea, 2005, pp. 221-232.

[Kang03] KANG In-Ho, KIM GilChang. *Query Type Classification for Web Document Retrieval*. In : Proceeding SIGIR '03. New York : ACM, 2003, pp.64-71.

[Lecoadic98] LE COADIC Yves-François. *Le besoin d'information : formulation, négociation, diagnostic*. Paris : ADBS Editions, 1998, 191 p.

[Mothe05] MOTHE Josiane, TANGUY Ludovic. *Linguistics features to predict query difficulty - A case study on previous TRES campaign*. In : ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications

[Ramirez06] RAMIREZ Goergina , DE VRIES Arjen P.. *Relevant contextual features in XML retrieval*. In : Proceedings of the 1st international conference on Information Interaction in Context. New York : ACM, 2006, pp. 95-110.

[Rose04] ROSE Daniel, LEVINSON Danny. *Understanding User Goals in Web Search*. In : Proceeding WWW '04 Proceedings of the 13th international conference on World Wide Web. New York : ACM, 2004, pp.13-19.

[Strohmaier08] STROHMAIER Markus, PRETTENHOFER Peter, LUX Mathias. *Different Degrees of Explicitness in Intentional Artifacts : Studying User Goals in a Large Search Query Log*. In CSKGOI'08, 2008, [10 p.]

Les modèles de description du verbe dans les travaux de Linguistique, Terminologie et TAL

Ornella Wandji Tchami
CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
ornella.wandjitchami@univ-lille3.fr

Résumé. Dans le cadre de notre projet de recherche, qui a pour but l'implémentation d'un outil de simplification des emplois spécialisés de verbes dans des corpus médicaux à partir de l'analyse syntaxico-sémantique de ces verbes en contexte, nous proposons une analyse de quelques approches et travaux qui ont pour objet principal la description du verbe dans les trois domaines de recherche à l'interface desquels se situe notre projet : linguistique, TAL et terminologie. Nous décrivons plus particulièrement les travaux qui peuvent avoir une incidence sur notre étude. Cet état de l'art nous permet de mieux connaître le cadre théorique dans lequel s'intègre notre projet de recherche et d'avoir les repères et références susceptibles de contribuer à sa réalisation.

Abstract. As part of our research project, which aims to implement a text simplification tool for the specialized usages of verbs in medical corpora using the syntactic and semantic analysis of these verbs in context, we propose an overview of some approaches and work whose main research object is the description of verbs, within the three research areas which interface our study is : linguistics, terminology and NLP. We pay a particular attention to studies that can have an impact on our work. This state of the art allows us to better understand the theoretical framework related to our research project. Moreover, it allows us to have benchmarks and references that might be usefull for the realization of our project.

Mots-clés : Verbe terminologique ou spécialisé, sémantique des cadres, sémantique lexicale, structure argumentale, étiquetage en rôles sémantiques.

Keywords: Specialized verb, Frame Semantics, lexical semantics, argumental structure, Semantic Role Labeling.

1 Introduction

Contexte général

L'intérêt porté au verbe change selon que l'on se situe dans le domaine de la linguistique, de la terminologie ou celui du Traitement Automatique des Langues (TAL). En effet, selon leurs objectifs respectifs, chacune de ces disciplines octroie au verbe une place différente reconnaissable à travers l'importance qui lui est donnée dans les diverses études et les différents cadres théoriques propres au domaine concerné. Quelles sont les éclairages proposés par les différentes approches (linguistique, terminologie et TAL) qui prennent le verbe comme objet d'étude ? Comment-est ce que le verbe est abordé dans ces travaux ? Est-il traité au même titre que les autres catégories grammaticales en l'occurrence le nom ? En quoi est-ce que le verbe et sa structure argumentale peuvent-ils être utiles en vue de la simplification des textes spécialisés ? Telles sont les questions auxquelles nous allons essayer de répondre dans ce travail qui a pour objectif de dresser un état de l'art de différents modèles de descriptions du verbe dans les trois disciplines concernées.

Travail envisagé

Le projet que nous entreprenons a pour objectif de proposer une méthode de simplification de textes médicaux, à partir d'une analyse syntaxico-sémantique des verbes en contexte. Au terme de ce travail, nous souhaitons implémenter un outil de simplification des textes écrits en français (et éventuellement en anglais), spécialisés en cardiologie (ou en d'autres domaines médicaux). L'outil devra repérer les emplois verbaux peu communs au discours des patients et devra ensuite proposer des emplois sémantiquement similaires, mais plus adaptés au niveau de spécialisation de ces utilisateurs. La méthode proposée est basée sur différentes hypothèses. En effet, nous pensons que le prédicat verbal peut être un excellent

point de départ pour cerner la sémantique des textes spécialisés puisqu'il sert à exprimer l'expertise portée par les mots qui l'entourent dans la phrase (L'Homme & Bodson, 1997). Par conséquent, nous considérons la structure argumentale du verbe comme une importante source d'informations sur les propriétés sémantique et syntaxique du verbe.

Ce projet de recherche s'inscrit dans le cadre de la simplification des textes spécialisés. Il s'agit d'une tâche du TAL qui consiste à cibler et à simplifier automatiquement les éléments, qui empêchent la compréhension aisée d'un texte, afin de faciliter l'accès au contenu de ce texte. Les travaux existants se focalisent sur la simplification syntaxique (Brouwers *et al.*, 2014), la simplification lexicale (Elhadad, 2006; Leroy *et al.*, 2012), la combinaison des fonctions lexicales, grammaticales, syntaxiques et discursives (Heilman *et al.*, 2007, 2008; Pitler & Nenkova, 2008), les caractéristiques de surface des textes (nombre de caractères et syllabes par mot), la capitalisation, la ponctuation et les ellipses (Tapas & Orr, 2009), ou la modélisation statistique de la langue (Thompson & Callan, 2004). L'approche que nous proposons se situe à mi-chemin entre la simplification lexicale et la simplification syntaxique et vise à réduire les difficultés de compréhension des textes médicaux fortement spécialisés à travers la simplification des constructions verbales. À notre connaissance, il n'existe pas de travaux en simplification de textes autant orientés sur l'analyse des verbes et de leurs argumentales. Une étude comparative du fonctionnement des verbes dans des textes de corpus médicaux rédigés par des experts et des non-experts en médécine a permis d'observer que les verbes ont tendance à s'entourer d'arguments fortement spécialisés dans les écrits des experts, rendant parfois leur compréhension difficile pour les non-experts (Wandji Tchami *et al.*, 2013). Notre travail de recherche vient donner une suite à cette observation. L'objectif principal étant d'améliorer certains aspects de la méthode (l'annotation automatique des arguments, l'analyse des verbes) et de la développer davantage, en y intégrant un travail de simplification.

Le travail présenté ici a pour objectif de nous aider à mieux cerner le cadre théorique dans lequel s'inscrit le projet de recherche que nous envisageons de réaliser, en nous donnant une idée précise des travaux et outils existants, centrés sur le verbe et susceptibles de nous aider pour réalisation du travail envisagé. Il est organisé autour de 4 grandes parties dont les trois premières sont consacrées à l'exploration des travaux portant sur le prédicat verbal, respectivement en linguistique (section 2), terminologie (section 3) et TAL (section 4). Dans la dernière partie (section 5), nous faisons une discussion de l'impact que les travaux de l'état de l'art peuvent avoir sur la réalisation de notre projet de recherche et nous abordons les perspectives de travail que nous envisageons d'explorer.

2 Les approches linguistiques dédiées au verbe

En linguistique, de nombreux cadres théoriques placent le verbe au coeur de leurs travaux. Nous nous attardons plus particulièrement sur les cadres théoriques qui s'intéressent au verbe en tant qu'élément régisseur (c'est-à-dire un élément dont la réalisation syntaxique et sémantique dépend grandement de la présence d'autres constituants qui lui sont subordonnés) et décrivent son rapport avec les autres constituants de la phrase. Il existe diverses approches de description du verbe, mais nous ne sommes pas en mesure de fournir une présentation exhaustive de toutes les approches théoriques existantes. Nous nous limitons à celles qui servent de bases à la réalisation de différentes tâches du TAL (sections 4.2 et 4.1), et à la conception des ressources (section 4.3).

2.1 Le nœud verbal au coeur de la syntaxe structurale

La syntaxe structurale (Tesnière, 1959) est la première théorie à avoir mis le verbe au centre de la phrase. En syntaxe structurale, l'ensemble des mots d'une phrase constitue une véritable hiérarchie au sein de laquelle les constituants sont liés les uns aux autres par des liens de dépendance. La phrase, encore appelée *stemma*, est décrite comme étant un schéma arborescent, ou un ensemble de nœuds. Le nœud quant à lui désigne un ensemble constitué d'un régissant et de tous ses subordonnés. Dans cette configuration, le nœud central correspond en général au noeud verbal. Le verbe, étant au centre du nœud verbal, est par conséquent au cœur de la phrase. Il est pour ainsi dire le régissant de toute la phrase. La notion de nœud verbal est définie à travers une métaphore du drame : *« le nœuds verbale ... exprime un tout petit drame. Comme un drame, ... il comporte obligatoirement un procès et plus souvent des acteurs et des circonstants »*. C'est dans cette optique que cette approche postule l'existence des actants ou participants au procès verbal (Tesnière, 1959). L'ensemble des actants d'un verbe constitue sa structure actancielle. Un verbe peut avoir zéro, un ou plusieurs actants, comme le montre l'exemple suivant : *Alfred donne le livre à Charles*. Dans cette phrase, le verbe *donner* a trois actants : *Alfred*, *le livre* et *Charles*. Chaque actant joue un rôle bien déterminé dans le procès verbal.

2.2 La théorie des cadres sémantiques

Encore appelée *Frame semantics*, la sémantique des cadres est une approche qui remonte aux années 1980. Elle est une extension de la grammaire des Cas (Fillmore, 1968), qui évoquait déjà l'existence des rôles sémantiques (agent, lieu, etc.) dans la structure syntaxique profonde du verbe. La sémantique des cadres (Fillmore, 1982) vise à l'origine à faciliter la compréhension des textes. Son principal objectif est de décrire la syntaxe et la sémantique des unités lexicales (noms, adjectifs, verbes). L'idée principale de Fillmore est que le sens d'un mot ne peut être interprété que si l'on a accès aux informations (linguistiques, extralinguistiques ou encyclopédiques) essentielles faisant référence à ce mot. Ces informations peuvent être accessibles grâce à un frame ou cadre au sein duquel les unités lexicales sont organisées. Le cadre est défini comme un scénario, un schéma ou une structure conceptuelle qui sous-tend l'utilisation d'un item lexical ainsi que son interprétation (Fontenelle, 2009). Il décrit une situation particulière ainsi que les participants *Frame elements (FE)* qui peuvent être obligatoires (*core elements*) ou facultatifs (*non core elements*). Un cadre est évoqué par une unité lexicale *(LU)*. Par exemple, le frame de la transaction commerciale (Fillmore, 1976) a plusieurs unités évocatrices : *acheter*, *vendre*, *payer*, *recupérer* et plusieurs participants : obligatoires (VENDEUR, ARGENT, BIEN, ACHETEUR) et facultatifs (MOYEN), etc. Lorsque l'unité évocatrice du cadre est un verbe, l'analyse est focalisée sur les arguments de ce dernier qui représentent les éléments du cadre.

2.3 Les classifications de verbes

2.3.1 La classification des verbes (anglais) selon Levin

Beth Levin (Levin, 1993) propose une classification lexico-sémantique de verbes anglais à partir d'une analyse de leur fonctionnement (syntaxe, classe sémantique des arguments sélectionnés, etc.). Les verbes qui affichent un ensemble d'alternances (de diathèses ou frames) identiques ou similaires dans la réalisation de leurs structures argumentales sont supposés partager certains éléments de sens et, de ce fait, sont regroupés dans une classe sémantiquement homogène. L'alternance de diathèses (la relation entre deux réalisations de surface d'un même prédicat), qui est le principal critère d'identification des classes verbales dans cette approche, est appuyée par des propriétés supplémentaires liées à la sous-catégorisation, à la morphologie et aux verbes ayant un sémantisme complexe. À partir de ces critères, la classification couvre 3 024 verbes, 4 186 sens, 240 classes de verbes construites autour de 79 alternances. Par exemple, la classe des prédicats dénotant une configuration spatiale contient les verbes suivants : *balance, bend, bow, crouch, dangle, flop, fly, hang, hover, jut, kneel, lean, lie, loll, loom, lounge, nestle, open, perch, plop, project, protude, recline, rest, rise, roost, sag, sit, slope, slouch, slump, sprawl, squat, stand, stoop, straddle, swing, tilt, tower* (Levin, 1993). Une extension substantielle de cette classification intègre 57 nouvelles classes pour les verbes qui n'ont pas été couverts initiallement (Korhonen & Briscoe, 2004). Parmi les nouvelles classes, *FORCE class* regroupe les verbes tels que *manipulate*, *pressure*, *force*.

2.3.2 Les classes d'objets de Gaston Gross

Une classe d'objets est un *« ensemble de substantifs, sémantiquement homogènes, qui détermine une rupture d'interprétation d'un prédicat donné, en délimitant un emploi spécifique »* (Gross, 2008). En d'autres termes, les classes d'objets déterminent l'interprétation donnée d'un prédicat parmi d'autres possibles. Elles sont induites par les prédicats (verbes et adjectifs) et permettent d'identifier en contexte les mots avec lesquels ils entretiennent une relation conceptuelle telle que la synonymie, l'antonymie, etc. Ces entités sont construites sur des bases syntaxiques et concernent particulièrement les compléments qui apportent beaucoup plus d'informations que le sujet dans l'interprétation d'un prédicat (Gross, 2012). Par exemple, la phrase *vous suivez* n'est pas assez précise, ce qui rend son interprétation difficile. Par contre, si l'on y ajoute un complément, l'interprétation sera plus aisée et la signification du verbe sera plus transparente. Ainsi, dans la phrase *vous suivez ce chemin*, l'objet *chemin* peut être remplacé par un autre substantif comme *route, rue, voie, sentier* et le verbe garde le même sens. Ces substantifs peuvent donc être considérés comme appartenant à une même classe d'objets, celle de <voies>. Par contre, si on remplace *chemin* par le mot *cours*, on est face à un autre emploi du verbe car *cours* appartient à une autre classe d'objets, appelée <enseignements>. Elle contient les mots comme *séminaire, stage, formation, cycle études, etc*. Le principal intérêt des classes d'objets est de rendre compte des différents emplois des prédicats, en déterminant leurs schémas d'arguments et en rattachant à ceux-ci un ensemble de propriétés qui les caractérisent (Gross, 2008).

2.3.3 Lexique-Grammaire des verbes français

Le lexique-grammaire des verbes du français (Gross, 1975) est un dictionnaire syntaxique électronique téléchargeable [1]. Il est organisé en plusieurs tables, chacune regroupant les verbes du lexique qui ont un fonctionnement comparable : constructions types, distribution des actants, sémantique, etc. Chaque table comprend un ensemble de propriétés, et un codage qui précise si l'élément a ou non cette propriété. Chaque entrée d'une table contient les informations suivantes : l'élément vedette, une construction type dans laquelle il peut apparaître, et des constructions associées à cette construction type. Les différents emplois des verbes, énumérés dans les tables, sont décrits grâces à des propriétés structurelles, distributionnelles et sémantiques. Par exemple, les tables de constructions sans compléments prépositionnels contiennent des constructions types, parmi lesquelles N_0 V et N_0 V N_1. La construction N_0 V accueille les verbes tels que *pleuvoir*, *bêtifier*, *bouillir*, *pisser* et selon le verbe, elle peut accepter un N_0 humain (*Luc bêtifie*), non humain (*l'eau bout*), impersonnel (*il pleut*), et le verbe peut être modifié par un adverbe (*ça ne pisse pas loin*) (Leclere, 1990).

3 Le verbe dans les travaux en terminologie

Les entités nominales ont longtemps occupé la place centrale dans les travaux sur les langues de spécialité au détriment des autres parties du discours, plus particulièrement des verbes, mis à l'écart pour diverses raisons. En effet, les travaux en terminologie se focalisent la plupart du temps sur la description des concepts, ou des entités nominales (particulièrement les noms) et la mise au jour des relations qu'elles partagent (genre-espèce, partie-tout, etc.). L'un des motifs principaux énoncés justifiant l'exclusion du verbe est la place accordée aux objets et à leurs dénominations dans l'approche de Wüster (Wüster, 1985). Cette situation trouve également une explication dans le fait que les entités nominales sont généralement utilisées pour le développement des terminologies, ontologies, thésaurus, glossaires, ou des vocabulaires. Ce constat s'explique également par les besoins croissants des applications : l'indexation et l'extraction d'informations sont des tâches typiquement basées sur les entités nominales. Pour ces raisons, la plupart des approches théoriques et méthodologiques sont adaptées aux entités nominales. Néanmoins, quelques travaux s'inspirant de la sémantique lexicale s'intéressent aux verbes et à leur mode de fonctionnement dans les domaines spécialisés. Ces travaux montrent que l'étude du verbe est quasi indispensable dans le cadre des activités comme l'extraction d'informations (Tateisi *et al.*, 2004), la conception des dictionnaires terminographiques (Tellier, 2008) ou encore la traduction spécialisée (Pimentel, 2011). Les structures argumentales des verbes peuvent également servir pour la détection automatique des relations sémantiques (Massimiliano *et al.*, 2008). Nous parlerons de deux approches d'analyse du verbe terminologique : l'approche conceptuelle (section 3.1) et l'approche lexico-sémantique (section 3.2).

3.1 L'approche conceptuelle

Le principe de l'approche conceptuelle stipule que l'on ne s'intéresse au verbe que s'il a l'aptitude de désigner un « concept d'activité », c'est-à-dire une activité (L'Homme, 2012). Telle est la condition qui détermine l'intégration des verbes dans des ressources terminologiques. Autrement dit, le verbe ne peut être considéré comme terme que s'il est fortement assimilable à un nom sur le plan conceptuel. Rey définit clairement le statut du verbe selon la perspective conceptuelle en ces termes : *« la terminologie ne s'intéresse aux signes (mots et unités plus grandes que le mot) qu'en tant qu'ils fonctionnent comme des noms dénotant des objets et comme des « indicateurs de notions » (de concepts) et dans cette optique, les verbes sont des noms de processus, d'actions »* (Rey, 1979). Cette conception justifie en partie la discrimination observée entre les parties du discours traitées dans un dictionnaire spécialisé. En général, on y compte très peu de verbes et d'adjectifs, mais beaucoup d'entités nominales. Les résultats d'une étude portant sur la présence des verbes dans les dictionnaires de spécialité évaluent à 2,44% (entre 0 et 4 verbes par dictionnaire) la moyenne d'apparition des verbes dans quatre dictionnaires terminologiques (L'Homme, 2003). L'approche conceptuelle a débouché de nos jours sur une démarche conceptuelle, incarnée par les ontologies, qui permet de distinguer les concepts d'activité, exprimés par les noms ou par les verbes, dans les domaines de spécialité. Ainsi, dans le domaine médical par exemple, les verbes tels que *traiter*, *observer* et *activer* peuvent devenir terminologiques puisqu'ils permettent de rendre compte des notions comme *traitement de la maladie, observation du patient* et *activation des cellules* (L'Homme, 2012).

1. http ://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Telechargement.html

3.2 L'approche lexico-sémantique

La sémantique lexicale est le cadre théorique qui a montré l'importance de la structure argumentale du verbe et du réseau lexical auquel le verbe appartient. Dans ce cadre, la caractérisation de la nature spécialisée du verbe est basée sur la description de sa structure argumentale ou son appartenance à un ou plusieurs réseaux lexicaux, (morpho-)sémantiques ou paradigmatiques. Ces tâches reposent sur l'observation et l'analyse des différentes occurrences du verbe en corpus.

3.2.1 La structure argumentale

L'analyse de la structure argumentale du verbe peut avoir pour but de démontrer sa nature terminologique. En effet, la nature prédicative du verbe fait qu'il a besoin des éléments qu'il régit pour la réalisation de son sens. Une étude propose de prendre en considération la nature des arguments du verbe qui détermine son degré de spécialisation (L'Homme, 1998). Ce raisonnement illustre l'hypothèse selon laquelle le verbe n'est pas spécialisé par lui même, mais grâce à la prise en compte de sa structure argumentale (L'Homme, 2012). C'est ce critère qui permet d'admettre *installer* comme verbe spécialisé dans l'exemple suivant :

L'utilisateur installe la nouvelle version du traitement de texte sur son PC.

Dans cette phrase, les termes (*utilisateur*, *version*, *PC*) qui représentent les têtes des arguments du verbe appartiennent au domaine de l'informatique. Par conséquent, *installer* peut être considéré comme verbe terminologique dans ce domaine.

L'analyse des arguments des verbes constitue également un critère de poids chez (Tellier, 2008) qui y trouve un moyen de sélection des verbes, à partir d'un corpus spécialisé relevant du domaine de l'infectiologie, représentant de bons candidats termes à ajouter dans un dictionnaire spécialisé. Ce critère est également utilisé dans d'autres travaux (Lerat, 2002; Pimentel, 2011). Cependant, la caractérisation des arguments du prédicat verbal n'a pas pour unique but l'identification des verbes terminologiques. D'autres objectifs peuvent être poursuivis : l'extraction d'informations dans les corpus spécialisés du domaine de la biologie moléculaire (Tateisi *et al.*, 2004), l'élaboration d'un dictionnaire juridique portugais-anglais (Pimentel, 2011), l'analyse contrastive des corpus médicaux de niveaux de spécialisation différents (expert vs profane)(Wandji Tchami *et al.*, 2013).

3.2.2 Le réseau lexical

Outre la nature des actants, d'autres paramètres peuvent être pris en compte par les chercheurs lors du repérage des verbes terminologiques. L'un de ces paramètres, qui revient très souvent, est le lien qu'un verbe peut avoir avec un nom. Ainsi, si le nom est terminologique, et si le verbe est sémantiquement et le plus souvent morphologiquement apparenté à celui-ci, alors, il est fort possible que le verbe soit spécialisé lui aussi (L'Homme, 2012). Ce critère s'observe avec les couples tels que *développement - développer*, *téléchargement - télécharger*, *rechauffement - rechauffer*, le verbe et le nom correspondant désignent tous les deux une activité. Cependant, il existe des cas où le sens du verbe et celui du nom sont distincts malgré le lien morphologique qui existe entre eux. C'est le cas du couple *programme - programmer*, où le nom *programme* désigne le résultat de l'activité que dénote le verbe *programmer*. Comme nous pouvons le constater, dans l'approche lexico-sémantique, les noms peuvent servir de point de départ à partir duquel les verbes spécialisés sont identifiés en fonction des liens qu'ils partagent avec eux. C'est d'ailleurs cette méthode qui permet de retenir les verbes *évoluer*, *excréter*, *infecter* et *sécréter* comme termes du domaine de l'infectiologie, de part leur parenté aux noms *évolution*, *excrétion*, *infection* et *sécrétion* (Tellier, 2008). Comme ces noms sont fortement spécialisés dans ce domaine, les verbes correspondant héritent de cette caractéristique. Toutefois, il est possible de déplacer le point de départ de l'analyse vers le verbe. Cette technique peut permettre de découvrir d'autres unités reliées au verbe et d'élargir ainsi le réseau lexical construit autour de ce dernier (L'Homme, 2012).

Cette démarche a été appliquée lors de la conception du DicoInfo (Dictionnaire fondamental de l'informatique et de l'Internet), une base de données lexicales contenant des termes (les verbes y compris) fondamentaux du domaine de l'informatique et de l'internet (L'Homme, 2009). L'approche utilisée s'inspire grandement des principes théoriques et méthodologiques de la Lexicologie explicative et combinatoire (Mel'cuk *et al.*, 1995) et permet de fournir pour chaque entrée différents types d'informations : la réalisation linguistique des actants, les liens lexicaux, les synonymes, les contextes d'apparition du terme, etc. Pour le verbe *programmer* par exemple, DicoInfo propose divers types d'unités lexicales appartenant au réseau lexical notamment, *programmation* (*action de programmer*), *programme* (*résultat de l'action de programmer*), *informaticien* (*agent de l'action de programmer*), *langage* (*instrument utilisé pour programmer*), *logiciel*

(*résultat l'action de programmer*), *écrire* (*synonyme de programmer*), *développer* (*synonyme de programmer*), etc. Cet exemple permet d'observer que les mots repérés sont liés au verbe par différentes relations exprimées de façon implicite à travers de courtes gloses explicatives.

4 Le verbe en TAL

Le traitement des verbes dans le domaine du TAL s'appuie le plus souvent sur la caractérisation de leur structure argumentale : la valence verbale (Eynde & Mertens, 2003), les possibilités combinatoires et les relations de dépendances (Marneffe *et al.*, 2006), les fonctions grammaticales et rôles sémantiques des arguments (Gildea & Jurafsky, 2002), la désambiguïstation du sens des verbes (Ide & Véronis, 1998; Ye & Baldwin, 2006; Wagner *et al.*, 2009; Brown *et al.*, 2011), l'acquisition de schémas de sous-catégorisation à partir de l'analyse automatique de gros corpus (Messiant *et al.*, 2010), etc. Dans cette section, nous nous focalisons sur deux types de travaux : l'étiquetage des rôles sémantiques (section 4.1) et la désambiguïsation du sens des verbes (section 4.2). Par la suite, nous faisons la description de quelques ressources dédiées au verbe (section 4.3).

4.1 Étiquetage des rôles sémantiques

L'étiquetage des rôles sémantiques (Gildea & Jurafsky, 2002; Palmer *et al.*, 2005; Swier & Stevenson, 2004; Ye & Baldwin, 2006), ou *Semantic Role Labeling (SRL)*, est une tâche du TAL qui consiste à identifier de façon automatique les relations ou les rôles sémantiques (agent, patient, recipient, etc.) que jouent les constituants d'une phrase dans un cadre sémantique donné. Cette tâche est nécessaire pour la conception de différents types d'applications, et plus particulièrement celles qui touchent la compréhension et l'interprétation de la langue. Il s'agit par exemple de systèmes de questions-réponses (Miller *et al.*, 1996), d'extraction d'informations (Surdeanu *et al.*, 2003), de traduction automatique (Boas, 2002), ou de résumé automatique (Melli *et al.*, 2005). Les unités prédicatives (verbes, noms, adjectifs) occupent généralement le coeur des études qui concernent la SRL. En ce qui concerne le verbe, l'annotation consiste généralement à identifier dans la phrase les limites de ses arguments et éventuellement des circonstants, et ensuite de leur associer des rôles sémantiques selon le contexte. La démarche la plus utilisée pour la réalisation d'une SRL comprend trois étapes principales : (1) l'identification des arguments du verbe, basée le plus souvent sur des heuristiques (Xue & Palmer, 2004) qui permettent de réduire le nombre de candidats ; (2) le calcul des probabilités pour chacune des étiquettes à représenter les rôles sémantiques possibles ; (3) l'attribution de scores à chaque étiquette, éventuellement combinée à d'autres facteurs de prédiction, pour assigner des étiquettes appropriées aux arguments des verbes.

De nos jours, les modèles d'apprentissage statistiques sont très sollicités pour l'annotation des textes en rôles sémantiques. L'un des travaux de référence propose un système de SRL statistique, qui peut être utilisé aussi pour l'analyse syntaxique, l'étiquetage des parties du discours (Church, 1988), et la désambiguïsation du sens des mots (Lapata & Brew, 2004). Ce système, conçu pour les verbes, les noms et les adjectifs, atteint 82% de précision sur des phrases pré-annotées manuellement, tandis qu'il montre 65% de précision et 61% de rappel sur des phrases non annotés (Gildea & Jurafsky, 2002). Il a d'ailleurs été utilisée dans le cadre du projet FrameNet.

4.2 Désambiguïsation du sens des verbes

La désambiguïsation du sens des verbes ou *Verb Sense Disambiguation (VSD)* est une sous-tâche de la WSD (word sense disambiguation). Elle consiste à sélectionner automatiquement, parmi ses différents sens, le sens le plus approprié d'un verbe polysémique, selon son contexte d'apparition. Par exemple, le verbe *read* (lire) a plusieurs sens. Pour faire la distinction entre les phrases telles que *I read a book* (*je lis un livre*) et *I read you loud and clear* (*je te comprends parfaitement*), il est nécessaire de désambiguïser le contexte d'apparition du verbe, en suivant une des méthodes existantes. La désambiguïsation du sens est une tâche nécessaire pour la traduction automatique (Carpuat & Wu, 2007) ou l'extraction d'informations (Schütze & Pedersen, 1995; Sanderson, 1994). Deux types d'approches sont utilisées habituellement pour la désambiguïsation du sens des mots : approche à base de règles et approche à base d'apprentissage. L'approche à base de règles requiert des ressources comme les bases de données lexicales, les dictionnaires électroniques, qui fournissent des descriptions lexicales, syntaxiques et sémantiques des mots. À partir de ces ressources, des règles sont définies pour déterminer le sens exact du mot parmi l'ensemble des sens possibles. En traduction automatique, un ensemble constitué de 63 règles est proposé comme source de connaissances (Specia *et al.*, 2005). L'approche la plus utilisée actuellement est

basée sur l'apprentissage automatique (Ye & Baldwin, 2006; Brown *et al.*, 2011; Yarowsky, 1995). L'apprentissage peut être supervisé (exigeant un ensemble d'exemples manuellement annotés) ou non supervisé (appliqué sur des textes non annotés). En effet, certains chercheurs proposent une technique non supervisée de désambiguïsation du sens des verbes, qui regroupent les verbes ayant les préférences sélectionnelles et de sous-catégorisation similaires (Wagner *et al.*, 2009). Cette méthode montre 57.06% de précision. D'autres travaux identifient les préférences sémantiques (Lapata & Brew, 2004) et les marques de sous-catégorisation (Lapata & Brew, 1999) des verbes apparaissant dans plusieurs classes de Levin.

En ce qui concerne les méthodes supervisées, les premières expériences se focalisaient sur les bi-grammes et les fonctions linguistiques et contextuelles (Pedersen, 2000, 2001; Hoa Trang & Palmer, 2002). Par la suite, les chercheurs se sont intéressés à l'apport des bases de connaissances fournissant les informations telles que la catégorie grammaticale des mots voisins, la forme morphologique, les collocations, la relation syntaxique verbe-objet, utiles pour lever certaines ambiguïtés (Yoong Keok & Hwee Tou, 2002). De plus en plus, les chercheurs abordent les rôles sémantiques des arguments des verbes comme des fonctions contribuant à l'amélioration des performances des systèmes lors de la désambiguïsation du sens des verbes (Hoa Trang & Palmer, 2005; Ye & Baldwin, 2006). Cette technique est d'ailleurs recommandée car les rôles sémantiques associés à un mot peuvent donner des indices pour la déduction de son sens, surtout lorsque ces rôles sont associés à des frames de sous-catégorisation syntaxique (Gildea & Jurafsky, 2002). Certains travaux suivent une approche supervisée basée sur connaissances extraites des ressources lexicales externes (VerbNet, WordNet, etc.) (*knowledge based WSD*) (Brown *et al.*, 2011). Une autre approche, inspirée par les travaux en psycholinguistique, propose de nouveaux critères de regroupement des sens d'un mot, en fonction de la différence faible ou importante qui existe entre ces mots (Brown, 2008). Pour le français, il existe une approche d'analyse sémantique des textes, basée sur des réseaux lexicaux et les relations de dépendance entre les mots ambigus et les autres mots de la phrase (Mouton, 2010).

La réalisation de la WSD sur des textes spécialisés est actuellement une tâche relativement difficile selon les domaines, à cause de l'absence des ressources terminologiques nécéssaires ou de l'insuffisance des données disponibles. Néanmoins, dans certains domaines comme l'informatique biomédicale, différentes études proposent des systèmes de WSD basés sur des méthodes non supervisées (Liu *et al.*, 2001) ou supervisées (Stevenson & Guo, 2010), utilisant des terminologies existantes.

4.3 Quelques ressources lexicales dédiées au verbe

Dans la suite de cette section, nous décrivons brièvement quelques ressources lexicales : FrameNet (section 4.3.1), VerbNet (section 4.3.2), VerbOcean (section 4.3.3) et WordNet (section 4.3.4). Nous nous intéressons particulièrement à la manière dont l'information sur le verbe est présentée.

4.3.1 FrameNet

FrameNet (Ruppenhofer *et al.*, 2006) est une base de données lexicales [2] initialement conçue pour l'anglais. Elle contient plus de 10 000 sens des unités lexicales décrits à travers plus de 1 000 cadres sémantiques liés hiérarchiquement les uns aux autres et illustrés par plus de 170 000 phrases. Le projet FrameNet propose une description des unités lexicales prédicatives (verbes, noms et adjectifs), basée sur l'annotation en cadres sémantiques (Fillmore, 1982) des phrases dans lesquelles ces unités apparaissent.

His *$20* TRANSACTION *with Amazon.com* *for a new TV* *had been very smooth.*

Dans cette phrase, chaque couleur représente un élément du cadre : bleu foncé=ACHETEUR, bleu ciel=ARGENT, rouge=VENDEUR, vert=BIEN.

Ces frames mettent en évidence des informations sémantiques nécessaires pour capturer les sens de l'unité lexicale clé. Ainsi, pour chacune de ses entrées, FrameNet est capable de fournir un cadre sémantique complet, une description du frame, ses éventuelles relations avec d'autres frames, une description des éléments du frame et une illustration des schémas valenciels de l'entrée à l'aide d'exemples (Ruppenhofer *et al.*, 2006).

2. https ://framenet.icsi.berkeley.edu/fndrupal/about

4.3.2 VerbNet

Contrairement à FrameNet, VerbNet [3] (Kipper *et al.*, 2000; Kipper-Schuler, 2005) est totallement focalisé sur les verbes. Cette ressource lexicale propose une description des verbes basée sur la classification de Levin (section 2.3.1). Elle consiste à regrouper les verbes en différentes classes, qui mettent en évidence leurs propriétés syntaxiques et sémantiques communes. Cette méthode de description permet de faire des généralisations sur le comportement des verbes. Par exemple, les verbes appartenant à la classe *Hit 18.1* : *bang, bash, hit, kick...* sont des transitifs direct. Ils exigent un agent et un patient, et peuvent être modifiés par des prédicats sémantiques exprimant la manière, la cause, la direction, etc.

VerbNet est donc un lexique hiérarchique de verbes anglais regroupés en classes, indépendamment des domaines de spécialités auxquels ils peuvent appartenir. Chaque classe est décrite à travers : l'ensemble d'arguments possibles, présentés sous forme de rôles thématiques ; les éventuelles restrictions de sélection d'arguments (comme *animé*, *humain*, *organisation*) ; les cadres, décrivant les possibles réalisations de surface de la structure argumentale (constructions transitives, intransitives, syntagmes prépositionnels, résultatives) ; les alternances de diathèse, c'est-à-dire les variations des différents cadres. Selon le site officiel, après son extension (Korhonen & Briscoe, 2004), VerbNet compte 274 classes de premier niveau, 23 rôles thématiques, 94 prédicats sémantiques, 55 restrictions syntaxiques, 5 257 sens des verbes et 3 769 lemmes.

4.3.3 VerbOcean

VerbOcean (Chklovski & Pantel, 2004) est une ressource lexicale qui propose un réseau sémantique de relations entre les verbes, et recense uniquement des paires de verbes sémantiquement proches. Elle contient 22 306 relations entre 3 477 verbes et identifie 5 types de relations : *similitude* (la similitude), *strenght* (la force), *antonymy* (l'antonymie), *enablement* (l'habilitation), et la relation temporelle *happens-before* (a lieu avant). L'approche appliquée pour la conception de cet outil est basée sur deux étapes : (1) la détection des paires de verbes qui apparaissent en co-ooccurrence fréquente, grâce à des requêtes effectuées sur le portail Google ; (2) pour chaque paire, le calcul du score de chaque relation possible, grâce à 35 schémas lexico-syntaxiques. Par exemple, les verbes *discover* (*découvrir*) et *refine* (*affiner, améliorer*) sont considérés comme une paire illustrant la relation *happens-before* si la chaîne *discovered and refined* (instantiant le schéma *Xed and then Yed*) est identifiée de façon très fréquente sur Google.

4.3.4 WordNet

WordNet (Fellbaum, 1998) est une base de données lexicale qui propose une description des verbes, mais également des noms et des adjectifs, sur la base de différentes relations sémantiques : la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, la méronymie, la troponymy et l'implication (Miller, 1995). Contrairement à VerbOcean qui s'intéresse uniquement aux paires de verbes sémantiquement proches, WordNet traite plusieurs catégories d'unités prédicatives et ces unités sont regroupées dans des *synsets*, 117 000 au total. Un synset est un groupe de mots (synonymes) sémantiquement homogènes. Il contient des pointeurs qui marquent ses relations conceptuelles avec d'autres synsets. En outre, un synset contient une brève définition et, dans la plupart des cas, une ou plusieurs courtes phrases illustrant l'utilisation des membres de ce synset. Les formes des mots ayant plusieurs significations sont représentées par autant de synsets distincts.

5 Discussion et travaux futurs

Comme indiqué plus haut, le projet que nous entreprenons a pour objectif de proposer une méthode de simplification de textes médicaux écrits en français, à partir d'une analyse syntaxico-sémantique des verbes en contexte. Nous avons vu dans la section 3 que les travaux sur les langues de spécialité sont le plus souvent focalisés sur les entités nominales et, par conséquent, les travaux sur les verbes terminologiques sont peu nombreux. De même, dans la section 4, nous démontrons que peu de travaux en TAL appliquent la sémantique des cadres à des textes spécialisés et qu'il existe encore des cadres théoriques dans lesquels le verbe et sa structure argumentale sont peu considérés. En rupture avec ces constats, nous proposons d'exploiter l'étude de la structure argumentale des verbes pour la simplification des textes spécialisés. Nous partons de l'hypothèse selon laquelle le verbe, en tant que prédicat central dans la phrase, peut être le point de départ pour cerner la syntaxe et la sémantique des textes spécialisés puisqu'il sert à articuler l'expertise et les connaissances portées

3. http ://verbs.colorado.edu/ mpalmer/projects/verbnet.html

par les mots qui l'entourent dans la phrase (L'Homme & Bodson, 1997). Les travaux de l'état de l'art peuvent nous venir en aide à différentes étapes de la méthode, dont les principales sont :

1. *Annotation automatique des arguments des verbes selon leurs catégories sémantiques.* Comme la SRL, cette annotation vise à détecter automatiquement les verbes et leurs arguments, et à associer des étiquettes sémantiques à ces arguments. Toutefois, dans notre travail, les arguments seront associés non pas à des rôles sémantiques, mais à des types sémantiques proposés par une ressource terminologique existante (Côté, 1996). Pour ce faire, nous allons implémenter un programme qui prend en entrée le fichier résultant de l'analyse morpho-syntaxique, la terminologie et éventuellement d'autres ressources. Afin d'optimiser l'identification des arguments et de repérer plus de termes, nous prévoyons de définir des heuristiques basées sur la coordination, la relation hypéronyme/hyponyme, les têtes lexicales, etc. De même, pour augmenter la couverture de cette annotation, nous allons enrichir la terminologie avec des entrées supplémentaires.
2. *Analyse des verbes.* L'analyse des verbes est effectuée grâce aux informations obtenues lors de l'annotation sémantique des arguments. Il s'agit typiquement d'analyser leur nombre et leurs types sémantiques. Des méthodes d'apprentissage non supervisé peuvent être utilisées pour distinguer entre les emplois des verbes. Plus particulièrement, à partir des annotations, nous souhaitons savoir si ces emplois véhiculent les sens spécialisés ou non. Pour cette étape aussi, nous pouvons nous appuyer sur les ressources existantes, qui viendront compléter ou renforcer les annotations obtenues à l'étape précédente. Bien que l'apport de ces ressources est limité lorsqu'il s'agit de traiter des textes spécialisés, elles restent utiles, plus particulièrement pour l'analyse des emplois non spécialisés des verbes. Par exemple, les ressources comme VerbNet et FrameNet peuvent fournir des informations standard sur les schémas valenciels des verbes. Les ressources de type classes d'objets ou WordNet peuvent fournir les séries d'arguments qui sont sémantiquement proches entre eux. L'analyse des verbes peut permettre d'effectuer plusieurs types d'appréciation : la complétude des annotations, l'importance de certains arguments des verbes, la déviation des schémas argumentaux qui peuvent être révélateurs des emplois non standard et spécialisés des verbes. Des heuristiques dédiées seront nécessaires pour combiner les annotations avec les ressources, et pour départager ces différents cas de figure ;
3. *Simplification.* Dans le cas des textes spécialisés, rédigés pour un public de non experts, les emplois spécialisés des verbes peuvent être considérés comme des sources de difficulté. Grâce à l'étape précédente, de tels emplois spécialisés peuvent être détectés automatiquement. La simplification a pour objectif de rendre ces emplois de verbes plus abordables pour les utilisateurs non spécialistes. A ce niveau, l'absence de ressources du type WordNet pour les langues de spécialité représente une difficulté cruciale que nous allons devoir affronter. Dans un premier temps, pour pallier à ce problème, les phrases simplifiées seront conçues sur un modèle que proposent les définitions des termes du DicoInfo (L'Homme, 2009). Il s'agira de fournir une définition typique de la construction verbale ambigüe dans laquelle entre le verbe. Cette définition sera enrichie par un ou plusieurs synonymes du verbe qui seront recherchés dans WordNet ou d'autres ressources qui proposent les synonymes des mots de la langue générale. Une étude comparative des corpus de textes spécialisés, écrits par des experts et ceux écrits par des non-experts, effectuée au préalable, sera utile lors de la simplification, pour l'identification, si possible, des constructions verbales synonymes. La simplification concernera également les constituants syntaxiques de la phrase et éventuellement des temps verbaux (Brouwers *et al.*, 2014). En exploitant la méthode appliquée dans FrameNet et grâce aux observations en corpus, nous pouvons détecter les arguments nécessaires (*core*) et non nécessaires (*non core*) et alléger les phrases en supprimant les éléments non nécessaires. De la même manière, si les éléments nécessaires à la compréhension sont absents, nous pouvons les déduire et compléter ainsi la structure argumentale de verbes, en espérant que cela facilite la compréhension des phrases.

Au terme de ces différentes étapes, nous pensons pouvoir améliorer la lisibilité du texte et de rendre le sens des verbes plus accessibles aux utilisateurs non spécialistes en médecine. Les résultats de notre approche lexico-syntaxique seront évalués et comparés à ceux des méthodes de simplification focalisées uniquement sur les entités nominales, c'est-à-dire sur les arguments des verbes.

6 Conclusion

Tout au long de ce travail, nous avons exploré les principaux cadres théoriques et travaux qui s'intéressent particulièrement au prédicat verbal dans trois domaines de recherche : terminologie, où le verbe a tardé à s'imposer comme unité pouvant exprimer des connaissances spécialisées, face à la place dominante des entités nominales ; linguistique, où le verbe a toujours fait partie des catégories grammaticales les plus étudiées ; TAL, où de nos jours, de nombreuses ressources

et méthodes se consacrent partiellement ou entièrement aux verbes à travers l'étude de sa structure argumentale ou de ses relations sémantiques avec d'autres verbes. De façon générale, notre travail met en évidence le fait que la frontière entre ces trois disciplines n'est pas étanche, car les techniques et approches utilisées en linguistique sont réutilisées en terminologie et en TAL, et de la même façon, la linguistique et la terminologie contemporaines font très souvent recours aux ressources, outils et applications développées en TAL. Nous allons nous servir de cette interdisciplinarité pour mener à bien notre projet.

Références

BOAS H. (2002). Bilingual framenet dictionaries for machine translation. In *LREC*, p. 1364–137, Las Palmas de Gran Canaria, Spain.

BROUWERS L., DELPHINE B., ANNE-LAURE L. & THOMAS F. (2014). Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, p. 47–56.

BROWN S. (2008). Choosing sense distinctions for wsd : Psycholinguistic evidence. In *Proceedings of ACL/HLT*, p. 249–252, Columbus, OH.

BROWN S., DLIGACH D. & PALMER M. (2011). Verbnet class assignment as a wsd task. In *9th International Conference on Computational Semantics*, Oxford, UK.

CARPUAT M. & WU D. (2007). Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*, p. 61–72.

CHKLOVSKI T. & PANTEL P. (2004). Verbocean mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

CHURCH K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, Austin, Texas.

CÔTÉ R. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

ELHADAD N. (2006). Comprehending technical texts : Predicting and defining unfamiliar terms. In *AMIA*, p. 239–243.

EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *French Language Studies*, **13**(1), 63–104.

FELLBAUM C. (1998). *WordNet : An Eletronic Lexical Database*, In M. PRESS, Ed., *Language Speech and Communications*.

FILLMORE C. (1968). *The case for case*, In UNIVERSALS, Ed., *Linguistic Theory*, p. 1–88.

FILLMORE C. (1976). *Topics in lexical semantics*, In I. U. PRESS, Ed., *Current Issues in Linguistic Theory*, p. 76–138.

FILLMORE C. (1982). *Frame Semantics*, In H. P. CO, Ed., *Linguistics in the morning calm*, p. 111–137.

FONTENELLE T. (2009). sémantique des cadres et lexicographie. *Lexique*, (19), 162–177.

GILDEA D. & JURAFSKY D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, **28**(3), 245–288.

GROSS G. (2008). Les classes d'objets. *Lalies*, (28), 111—165.

GROSS G. (2012). *Manuel d'analyse linguistique : approche sémantico-syntaxique du lexique*. Villeneuve-d'Ascq : Presses universitaires du Septentrion.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann.

HEILMAN M., THOMPSON C., CALLAN J. & ESKENAZI M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *HLT-NAACL*, p. 460–467, Rochester, New York.

HEILMAN M., THOMPSON C. & ESKENAZI M. (2008). An analysis of statistical models and features for reading difficulty prediction. In COLUMBUS, Ed., *Third Workshop on Innovative Use of NLP for Building Educational Applications*, p. 71–79, Ohio.

HOA TRANG D. & PALMER M. (2002). Combining contextual features for word sense disambiguation. In A. FOR COMPUTATIONAL LINGUISTICS, Ed., *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation : Recent Successes and Future Directions*, p. 88–94, Stroudsburg, PA, USA.

HOA TRANG D. & PALMER M. (2005). The role of semantic roles in disambiguating verb senses. In A. FOR COMPUTATIONAL LINGUISTICS, Ed., *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 42–49, Stroudsburg, PA, USA.

IDE N. & VÉRONIS J. (1998). Introduction to the special issue on word sense disambiguation : The state of the art. *Compututational Linguistic*, **24**(1), 2–40.

KIPPER K., DANG H. & PALMER M. (2000). Class-based construction of a verb lexicon. In *The Seventh National Conference on Artificial Intelligence AAAI/IAAI*, p. 691–696.

KIPPER-SCHULER K. (2005). *VerbNet : A broad-coverage comprehensive verb lexicon*. Thèse de doctorat, niversity of Pennsylvania, Philadelphia, PA.

KORHONEN A. & BRISCOE T. (2004). Extended lexical-semantic classification of english verbs. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, Boston, MA.

LAPATA M. & BREW C. (1999). Using subcategorization to resolve verb class ambiguity. In *JOINT SIGDAT CONFERENCE ON EMPIRICAL METHODS IN NLP AND VERY LARGE CORPORA*, p. 266–274.

LAPATA M. & BREW C. (2004). Verb class disambiguation using informative priors. *COMPUTATIONAL LINGUISTICS*, p. 45–73.

LECLERE C. (1990). Organisation du lexique-grammaire des verbes français. *Langue française*, (87), 112–122.

LERAT P. (2002). Qu'est-ce que le verbe spécialisé ? le cas du droit. *Cahiers de Lexicologie*, **80**, 201–211.

LEROY G., ENDICOTT J., MOURADI O., KAUCHAK D. & JUST M. (2012). Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *American Medical Infomatics Association*.

LEVIN B. (1993). *English Verb Classes and Alternation, A Preliminary Investigation*. The University of Chicago : Press.

L'HOMME M. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, **73**(2), 61–84.

L'HOMME M. (2003). Capturing the lexical structure in special subject fields with verbs and verbal derivatives a model for specialized lexicography. *IJL*, **16**(4), 403–422.

L'HOMME M. (2009). *Le DiCoInfo. Dictionnaire fondamental de l'informatique et de l'Internet*. Rapport interne, Observatoire de linguistique Sens-Texte (OLST).

L'HOMME M. (2012). Le verbe terminologique un portrait de travaux récent. In *Congrès Mondial de Linguistique Française-CMLF*, p. 93–107.

L'HOMME M. & BODSON C. (1997). Modèle de description des verbes specialises combinant base de connaissances et hypertexte. In *Congres international de terminologie*, p. 381–398, San Sebastian, Espagne.

LIU H., LUSSIERB Y. & FRIEDMAN C. (2001). Disambiguating ambiguous biomedical terms in biomedical narrative text : An unsupervised method. *Journal of Biomedical Informatics*, **34**, 249–261.

MARNEFFE M., MACCARTNEY B. & MANNING C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, p. 449–454.

MASSIMILIANO C., ALDO G., ESTHER R., J S. & ISABEL R. (2008). Unsupervised learning of semantic relations for molecular biology ontologies. In *Proceeding of the 2008 conference on Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, p. 91–104.

MELLI G., WANG Y., LIU Y., KASHANI M., SHI Z., GU B., SARKAR A. & POPOWICH F. (2005). Description of squash the sfu question answering summary handler for the duc-2005 summarization task. In *HLT/EMNLP*.

MEL'CUK I., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot / Aupelf-UREF.

MESSIANT C., GÁBOR K. & POIBEAU T. (2010). Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues*, **51**(1), 65–96.

MILLER G. A. (1995). Wordnet : A lexical database for english. *Communication ACM*, **38**(11), 39–41.

MILLER S., STALLARD D., BOBROW R. & SCHWARTZ R. (1996). A fully statistical approach to natural language interfaces. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, p. 55–61, Stroudsburg, PA, USA.

MOUTON C. (2010). *Ressources et méthodes semi-supervisées pour l'analyse sémantique de texte en français*. PhD thesis, Université Paris Sud-Paris XI.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank an annotated corpus of semantic roles. *Computational Linguistics*, **31**(1), 71–105.

PEDERSEN T. (2000). A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. *CoRR*, **cs.CL/0005006**.

PEDERSEN T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, p. 1–8, Stroudsburg, PA, USA.

PIMENTEL J. (2011). Description de verbes juridiques au moyen de la sémantique des cadres. In *TOTH*.

PITLER E. & NENKOVA A. (2008). Revisiting readability : A unified framework for predicting text quality. In *EMNLP*, p. 186–195, Waikiki,Honolulu,Hawaii.

REY A. (1979). *La terminologie : noms et notions*, In P. UNIVERSITAIRES DE FRANCE, Ed., *"Que sais-je ?"*.

RUPPENHOFER J., ELLSWORTH M., PETRUCK M., JOHNSON C. & SCHEFFCZYK J. (2006). *FrameNet II Extended Theory and Practice*. Berkeley, California : International Computer Science Institute. Distributed with the FrameNet data.

SANDERSON M. (1994). Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 142–151, New York, NY, USA.

SCHÜTZE H. & PEDERSEN J. (1995). Information retrieval based on word senses. In *Proceedings of the 4th Annual symposium on document analysis and information retrieval*, p. 161–175, Las Vegas.

SPECIA L., DAS GRAÇAS V NUNES M. & STEVENSON M. (2005). Exploiting rules for word sense disambiguation in machine translation.

STEVENSON M. & GUO Y. (2010). Disambiguation of ambiguous biomedical terms using examples generated from the umls metathesaurus. *Journal of Biomedical Informatics*, **43**, 762–773.

SURDEANU M., HARABAGIU S., WILLIAMS J. & AARSETH P. (2003). Using predicate-argument structures for information extraction. In A. FOR COMPUTATIONAL LINGUISTICS, Ed., *Proceedings of the ACL*, p. 8–15, Sapporo, Japan.

SWIER R. & STEVENSON S. (2004). Unsupervised semantic role labelling. In *EMNLP*.

TAPAS K. & ORR D. (2009). Predicting the readability of short web summaries. In *WSDM*, p. 202–211, Barcelona, Spain.

TATEISI Y., OHTA T. & TSUJII J. (2004). Annotation of predicate-argument structure on molecular biology text. In SPRINGER, Ed., *In Proceedings of the Workshop on the 1st International Joint Conference on Natural Language Process (IJCNLP*, Hainan Island, China.

TELLIER C. (2008). *Verbes spécialisés en corpus médicale : une méthode de description pour la rédaction d'articles terminologiques*. Thèse de doctorat, Université de Montréal.

TESNIÈRE L. (1959). *Éléments de syntaxe structurale*. Paris : Klincksieck.

THOMPSON C. & CALLAN P. (2004). A language modeling approach to predicting reading difficulty. In *HTL-NAACL*, p. 193–200.

WAGNER W., SCHMID H. & SCHULTE IM WALDE S. (2009). Verb sense disambiguation using a predicate-argument-clustering model. In *In Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts*.

WANDJI TCHAMI O., L'HOMME M. & GRABAR N. (2013). Discovering semantic frames for a contrastive study of verbs in medical corpora. In *Terminologie et intelligence artificielle (TIA)*, Villetaneuse.

WÜSTER E. (1985). *Introduction to the General Theory of Terminology and Therminological Lexicography*.

XUE N. & PALMER M. (2004). Calibrating features for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, p. 189–196, Stroudsburg, PA, USA.

YE P. & BALDWIN T. (2006). Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler. In *Australasian Language Technology Workshop*, p. 141–148, Sydney, Australia.

YOONG KEOK L. & HWEE TOU N. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, p. 41–48, Stroudsburg, PA, USA.

Réseau de neurones profond pour l'étiquetage morpho-syntaxique

Jérémie Tafforeau[1]
(1) LIF, AMU, CNRS, 163 avenue de Luminy, 13288 Marseille Cedex 9
jeremie.tafforeau@lif.amu-mrs.fr

Résumé. L'analyse syntaxique et sémantique de langages non-canoniques est principalement limitée par le manque de corpus annotés. Il est donc primordial de mettre au point des systèmes robustes capables d'allier références canoniques et non-canoniques. Les méthodes exploitant la théorie des réseaux de neurones profonds ont prouvé leur efficacité dans des domaines tels que l'imagerie ou les traitements acoustiques. Nous proposons une architecture de réseau de neurones appliquée au traitement automatique des langages naturels, et plus particulièrement à l'étiquetage morpho-syntaxique. De plus, plutôt que d'extraire des représentations empiriques d'une phrase pour les injecter dans un algorithme de classification, nous nous inspirons de récents travaux portant sur l'extraction automatique de représentations vectorielles des mots à partir de corpus non-annotés. Nous souhaitons ainsi tirer profit des propriétés de linéarité et de compositionnalité de tels plongements afin d'améliorer les performances de notre système.

Abstract. Syntactic and semantic parsing of non-canonical languages is mainly restricted by the lack of labelled data sets. It is thus essential to develop strong systems capable of combining canonical and non-canonical text corpora. Deep Learning methods proved their efficiency in domains such as imaging or acoustic process. We propose neural network architecture applied to natural languages processing. Furthermore, instead of extracting from the sentence a rich set of hand-crafted features wich are the fed to a standard classification algorithm, we drew our inspiration from recent papers about the automatic extraction of word embeddings from large unlabelled data sets. On such embeddings, we expect to benefit from linearity and compositionality properties to improve our system performances.

Mots-clés : TALN, Étiquetage morpho-syntaxique, Apprentissage Automatique, Réseau de Neurones Profond, Plongements.

Keywords: NLP, Part-of-Speech Tagging, Machine Learning, Deep Neural Network, Embeddings.

1 Introduction

Dans le cadre du projet européen SENSEI fondé sur l'étude des conversations humaines, nous allons être amené à analyser syntaxiquement et sémantiquement des corpus de langages non-canoniques tels que des transcriptions de conversations téléphoniques ou des commentaires web. Le manque de corpus annotés dans ces domaines force l'utilisation de corpus alliant références canoniques et non-canoniques. Les performances des approches statistiques sont particulièrement affectées par un aussi radical changement de contexte (Giesbrecht & Evert, 2009). Nous nous proposons de mettre au point un réseau de neurones profond égalant les résultats de références sur des données journalistiques avant de considérer, dans de futurs travaux, son adaptation aux corpus non-canoniques.

Les méthodes exploitant la théorie de l'apprentissage automatique de réseaux de neurones profonds ou *Deep Learning* ont prouvé leur robustesse sur des tâches complexes du domaines de l'imagerie et du traitement acoustique. Nous nous situons dans la lignée de récents travaux misant sur l'auto-adaptabilité de tels systèmes dans le cas de traitement d'informations partielles et bruitées (Collobert *et al.*, 2011). Nous proposons une architecture de réseau de neurones profond appliquée aux traitements automatiques des langages naturels. Afin d'éprouver notre méthodologie, nous nous fixons comme contexte expérimental l'étiquetage morpho-syntaxique ou *POS tagging* du corpus Penn TreeBank. Il s'agit en effet d'une tâche bien connue du traitement automatique de la langue (Manning, 2011) pour laquelle existent de nombreux résultats de référence.

Après quelques rappels généraux sur la théorie des réseaux de neurones, nous exposerons de récents travaux portant sur l'extraction de représentations vectorielles des mots d'un texte (Mikolov *et al.*, 2013a). Ces plongements ou *embeddings* présentent d'intéressantes propriétés de linéarité et de compositionnalité (Mikolov *et al.*, 2013b) que nous souhaitons capturer par l'apprentissage au plus profond du réseau de neurones. En résulte un gain en terme de rapidité de convergence ainsi qu'une approche plus fine des mots inconnus.

2 Réseaux de Neurones

L'approche conventionnelle afin d'étiqueter une séquence de mots est la suivante : extraire des représentations de la phrase en cours de traitement et les injecter dans un algorithme de classification, par exemple une machine à vecteurs de support (*SVM*), bien souvent à l'aide d'un *kernel trick*. Nous envisageons ici une approche radicalement différente : nos données seront aussi peu pré-traitées que possible afin de permettre à notre réseau de neurones d'en extraire, couche après couche, ses propres représentations internes qui seront entrainées par *rétro-propagation* afin de gagner en pertinence tout au long de l'apprentissage.

2.1 Architecture & Notations

Il existe un grand nombre d'architectures possibles lorsqu'il s'agit d'assembler des neurones en réseau. On pourra citer les *perceptrons multi-couches*, les *réseaux de neurones convolutionels* (Collobert *et al.*, 2011), les *machines de Boltzmann restreintes* (Fischer & Igel, 2012) ou encore d'autres types de *réseaux de neurones récurrents* (Toutanova *et al.*, 2003). Si les propriétés de tels réseaux sont connues depuis des décennies, les architectures comprenant plus de trois couches de neurones ne présentaient pas les améliorations attendues. C'est désormais le cas grâce à de récentes évolutions matérielles et algorithmiques. On parle depuis de réseaux de neurones profonds ou *deep neural networks*.

Considérons un réseau de neurones $f_\theta()$ de paramètres θ. Tout réseau de neurones en aval ou *feed-forward neural networks* à L couches peut être interprété comme la composition des fonctions $f^l_\theta()$ associées à chaque couche l :

$$f_\theta() = f^L_\theta(f^{L-1}_\theta(...f^1_\theta()...)) = f^L_\theta \circ f^{L-1}_\theta \circ ... \circ f^1_\theta()$$

En cela, un réseau de neurones peut être considéré comme un système d'approximation fonctionnelle. Dans toute la suite, nous noterons $(|input| \times |output|)$ afin de caractériser une couche neuronale ayant $|input|$ neurones d'entrée et $|output|$ neurones de sortie. Ce même formalisme permet de décrire des réseaux aux structures neuronales plus complexes, cf. FIGURE 1.

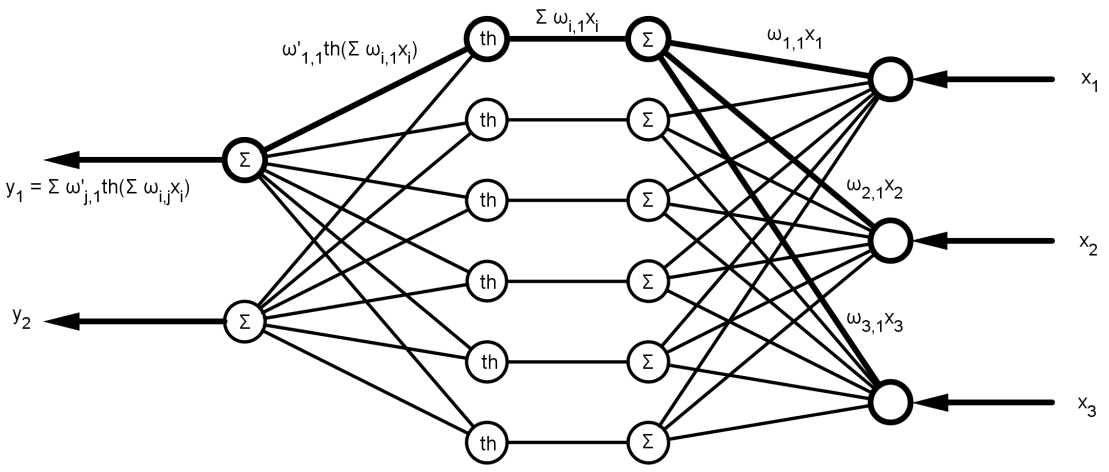


FIGURE 1 – Exemple de réseau de neurones ($3 \times 6 \times 2$)

Chaque couche neuronale réalise un plongement entre des espaces de dimensions variables en associant un vecteur de sortie à chaque vecteur d'entrée. Ainsi, compressions et dilatations des représentations internes du réseau se composent afin d'approximer au mieux une fonction pouvant permettre la classification de nos données d'apprentissage.

2.2 Structure Neuronale

Nous considérons par la suite une structure classique de réseau neuronal basée sur celle du *perceptron multi-couches* ou *multi-layer perceptron*. Elle se compose de couches linéaires réalisant une transformation affine de leur vecteur d'entrée, et de couches non-linéaires, ayant une tangente hyperbolique comme fonction d'activation, cf FIGURE 2. On alterne couches linéaires et non-linéaires afin de pouvoir approximer toute fonction, même hautement non-linéaire. En effet, si aucune non-linéarité n'est introduite dans le réseau, alors ce dernier se résumerait en un simple modèle linéaire. Enfin, la dernière couche de notre réseau compte autant de neurones que le problème considéré ne compte de classes. Chaque sortie peut ainsi être interprétée comme le score de la classe correspondante pour un exemple donnée en entrée du réseau.

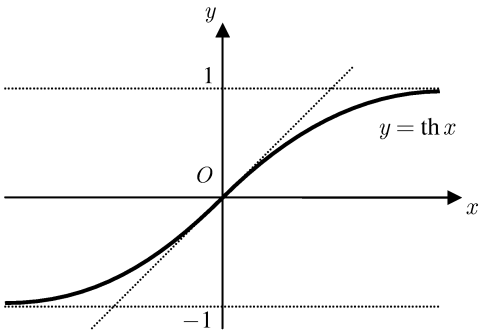


FIGURE 2 – tangente hyperbolique

2.3 Apprentissage

Tous nos réseaux ont été entrainés par maximum de vraisemblance sur l'ensemble d'apprentissage par descente de gradient stochastique. Si on note θ l'ensemble des paramètres du réseau, entrainé sur l'ensemble d'apprentissage $\mathcal{T}$, on souhaite maximiser la log-vraisemblance :

$$\theta \mapsto \sum_{(x,y)\in\mathcal{T}} log\ p(y|x,\theta),$$

où x correspond à un vecteur de *features* et y représente l'étiquette correspondante. Soit un exemple x donné en entrée, on note $[f_\theta(x)]_i$ le score associé à la i$^{\text{ième}}$ étiquette par notre réseau de paramètre θ. La probabilité $p()$ est calculé à partir des sorties du réseau de neurones au moyen d'un *softmax* :

$$p(i|x,\theta) \ = \ \frac{e^{[f_\theta(x)]_i}}{\sum_j e^{[f_\theta(x)]_j}}$$

Ce qui nous permet d'exprimer aisément la log-vraisemblance :

$$log\ p(y|x,\theta) \ = \ [f_\theta(x)]_y - log(\sum_j e^{[f_\theta(x)]_j})$$

Maximiser cette log-vraisemblance au moyen d'un gradient stochastique s'effectue en sélectionnant aléatoirement un exemple d'apprentissage (x, y) et en réalisant une descente de gradient :

$$\theta \leftarrow \theta \ + \ \lambda \frac{\delta\ log\ p(y|x,\theta)}{\delta\theta},$$

où λ est le taux d'apprentissage ou *learning rate* considéré. Cela revient à explorer l'espace des fonctions approximables par notre réseau en minimisant pas à pas l'erreur de classification.

3 Représentations vectorielles du texte pour le *POS tagging*

3.1 Le *POS tagging*, un problème de classification

Notre contexte expérimental est le *POS tagging* qui a pour but d'étiqueter chaque mots en fonction de son rôle syntaxique, par exemple, nom, adjectif, adverbe, etc. Les sections 0-18 ainsi que 19-21 du Penn TreeBank ont été respectivement utilisées comme ensemble d'apprentissage et de développement, tandis que les sections 22-24 furent réservées à la phase de test. Afin de minimiser le nombre de mots composant notre vocabulaire, nous avons modifié notre corpus afin qu'il ne comporte que des mots en minuscules. De même, toutes les séquences de chiffres formant un nombre entier ont été remplacées par la chaîne _NUMBER_ (*ex. 1 ; 1,000 ; 1,000,000*) et celles formant un nombre réel par la chaîne _REAL_ (*ex. 3.14 ; 6,378.137*).

Dataset	#phrases	#mots	#mots inconnus	#classes
Training	38.219	912.344	0	48
Development	5.527	131.768	4.467	48
Test	5.462	129.654	3.649	48

TABLE 1 – Caractéristiques du corpus Penn TreeBank

Les meilleurs systèmes pour cette tâche sont généralement entraînés sur une fenêtre de texte et servent de fondations à des algorithmes bidirectionnels apportant une information de séquence, cf.*HMM & Viterbi*. Parmis les *features* récurrentes, on trouve les *tags* contextuels i.e les étiquettes des mots suivant ou précédant le mot à étiqueter, les *n-grams* ainsi que des *features* empiriques ou linguistiquement motivées, dédiées à la gestion des mots inconnus. Comme dans la plupart des travaux nous ayant précédés, nous abordons le *POS Tagging* comme un problème de classification. Mettant de côté l'aspect séquentiel, nous souhaitons attribuer indépendamment à chaque mot d'une phrase son étiquette morpho-syntaxique correspondante. Il nous est donc nécessaire de fixer une représentation vectorielle propre à chaque mot qui constitueront les entrées de notre classifieur.

Model	All Words Accuracy	Unknown Words Accuracy
Hidden Markov model (Brants, 2000)	96,46%	85,86%
Support machine vectors (Giménez & Márquez, 2004)	97,16%	89,01%
Maximum entropy cyclic dependency network (Manning, 2011)	97,32%	90,79%

TABLE 2 – Résultats de référence pour le *POS tagging* sur le Penn TreeBank

3.2 Représentations en vecteur d'index lexical

Par soucis d'efficacité, les mots sont fournies au réseau sous forme d'indices tirés d'un lexique. C'est pourquoi la première étape de notre modèle est d'établir un lien entre indices et représentations vectorielles au moyen d'une table d'association. Soit $|V|$ la taille de notre vocabulaire. Nous considérons dans un premier temps comme représentation du mot w d'indice lexical n_w le vecteur *"creux"* :

$$W = \begin{cases} W_{n_w} & = & 1 \\ W_i & = & 0 \quad \forall i \neq n_w \end{cases}$$

Ces représentations vectorielles doivent par la suite être combinées afin d'étiqueter un par un chaque mot composant la phrase en cours de traitement. Prenons comme hypothèse que seuls les mots voisins du mot considéré influent sur son étiquetage. Soit un mot w a étiqueter, on considère une fenêtre de taille fixe d_{win} centrée sur w. En associant à chaque mot composant cette fenêtre sa représentation vectorielle, nous obtenons une matrice de taille $d_{win} \times |V|$, cf FIGURE 3. Cette matrice peut être vue comme un vecteur de dimension $d_{win}.|V|$ par concaténation des colonnes, qui peut être injecté en entrée du réseau de neurones.

w_1	w_2	w_3	
0	0	0	0
.	.	.	
.	.	.	
1	0	0	n_{w_1}
.	.	.	
.	.	.	
0	0	1	n_{w_3}
.	.	.	
.	.	.	
0	1	0	n_{w_2}
.	.	.	
.	.	.	
0	0	0	$\|V\|$

FIGURE 3 – Représentation vectorielle associée à un triplet de mot (w_1, w_2, w_3)

3.3 Extraction automatique de représentations vectorielles

Il est possible d'extraire à partir de données non-annotées des plongements, ou *embeddings*, de plus petite dimension afin d'obtenir de meilleurs représentations vectorielles de nos mots. Dans nos expériences, nous nous basons sur l'approche WORD2VEC (Mikolov *et al.*, 2013a). Cette dernière consiste à entraîner un réseau de neurones linéaire i.e sans couche cachée non-linéaire. La matrice des poids de la couche linéaire peut ainsi être interprétée comme une projection linéaire, permettant de passer de l'espace des mots à un espace de dimension réduite. Deux heuristiques d'apprentissage ont été proposées afin de calibrer ce plongement, cf. FIGURE 4.

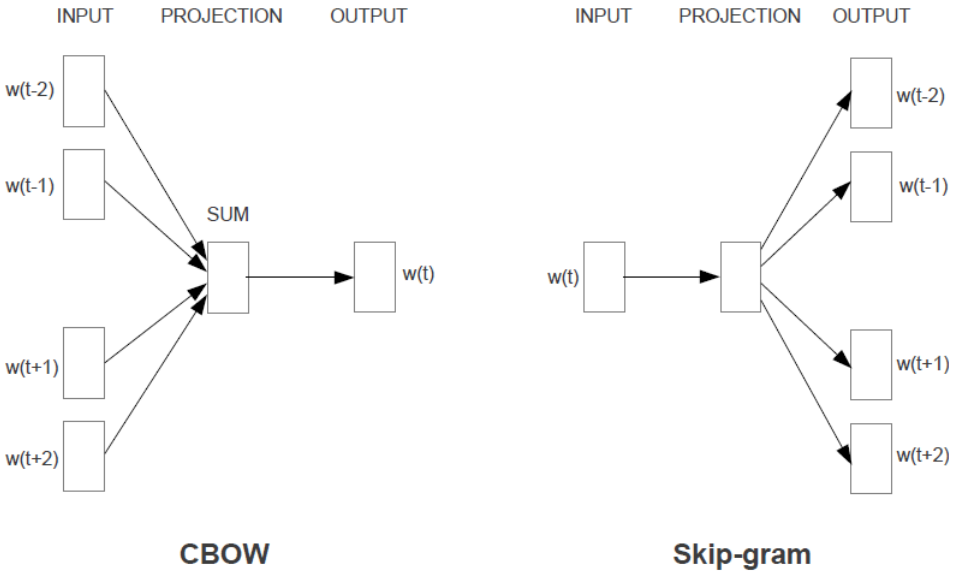


FIGURE 4 – WORD2VEC Architectures

L'approche *Continuous Bag of Words* (CBOW) consiste à entraîner le réseau à prédire un mot à partir de son contexte. À l'inverse, l'approche *Skip-gram* se propose de prédire les mots encadrant un mot donné en entrée. Une fois l'un de ses deux entraînements arrivé à son terme, les plongements sont extraits des poids de la couche linéaire du réseau. La dimensions n de ces plongements est ainsi librement paramétrable car totalement déterminée par le nombre de neurones composant cette couche de projection. Il s'agit donc d'un hyper-paramètre de notre système.

Ainsi, deux mots proches dans ce nouvel espace de représentations sont syntaxiquement et/ou sémantiquement proches car c'est la promiscuité de leur contexte respectif qui les a rapprochés dans ce nouvel espace. Il est bon de noter que ces dernières présentent également des propriétés de linéarité et de compositionalité (Mikolov *et al.*, 2013b) alors qu'aucune contrainte ne fut fixée en ce sens au cours de l'apprentissage. Notre architecture est capable de tirer partie de telles *features*.

4 Architectures proposées

L'ensemble des architectures réseaux proposées ont été réalisées à l'aide de la librairie *Torch7* car en plus de proposer une implémentation intuitive, elle a montré de bonnes performances dans les tests de (Collobert *et al.*, 2012).

4.1 Perceptron Multi-Couche Naïf

Il s'agit d'une implémentation standard d'un *Multi-Layer Perceptron* (*MLP*). Les entrées de ce réseau sont les concaténations de vecteurs *"creux"* d'index lexical. La structure de la couche cachée non-linéaire est un hyper-paramètre du réseau. Dans la suite et par soucis de clarté, nous considèrerons qu'elle ne se compose que d'une seule couche non-linéaire constituée de $d_{win} * n$ neurones tandis qu'elle sera bien plus complexe lors de l'exposé de nos expérimentations. En vu de permettre au réseau d'appréhender les mots de l'ensemble de test absents de l'ensemble d'apprentissage (en d'autres termes les mots inconnus), nous avons considéré les mots n'ayant qu'une seule occurrence au sein de notre corpus d'apprentissage comme inconnus. Ces derniers ont ainsi été remplacés dans notre corpus par la chaîne _UNK_.

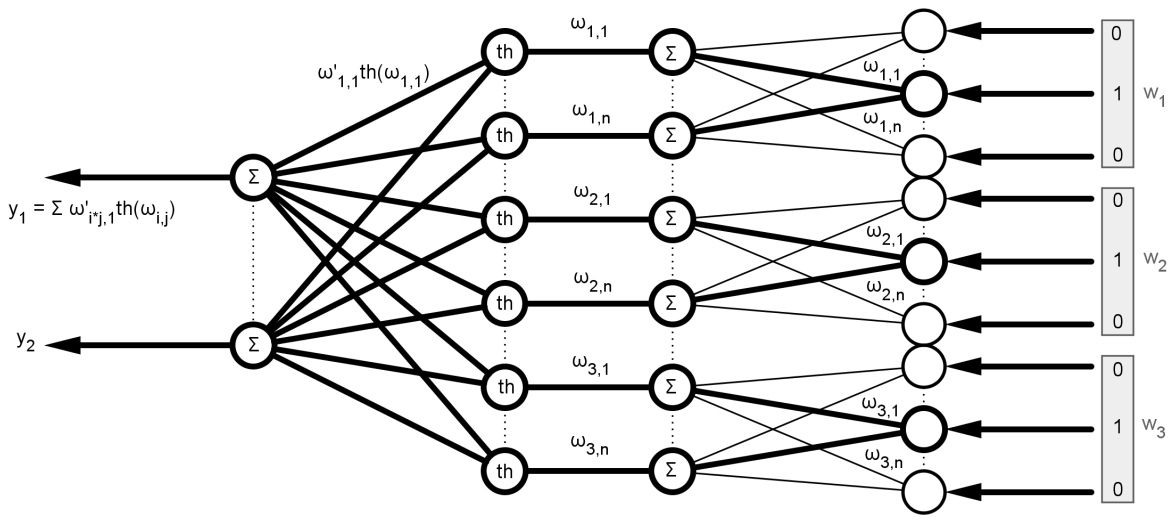


FIGURE 5 – Exemple de *MLP* à une couche cachée dans le cas d'une classification à 2 classes et d'une fenêtre de taille 3. On applique un *softmax* à nos sorties afin d'obtenir des scores assimilables à des probabilités . Les neurones $\sum$ réalisent une combinaison linéaire pondérée de leurs entrées tandis que les neurones *th* forment notre couche non-linéaire appliquant une tangente hyperbolique. On a représenté en gras les connexions réellement sollicitées par l'exemple d'apprentissage considéré.

On remarque immédiatement que la taille de la couche d'entrée, constituée de $d_{win} * |V|$ neurones, est directement proportionnelle au nombre de mots composant notre vocabulaire. Pourtant, pour un exemple d'apprentissage donné, très peu de connexions sont réellement sollicitées au sein de cette couche puisque seules d_{win} dimensions d'entrée sont effectivement actives i.e de valeur non-nulle. Dans la pratique, cette implémentation n'est pas efficace du fait du trop grand nombre de connexions qui la composent.

4.2 Initialisations des poids de la couche d'entrée par nos plongements

Nous souhaitons initialiser les poids de notre couche d'entrée à l'aide des plongements extraits, grâce à l'approche CBOW de WORD2VEC, d'un important corpus non-annoté tiré de *Wikipédia*, cf TABLE 3. Nous espérons obtenir ainsi une meilleur

convergence de notre modèle. De plus, nous nous proposons de réaliser une couche qui minimise le coût d'une descente de gradient dans le cas de vecteurs d'entrée *"creux"* afin de diminuer la complexité de la phase d'apprentissage. On expose pour ce faire deux implémentations distinctes.

#phrases	#mots	$\|V\|$
84.805.805	1.709.107.335	1.556.817

TABLE 3 – Caractéristiques du corpus d'apprentissage de WORD2VEC

4.2.1 Initialisation Statique (*SMLP*)

Nous considérons dans un premier temps un réseau simulant une couche d'entrée statique dont les poids ont été initialisés par nos plongements, cf. FIGURE 6. La couche d'entrée initiale y a été remplacée par un pré-traitement des entrées qui sont désormais les concaténations des plongements associés à chaque mot composant la fenêtre initialement injectée en entrée. Ce réseau simule un *perceptron multi-couches* à la différence que la *rétro-propagation* n'atteint pas les poids de la couche d'entrée qui restent par conséquent constants tout au long de l'apprentissage.

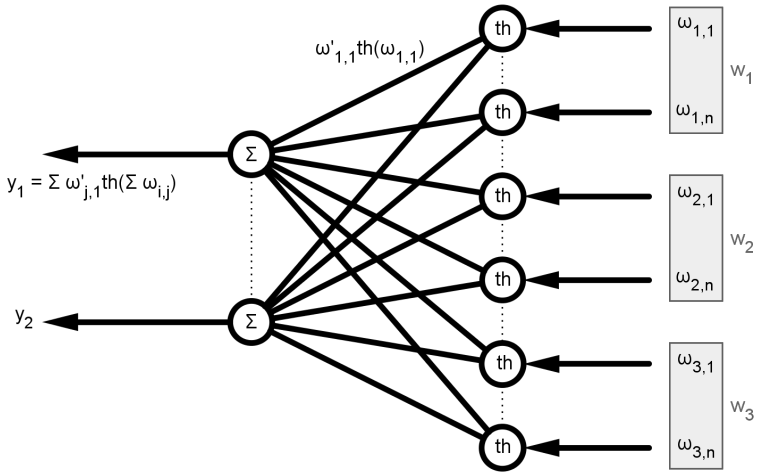


FIGURE 6 – Architecture *SMLP* équivalente au *MLP* décrit en FIGURE 5. Chaque mot $w_{i\in[1..3]}$ composant notre exemple d'apprentissage est désormais représenté comme un vecteur ω_i de dimension n. C'est la concaténation de ces plongements qu'on injecte en entrée du réseau. On obtient ainsi un réseau au comportement analogue au précédent, s'abstrayant des connexions inactives au prix du gel des paramètres de la couche d'entrée initiale.

4.2.2 Initialisation Dynamique (*DMLP*)

Nous considérons désormais un réseau permettant à la *rétro-propagation* de raffiner nos plongements, cf. FIGURE 7. À chaque mot w de notre dictionnaire, nous associons, dans une bibliothèque L de couches linéaires, une couche L_w de structure $(1 \times n)$ dont les n poids représentent nos plongements. Lorsqu'on souhaite injecter une fenêtre de mots (w_1,w_2,w_3) dans notre réseau, on remplace la couche d'entrée par la concaténation des couches linéaires associées L_{w_1}, L_{w_2} et L_{w_3}. La descente de gradient stochastique se propage ainsi jusque dans notre nouvelle couche d'entrée, modifiant par la même les poids des couches de notre bibliothèque et donc nos plongements.

Les entrées sont forcées à 1 et représentent les dimensions actives des vecteurs d'index lexical initiaux. De cette manière, la couche d'entrée simule les connexions sollicitées pour un mot donné et s'abstrait ainsi de toutes celles inactives dans le cas de vecteur *"creux"*. Cette implémentation est réalisée à l'aide de partages dynamiques de paramètres afin de manipuler nos couches d'entrées interchangeables. Cette architecture permet un raffinage des plongements initiaux tout au long de l'apprentissage en fonction de la tâche considérée tout en minimisant la taille de la structure neuronale nécessaire et donc le coût d'une descente de gradient.

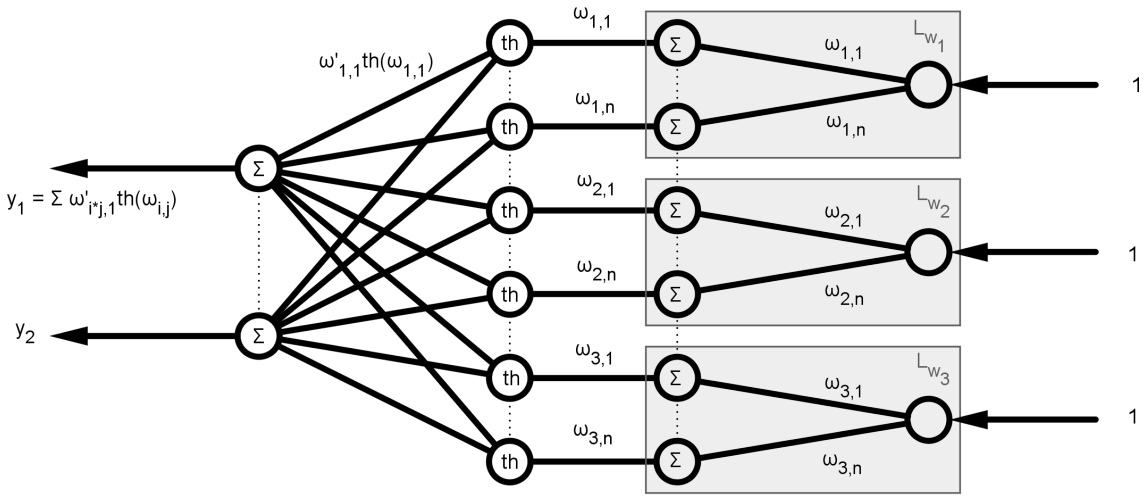


FIGURE 7 – Architecture *DMLP* équivalente au *MLP* décrit en FIGURE 5. Chaque mot $w_{i\in[1..3]}$ composant notre exemple d'apprentissage est désormais représenté par une couche linéaire L_{w_i}. Le vecteur de poids ω_i de cette couche étant le plongement associé au mot w_i considéré, ces derniers deviennent des paramètres du réseau. Lorsqu'on souhaite soumettre au réseau un nouvel exemple d'apprentissage $w'_{i\in[1..3]}$, il suffit de remplacer les vecteurs poids ω_i par les plongements ω'_i associés. Par soucis d'efficacité, cette opération est réalisée à l'aide d'un partage dynamique de paramètres entre notre couche d'entrée et notre bibliothèque L de couches linéaires contenant nos plongements raffinés au cœur de leur vecteur poids.

4.3 Utilisation de connaissances *a priori*

Nous désirons utiliser les connaissances *a priori* pouvant être extraites de notre corpus d'apprentissage afin de guider notre classification. Soit un mot w, nous nous concentrons sur les classes observées à chaque occurrence de w pour obtenir un *filtre* à appliquer aux prédictions. Ainsi, lors de la comparaison des scores des différentes classes, seules les classes observées sont mise en compétition. Dans le cas où w serait absent du corpus d'apprentissage, nous considérons l'ensemble des classes comme ayant été observées, ce qui revient à ne pas filtrer les prédictions de notre réseau de neurones.

5 Expérimentations & Résultats

Afin d'obtenir des performances directement comparables aux nôtres, nous avons réalisé une évaluation d'une implémentation des *Conditional random fields* (Nasr *et al.*, 2014) avec différents ensembles de *features*. Habituellement, à chaque mot sont associés :

– les *n-grams* le contenant pour n allant de 1 à 5 ;
– les *bigrams* d'étiquettes *POS* apportant une information de séquence ;
– ses attributs morphologiques (préfixes, suffixes ainsi que des *features* de présence de majuscules ou de symboles).

Model	All Words Accuracy	Unknown Words Accuracy
Conditional random field (3-gram uniquement)	95,55%	66,02%
Conditional random field (3-gram et toutes les *features*)	97,30%	87,59%
Conditional random field (5-gram et toutes les *features*)	97,43%	88,36%

TABLE 4 – Évaluations d'une implémentation des *CRF* pour le *POS tagging* (Nasr *et al.*, 2014) sous restrictions

Sur la première évaluation, les *features* sont restreintes aux *3-grams*. Ce système est par conséquent comparable au notre en terme d'information à disposition. La dernière, réalisée quant à elle sans limitations, égale nos résultats de références. On peut ainsi observer le gain notable apporté par l'ajout de *features* de séquence et de morphologie, notamment sur la gestion de mots inconnus.

d_{win}	n	Structure des couches cachées	*Learning Rate*
3 et 5	300	$((d_{win} * n) \times (d_{win} * 100))$	0.01

TABLE 5 – jeux d'hyper-paramètres considérés, où d_{win} et n représentent respectivement la taille des fenêtres de mots envoyées en entrée du réseau et la dimension des *embeddings* extraits par WORD2VEC. Nous avons choisi d'utiliser deux couches cachées afin de permettre au réseau d'extraire des représentations internes issues de la combinaison des $features$ propres à chaque mot de la fenêtre.

Dans l'ensemble, nos résultats expérimentaux sont significativement proches de nos références, cf. TABLE 2. L'approche *DMLP* obtient globalement de meilleurs résultats car elle permet un raffinage des plongements guidé par la tâche. Elle présente également de meilleures performances que le *CRF* dépourvu d'informations de séquence et morphologiques.

Toutefois, nos résultats se trouvent dans la plage basse des résultats de références car n'a été utilisé aucune information de séquence ou d'optimisation globale visant à prendre toute décision en se servant explicitement d'une décision prise antérieurement. En effet, dans nos approches les mots sont étiquetés un par un, indépendamment les uns des autres, sans utiliser, par exemple, les $tags$ prédits des mots précédents ou suivants. L'évaluation du *MLP* naïf quant à elle n'a pas été mené à son terme à cause du trop important temps de traitement qu'elle nécessiterait. Elle permet cependant une estimation de l'efficience de notre approche.

Model	All Words Accuracy	Unknown Words Accuracy	ms/exemple
3win-MLP	na	na	350ms
5win-MLP	na	na	973ms
3win-SMLP	95,87%	84,14%	1,09ms
5win-SMLP	96,09%	84,64%	3,11ms
3win-DMLP	96,51%	85,26%	1,19ms
5win-DMLP	96,79%	86,20%	3,33ms

TABLE 6 – Résultats expérimentaux

6 Conclusion

Issue de la théorie des réseaux de neurones profonds, l'approche que nous proposons égale nos résultats de référence sur le Penn TreeBank. Cette dernière se base sur l'extraction automatique de représentations vectorielles des mots à partir d'importants corpus non-annotés ainsi que sur les théories du *Deep Learning* permettant de raffiner, couche après couche, ces représentations internes afin qu'elles gagnent en pertinence tout au long de l'apprentissage.

Les performances prometteuses observées sur le *POS Tagging* confirment le bien-fondé de notre méthodologie n'utilisant pourtant aucune information de séquence, contrairement à nos résultats de référence. De plus, la comparaison des performances que nous avons obtenues à celles d'un *CRF* n'utilisant également aucune information de séquence met en évidence le potentiel de notre approche. En effet, il n'y a qu'un pas pour doter notre architecture d'une structure récurrente afin de lui permettre d'étiqueter simultanément tous les mots d'une même phrase. C'est ce que nous considèrerons dans de futurs travaux au même titre que l'ajout de *feature* morphologiques tels que les suffixes ou les préfixes.

C'est donc confiants que notre attention se porte désormais sur des tâches plus complexes du traitement automatique des langages naturels, telles que les traitements joints et la détection de disfluences dans des corpus oraux non-canoniques, afin de tester la robustesse de notre modèle au changement de domaine.

Remerciements

Ces travaux de recherche ont été financés en partie par l'Union Européenne à travers le projet SENSEI [1] (FP7/2007-2013 - n° 610916 – SENSEI).

1. http ://www.sensei-conversation.eu

Références

BRANTS T. (2000). TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the 6th Applied Natural Language Processing*, ANLC '00, p. 224–231 : Association for Computational Linguistics.

COLLOBERT R., KAVUKCUOGLU K. & FARABET C. (2012). Implementing Neural Networks Efficiently. In G. MONTAVON, G. ORR & K.-R. MULLER, Eds., *Neural Networks : Tricks of the Trade*.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural Language Processing (Almost) from Scratch. In *the Journal of Machine Learning Research 12*, p. 2461–2505.

FISCHER A. & IGEL C. (2012). An introduction to restricted Boltzmann machines. In L. ALVAREZ, M. MEJAIL, L. GOMEZ, & J. JACOBO, Eds., *Proceedings of the 17th Iberoamerican Congress on Pattern Recognition, volume 7441 of LNCS*, p. 14–36.

GIESBRECHT E. & EVERT S. (2009). Part-of-Speech (POS) Tagging - a solved task ? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*, p. 27–35, San Sebastian : Elhuyar Fundazioa.

GIMÉNEZ J. & MÁRQUEZ L. (2004). SVMTool : A general POS tagger generator based on Support Vector Machines. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, p. 43–46.

MANNING C. (2011). Part-of-Speech Tagging from 97% to 100% : Is it time for some linguistics ? In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing, 12th International Conference (CICLing), Part I. Lecture Notes in Computer Science 6608*, p. 171–189.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient Estimation of Word Representations in Vector Space. volume abs/1301.3781.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Deep Learning Workshop at the 2013 Conference on Neural Information Processing Systems (NIPS)*.

NASR A., BECHET F., FAVRE B., BAZILLON T., DEULOFEU J. & VALLI A. (2014). Automatically enriching spoken corpora with syntactic information for linguistic studies. In *LREC*.

TOUTANOVA K., MANNING C., KLEIN D. & SINGER Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAACL*.

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations : a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 384–394 : Association for Computational Linguistics.

Extraction terminologique : vers la minimisation de ressources

Yuliya Korenchuk[1, 2]
(1) LiLPa (Linguistique, Langues, Parole), EA 1339, Université de Strasbourg
(2) Rebuz SAS, Strasbourg
korenchuk@unistra.fr

Résumé. Cet article présente une méthode ayant pour objectif de minimiser l'apport extérieur nécessaire à la tâche d'extraction terminologique (ET) et de rendre cette tâche moins dépendante de la langue. Pour cela, la méthode prévoit des ressources morphologiques et morphosyntaxiques simplifiées construites directement à partir d'un corpus lemmatisé. Ces ressources endogènes servent à la création d'un système de filtres qui affinent les calculs statistiques et à la génération de patrons pour l'identification de candidats termes polylexicaux. La méthode a été testée sur deux corpus comparables en chimie et en télécommunication, en français et en anglais. La précision observée sur les 100 premiers candidats termes monolexicaux fluctue entre 71% et 87% pour le français et entre 44 % et 69 % en anglais ; celle des candidats termes polylexicaux s'élève à 69-78 % en français et 69-85 % en anglais en fonction du domaine.

Abstract. The article presents the method which aims to minimize the use of external resources for the terminology extraction task and to make this task less langage dependent. For that purpose, the method builds simplified morphological and morphosyntactic resources directly from a lemmatized corpus. These endogenous resources are used both in filters, which refine the statistical calculations, and in patterns for polylexical terms identification. The method was tested on two comparable corpora in chemistry and in telecommunication in French and in English. The precision observed on the first 100 monolexical terms fluctuates between 71% and 87% for French and between 44% and 69% in English ; for polylexical terms the precision was 69-78% in French and 69-85% in English depending on the domain.

Mots-clés : extraction terminologique, ressources endogènes, apprentissage automatique.

Keywords: terminology extraction, endogenous resources, machine learning.

1 Introduction

L'extraction terminologique (ET) a de nombreuses applications dans la construction de ressources linguistiques et sémantiques (par exemple, des glossaires et des bases terminologiques, des ontologies spécialisées, etc.). Ses résultats peuvent servir directement à des utilisateurs humains (des traducteurs ou des terminologues) ou bien être exploités par des moteurs de recherche ou des systèmes de traduction automatique. Les méthodes et les outils élaborés pour l'ET sont très variés. Leur diversité se manifeste à plusieurs niveaux, tels que l'approche plus ou moins dépendante de la langue, les exigences par rapport au choix du corpus (sa taille, son type, un prétraitement spécifique, etc.) et d'autres ressources nécessaires. Du point de vue de l'application industrielle, chacun de ces aspects a un coût, et les coûts de préparation et d'adaptation des ressources peuvent s'avérer assez importants.

Une des principales motivations de la méthode proposée est d'identifier automatiquement des régularités dans les textes analysés pour créer des ressources endogènes à partir du corpus. Cette approche est :
– applicable à plusieurs langues ;
– indépendante du domaine de spécialité ;
– capable d'extraire des candidats termes mono et polylexicaux ;
– capable d'apprendre les informations nécessaires pour l'analyse à partir du corpus.
Cet objectif a été atteint par le biais de patrons morphologiques et morphosyntaxiques endogènes qui résultent d'une approche multi-étapes à base de n-grammes et qui sont générés directement au cours de l'analyse à partir du corpus.

Les résultats obtenus ont confirmé la pertinence de la méthode pour sa tâche principale. En outre, ses caractéristiques permettent d'envisager l'application de la méthode à l'enrichissement multilingue des ontologies de spécialité.

La première partie de cet article propose une brève description de méthodes existantes et de ressources utilisées en ET. La

deuxième partie décrit les corpus de test et la méthode proposée. Ensuite viennent l'évaluation des résultats pour les deux langues et la conclusion.

2 État de l'art

L'ET connaît son essor dans les années 1990 et de nombreux systèmes et méthodes apparaissent presque simultanément à cette période. Depuis, le domaine est en plein développement. Les systèmes d'ET cherchent à extraire des unités lexicales simples ou complexes qui sont susceptibles d'être des termes. Plusieurs traits peuvent être pris en compte (la fréquence, la structure, le contexte ou la comparaison du corpus analysé avec un corpus de référence).

La notion du candidat terme est utilisée en extraction terminologique pour désigner les unités lexicales repérées par les systèmes automatiques (Drouin et Langlais, 2006). Cette notion sécurise le travail avec les systèmes d'ET, car le fait d'insister sur le caractère incomplet de résultats du traitement incite à valider ces résultats avant de les réutiliser.

La palette des outils disponibles pour le français comprend des systèmes récents comme TTC TermSuite (Morin et Daille, 2012) et YaTeA (Aubin et Hamon, 2006), ainsi que leurs prédécesseurs : ACABIT (Daille, 1996, 2003; Morin et Daille, 2006), FASTR (Jacquemin, 1997) et LEXTER (Bourigault *et al.*, 1996). Ces outils sont basés sur des méthodes et des ressources différentes et nous avons choisi d'en présenter les plus pertinentes par rapport à notre projet.

2.1 Méthodes d'extraction terminologique

En général, les classifications des méthodes d'ET (Bernhard, 2006) distinguent les méthodes basées sur des mesures statistiques, les méthodes basées sur des éléments linguistiques et les méthodes mixtes.

Les mesures statistiques, telles que la fréquence absolue ou pondérée, TFxIDF, l'information mutuelle ou l'indice de Jaccard mettent en évidence des unités mono ou polylexicales caractéristiques d'un document ou d'un corpus. Ces méthodes sont indépendantes de la langue et nécessitent uniquement un ou plusieurs corpus. Toutefois, leurs résultats sont meilleurs sur des corpus de taille importante. Certaines méthodes permettent d'évaluer le potentiel terminologique de candidats termes (Drouin et Langlais, 2006), ce qui peut servir de filtre pour d'autres méthodes.

La deuxième catégorie de méthodes fait appel à des éléments linguistiques parmi lesquels on peut distinguer deux groupes : les patrons morphosyntaxiques et les formants de langues classiques. L'identification des termes de certains domaines peut utiliser des traits morphologiques spécifiques à leurs nomenclatures. Tel est le cas de la nomenclature en chimie, en physique, en biologie ou encore en médecine. Cette méthode est adaptée pour des termes qui contiennent des composants savants, c'est-à-dire, des morphèmes provenant du grec ou du latin. En effet, ces morphèmes sont très productifs ; par exemple, le radical *AZOT* est à base des termes suivants : *azote, azotate, azoté, azoteux, azotique, azotite, azotémie, azothydrique, azoture, azoturie, etc.* Appliquée aux domaines listés ci-dessus, cette méthode est efficace à 87,70 % selon Estopà *et al.* (2000).

Le deuxième groupe s'approche des méthodes mixtes, car les méthodes à base de patrons morphosyntaxiques ont recours à des calculs statistiques pour affiner leurs résultats. Cette approche est utilisée dans les méthodes de (Daille, 2003; Morin et Daille, 2012), de Bourigault (Bourigault et Fabre, 2000; Bourigault, 2002) et la méthode C-value/NC-value de Frantzi *et al.* (2000). Orobinska *et al.* (2013) propose une approche plus souple en apprenant les patrons morphosyntaxiques caractéristiques à partir du corpus étiqueté. L'un des avantages des patrons morphosyntaxiques est la possibilité d'identifier la variation au sein des candidats termes polylexicaux et de définir des règles de transformation afin d'améliorer l'organisation des résultats (Bourigault et Jacquemin, 1999).

Dans notre travail, nous reprenons certains éléments des deux méthodes mixtes qui ne nécessitent pas de prétraitement de corpus. La première méthode proposée par Vergne (2003, 2004, 2005) permet d'annoter le corpus par des mots informatifs et vides pour ensuite en extraire des termes à structure contrôlée. Nous allons détailler cette méthode dans la partie 3.2. La deuxième méthode, qui s'appelle ANA (Apprentissage naturel Automatique), est développée par Enguehard (1993). A l'étape de la familiarisation, le logiciel parcourt le corpus et en extrait les listes des mots fonctionnels (*a, alors, après, etc.*), des mots fortement liés (*de la, de l', etc.*) et des mots de schéma (*en, de, du, d, de la, des*). Ces deux derniers groupes servent de liaison dans les candidats termes polylexicaux et permettent de les identifier dans le texte. Ensuite, le programme analyse le corpus de manière itérative, en identifiant les mots qui apparaissent fréquemment avec les termes de bootstrap (un ensemble de quelques termes du domaine prédéfinis manuellement dont il est question dans le corpus de

textes), et en les rajoutant dans le bootstrap.

Une autre approche intéressante consiste à utiliser le lexique transdisciplinaire comme indicateur et délimiteur des unités terminologiques (Jacquey *et al.*, 2013; Tutin, 2007). En effet, des mots comme *concept, méthode, technologie, analyser, etc.* servent souvent à introduire des termes d'un domaine sans appartenir à ce domaine. La détection de ces mots ou expressions et leur classification facilitent l'ET. Jacquey *et al.* (2013) combinent le lexique transdisciplinaire projeté sur le corpus avec l'analyse syntaxique pour identifier les cas où une unité du lexique introduit un candidat terme. Leur hypothèse a été confirmée dans environs 74 % des cas (Jacquey *et al.*, 2013).

La plupart des méthodes combinent plusieurs paramètres pour augmenter leur efficacité. Cela s'explique en partie par les particularités des termes (Estopà *et al.*, 2000) ou par les limites internes de chaque approche. Nous allons maintenant décrire les avantages et les limites des différents types de ressources employées par les méthodes d'ET.

2.2 Ressources pour l'extraction terminologique

Un corpus de domaine est la première ressource indispensable à l'ET. Il conditionne le choix de la méthode, car certaines méthodes se montrent plus efficaces sur des corpus de taille importante. Or, il n'est pas toujours facile de trouver un corpus suffisamment grand pour un domaine donné, surtout pour des langues possédant moins de ressources que le français ou l'anglais. Pour les corpus parallèles, le choix est davantage limité. Il est donc intéressant de pallier la contrainte de la taille du corpus soit par la minimisation de l'importance des opérations statistiques (en démultipliant les traits linguistiques pris en compte), soit par l'emploi de corpus de référence pour augmenter le contraste entre les termes et les mots de la langue générale.

Les ressources morphosyntaxiques, notamment les patrons d'identification de candidats polylexicaux, sont développées pour chaque méthode à part. Elles dépendent de l'étiqueteur mophosyntaxique disponible et reposent sur les résultats de ce dernier. Cependant, l'étiqueteur risque de perdre en précision lorsqu'il s'agit d'un domaine très spécifique, comme la médecine ou la chimie. Les méthodes basées sur ces patrons sont performantes, mais limitées à des langues qui ont déjà un étiqueteur disponible.

En ce qui concerne les ressources morphologiques, des listes de formants grecs et latins sont disponibles sur Internet. Cependant, elles nécessitent une vérification manuelle et une adaptation de format avant d'être introduites dans le programme. Les méthodes qui comparent des fréquences de formants dans un corpus de spécialité et dans un corpus de langue générale sont applicables pour affiner les listes existantes (Bernhard, 2006). La limite évidente des formants est leurs productivité dans le domaine analysé : certains domaines ont très peu de recours à ces morphèmes, en préférant des emprunts ou la néologie.

Enfin, des ressources lexicales, comme des anti-dictionnaires (*stop words lists*) ou le lexique transdisciplinaire (Tutin, 2007; Jacquey *et al.*, 2013), sont évidemment liées à la langue pour laquelle elles sont développées. De ce point de vue, les travaux de Drouin (2007), qui portent sur l'identification automatique du lexique transdisciplinaire, sont très intéressants. Une autre méthode permettant d'éviter l'utilisation d'un anti-dictionnaire est l'annotation de Vergne (2003, 2004, 2005). En effet, les mots étiquetés comme vides sont dans la majorité des cas éligibles pour un anti-dictionnaire.

Lever les contraintes existantes dans l'ET signifie minimiser les ressources nécessaires pour cette tâche tout en améliorant les résultats. Nous allons présenter une méthode qui tend à réunir les aspects forts des approches décrites ci-dessus tout en minimisant les ressources nécessaires pour arriver à des résultats satisfaisants.

3 Méthodologie

Les méthodes qui se montrent efficaces nécessitent des ressources externes assez importantes, comme par exemple les listes de formants savants, les patrons morphosyntaxiques, les dictionnaires de référence, etc. Les systèmes basés sur de telles méthodes sont dépendants vis-à-vis de la langue et de ces ressources. Or, ces ressources ne font que refléter des régularités linguistiques. De ce point de vue, il doit être possible de générer ces ressources à partir du corpus-même.

Ainsi, l'objectif principal de la présente méthode est de contourner la nécessité de fournir des ressources morphologiques ou morphosyntaxiques pour obtenir des candidats termes mono et polylexicaux en se basant sur les données les plus fiables obtenues à chaque étape.

3.1 Présentation des corpus

Pour notre projet, nous avons choisi deux corpus comparables bilingues (FR/EN) pour deux domaines bien distincts : la chimie et les télécommunications. Le corpus en chimie est composé de mémoires et de thèses en chimie des métaux sélectionnés manuellement à l'aide des options de recherche avancées sur Google Scholar. Tous les textes sont convertis en UTF-8 et les fragments en langues étrangères ont été éliminés.

Le corpus en télécommunication, plus précisément en technologies mobiles, vient du projet TTC [1]. Nous n'avons donc pas eu à le construire, mais nous avons pu constater qu'il contient essentiellement des documents techniques.

La différence fondamentale entre les deux domaines permet d'apporter un regard critique sur notre méthode, car les résultats peuvent varier en fonction de la nomenclature du domaine (formants savants, emprunts, etc.). Toutes les caractéristiques fournies dans la table 1 sont quantifiées après la tokenisation du corpus par le script du projet Europarl [2]. Chaque paire de corpus est assez équilibrée. Nous pouvons donc nous attendre à des résultats fiables et comparables.

Corpus/ Langue	Chimie				Télécommunication			
	Taille	*Tokens*	*Types*	*Lemmes*	*Taille*	*Tokens*	*Types*	*Lemmes*
Français	4,13 Mo	841 843	49 948	28 717	2,94 Mo	526 240	24 419	14 758
Anglais	3,49 Mo	713 369	44 684	27 274	1,95 Mo	349 656	25 425	17 186

TABLE 1 – Caractéristiques des corpus

3.2 Prétraitement et annotation automatique du corpus

Le système fait appel au script de tokenisation du projet Europarl [3] et à l'étiqueteur morphosyntaxique TreeTagger (Schmid, 1994) qui est utilisé comme lemmatiseur. La suite des traitements est effectuée sur le corpus lemmatisé.

Le corpus est traité par l'algorithme d'annotation par des mots informatifs et non-informatifs de Vergne (2003, 2004, 2005). Cet algorithme combine deux propriétés, la fréquence et la longueur de mots, pour détecter les mots vides à partir de corpus brut indépendamment de la langue de ce dernier. Selon la méthode de Vergne, le mot informatif est sélectionné selon trois critères :

1. Longueur importante
2. Fréquence réduite
3. Entourage par des mots plus courts et plus fréquents

Le reste de mots est considéré comme des mots vides. Cependant, l'application de la méthode, telle qu'elle est présentée par Vergne, rapproche les mots informatifs des candidats termes. Or, les exemples de résultats obtenus pour cette méthode sur des corpus constitués d'articles de presse (Vergne, 2005) contiennent peu de candidats termes tels que nous pouvons exploiter dans un système de gestion de la terminologie. Par exemple, dans *une nouvelle résolution de l'ONU*, les mots informatifs sont : *nouvelle, résolution, ONU*. Les mots *ONU* et *résolution* peuvent effectivement être des candidats termes et même former un candidat terme poly lexical *résolution de l'ONU*, mais le mot *nouvelle* ne s'inscrit pas dans cette terminologie. Dans d'autres exemples, les mots informatifs incluent également des mots de la langue générale, comme *cherche, utiliser, tonnes, L'or*, etc.

Nous avons remarqué que cette annotation n'est pas homogène [4] : un même mot peut être annoté comme informatif s'il est entouré par des mots plus courts ou comme non informatif s'il est à côté d'un mot plus long et encore moins fréquent. Pour cette raison, nous avons appliqué un coefficient pour corriger l'annotation. Le mot est validé comme informatif s'il a été annoté comme tel dans 90 % d'occurrences. Cette amélioration de l'algorithme de base a une conséquence positive sur les résultats de tête de la liste de fréquence absolue (table 2), mais en même temps elle neutralise le caractère local du

1. `http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html`, consulté le 3 mars 2014
2. `http://www.statmt.org/europarl/`, consulté le 3 mars 2014
3. Le choix de tokeniseur est susceptible de changer dans les futures versions du programme.
4. Il est difficile d'évaluer la performance de cette annotation, car la frontière entre les mots vides et informatifs n'est pas clairement définie. Les résultats de l'évaluation faite par Vergne (2003) ne sont pas comparables à l'usage que l'on fait de sa méthode. En effet, ils contiennent des candidats pour plusieurs langues mélangées et ne permettent pas de déduire les critères de validation.

calcul qui permettait de tenir compte des homographes (Vergne, 2004). Le corpus annoté par les mots informatifs et les mots vides servira de base pour l'extraction de candidats termes monolexicaux.

Baseline	M(I)>M(n) [5]	coef. 40 %	coef. 60 %	coef. 80 %	coef. 90 %
(	avoir	complexe	complexe	complexe	complexe
être	complexe	figure	figure	j.	énergie
en	figure	il	il	énergie	surface
ce	il	plus	deux	surface	système
dans	plus	deux	nous	système	métal
par	deux	nous	j.	différent	monsieur
au	nous	pouvoir	énergie	métal	état
que	pouvoir	j.	surface	monsieur	solution
avec	j.	on	système	état	atome
complexe	on	énergie	utiliser	solution	acide

TABLE 2 – Influence du coefficient sur le tri par fréquence absolue

3.3 Extraction de termes monolexicaux et apprentissage des ressources morphologiques endogènes

3.3.1 Termes monolexicaux

Les résultats obtenus par différentes mesures statistiques ne seront pas identiques, surtout ceux de tête de la liste triée. Pour cette raison, nous avons basé l'extraction de candidats termes monolexicaux sur le calcul de la fréquence absolue (nombre d'occurrences dans le corpus, formule (1)) et sur une formule qui retrouve des termes fréquents qui apparaissent dans un nombre réduit de documents (formule(2)). Les deux calculs tiennent compte des mots informatifs uniquement.

$$tf_i = \sum_{j=1}^{n} tf_{ij} : dj \in D \qquad (1)$$

où tf_{ij} est la fréquence d'un mot informatif dans d_j,
d_j est un élément de l'ensemble de documents D,
n est le nombre de documents

$$TF\mathrm{x}IDF(m_i) = \frac{tf(m_i)}{tf(m_{max})} \times \log \frac{D}{d} \qquad (2)$$

où m_{max} est le mot informatif le plus fréquent,
d est le nombre de documents contenant m_i

La deuxième formule appartient à la famille de TFxIDF, mais en représente une version adaptée pour être appliquée sur à une autre unité textuelle (le corpus). L'idée d'utiliser le coefficient TFxIDF pour l'extraction des termes n'est pas neuve : Witschel (2005) suggère d'identifier pour chaque document les termes fréquents qui apparaissent dans peu de documents du corpus. Nous avons modifié cette méthode en calculant la fréquence pondérée d'un mot lemmatisé par rapport à la fréquence maximale d'un mot informatif dans le corpus entier. Cette modification s'explique par la volonté d'éviter un bruit possible si un document du corpus ne porte pas vraiment sur le domaine donné, risquant ainsi de polluer les résultats de TFxIDF calculés pour chaque document.

La table 3 illustre la différence d'approche : la fréquence absolue maximale met en valeur les mots les plus fréquents dans les textes du corpus, tandis que TFxIDF identifie les mots fréquents, mais présents dans un nombre restreint de documents. Nous avons constaté que les mots qui sont en tête de la liste TFxIDF se trouvent au milieu de la liste triée par fréquence décroissante.

Paramètre	Top-10
MaxFreq	complexe, énergie, surface, système, métal, monsieur, état, solution, atome, acide
TFxIDF	perr, pile, capteur, accumulateur, peptide, inhibiteur, tension, 2phen, clo4, clhp

TABLE 3 – Top-10 résultats pour la fréquence absolue maximale et TFxIDF

Les deux listes partagent certains candidats termes, mais il est intéressant de traiter chaque liste à part : comme la précision

5. Le mot a été annoté comme informatif M(I) plus souvent que comme un non informatif M(n)

est maximale en tête de la liste triée par ordre décroissant, le premier tiers [6] de chaque liste sera retenu dans la liste des candidats termes monolexicaux qui seront utilisés pour générer des ressources morphologiques endogènes.

3.3.2 Ressources morphologiques endogènes

Notre méthode prévoit la génération de ressources morphologiques qui servent à l'apprentissage des patrons morphosyntaxiques. Ces ressources sont basées sur les n-grammes de caractères extraits à partir de candidats termes monolexicaux. En effet, les n-grammes correspondent à des quasi-morphèmes selon leurs positions, ce qui a notamment été exploité pour la tâche de tokenisation (McNamee *et al.*, 2009; McNamee, 2008).

Dans notre méthode, nous prenons en compte trois positions de n-grammes : au début du mot, au milieu (dans la fenêtre entre le deuxième et l'avant-dernier caractère du mot) et à la fin. Les listes obtenues confirment l'hypothèse que ces n-grammes se rapprochent des morphèmes de la langue. Nous avons testé le système sur les 2-grammes, 3-grammes et 4-grammes (table 4) à partir de mots dont la longueur est supérieure à 4 caractères. Les résultats de cette comparaison permettent de faire quelques observations sur le comportement de n-grammes selon leur position.

Position	Français						Anglais					
	Chimie			Télécommunication			Chimie			Télécommunication		
n	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**
Début	co	con	comp	co	con	inte	co	con	inte	co	con	inte
	in	pro	cons	in	com	cons	in	int	comp	re	int	comp
	pr	com	élec	re	pro	comp	re	com	elec	in	com	cont
	dé	tra	inte	pr	int	cont	di	pro	nano	pr	pro	cons
	di	int	phot	dé	tra	comm	pr	dis	spec	de	mul	tran
Milieu	at	ect	ectr	at	ent	icat	at	ect	ectr	en	ter	tion
	ti	rat	ctro	ti	rat	tion	er	ter	ctro	ti	ica	icat
	ct	ctr	tion	em	ica	isat	ti	ica	izat	at	ent	atio
	em	tro	isat	ra	ter	fica	en	tro	lect	en	ect	izat
	ro	ent	omét	te	ect	enta	ct	ctr	tion	ra	tio	erat
Fin	er	ion	tion	er	ion	tion	on	ion	tion	on	ion	tion
	on	ent	ment	on	ent	ment	ly	ate	ally	ng	ing	ally
	nt	que	ique	nt	ter	ique	er	ing	tive	ly	ent	tive
	re	eur	aire	re	eur	teur	al	ent	ical	er	ate	able
	ue	ire	teur	ur	que	aire	te	ity	onal	ed	ity	lity

TABLE 4 – Top-5 de n-grammes dans les corpus

En début de mots français, les bi-grammes et les tri-grammes contiennent des préfixes (*co-, in-, dé-, re-, ré-, pro-, com-, dis-, pré-*) ; ils sont largement partagés par les deux corpus, ce qui permet d'affirmer que ce sont plutôt les éléments de la langue générale. Les quadri-grammes, au contraire, sont plus caractéristiques de chaque corpus (5 sur 10 résultats sont différents) ; ces éléments contiennent notamment les formants classiques (*hydr-, phot-, micr-, mult-, télé-,* etc.)

Au milieu, les bi-grammes ne représentent pas d'intérêt, tandis que les tri-grammes et les quadri-grammes sont bien propres à chaque corpus. Dans ces listes, il est possible de deviner les racines fréquentes (*-électr-, communic-,* etc.)

A la fin, la quasi-totalité des n-grammes est partagée par les deux corpus. Nous y retrouvons les suffixes propres aux verbes (*-er, -ir*), aux substantifs (*-ion, -ité, -eur, - ment*) et aux adjectifs (*-que*). Certes, une partie de ces suffixes peut correspondre aux adverbes (*-ment*) ou aux participes présent (*-ant*), mais l'usage que l'on en fait minimise l'influence de ces éléments.

Le choix entre les tri-grammes et les quadri-grammes n'est pas simple : d'un côté il est possible que les quadri-grammes puissent donner plus de précision, mais le nombre de candidats termes sera assez réduit ; les tri-grammes peuvent générer plus d'imprécision, mais le nombre de candidats termes sera plus élevé. Nous optons pour les tri-grammes qui se placent dans le premier tiers de la liste triée par fréquence décroissante, combinés avec un score de confiance : chaque tri-gramme du début et du milieu apporte un point ; les candidats dont le score est supérieur à trois pourront être retenus. Pour illustrer

6. Nous avons pris les premiers 30 % des deux listes.

la méthode sur le tableau 4, le mot *électronique* aura un score égal à 3, car il contient les tri-grammes *ect, ctr, tro*. Ce score s'élèvera à 4 si le tri-gramme *éle* est ajouté à la liste des tri-grammes du début.

Ce filtre sera appliqué au reste des deux listes de mots pour en extraire des candidats termes monolexicaux qui ne sont pas en tête de la liste par le biais de fréquences.

Les tri-grammes de la fin des mots qui jouent le rôle de suffixes seront utilisés pour l'apprentissage de patrons morpho-syntaxiques endogènes.

3.4 Patrons morphosyntaxiques endogènes et extraction de termes polylexicaux

Nous avons observé deux types de patrons pour l'extraction de candidats termes polylexicaux : les patrons morphosyntaxiques et les structures contrôlées de Vergne (2005).

Les deux types de patrons représentent des inconvénients. Les patrons morphosyntaxiques sont dépendants de la langue et l'efficacité de la méthode dépend de la richesse de la liste fournie, tandis que les structures contrôlées basées sur les annotations mot informatif – mot vide manquent de cohérence : n'importe quel mot informatif peut se trouver en tête de l'expression.

Notre méthode permet de profiter des points forts des deux approches sans avoir à résoudre les problèmes cités ci-dessus. Tout d'abord, la structure de patrons endogènes n'est pas figée : ils sont appris dans la fenêtre de 5 mots. Les patrons validés satisfont les conditions de commencer et de terminer par des mots informatifs, et de ne pas contenir de chiffres ou de signes de ponctuation, à l'exception de l'apostrophe qui est à ce stade annoté comme un mot vide.

La deuxième particularité de ces patrons est d'utiliser les tri-grammes à la fin de mots informatifs pour remplacer les étiquettes morphosyntaxiques. En effet, l'expérience démontre que les patrons obtenus contiennent des syntagmes nominaux, verbaux et adjectivaux. Par exemple, les patrons les plus fréquents pour le corpus français en chimie : *ion n*[7] *n ion (632), ion n n n ion (406), ion n ion (268), ion que (261), ent n n ion (219), que n n ion (201), ide que (197), tre n n ion (157), etc.*, correspondent aux patrons morphosyntaxiques suivants :
– NOM (PREP ?(DET ?)) NOM
– NOM ADJ
– VER (PREP | DET) NOM
L'apprentissage de patrons endogènes ne nécessite pas d'étiquettes morphosyntaxiques : les tri-grammes remplissent cette fonction. Certes, les patrons endogènes n'excluent pas un certain niveau de bruit. Afin d'éviter ce bruit, nous pouvons restreindre les mots vides valides à la liste de mots de schéma[8] générée à partir du corpus à l'étape de l'acquisition des patrons (table 5). Pour construire cette liste, nous avons repéré les mots non-informatifs les plus fréquents qui apparaissent entre les mots informatifs à l'intérieur des patrons.

Corpus	Mots de schéma
Chimie	le, de, du, ', un, être, et, à, en, l
Télécommunication	le, de, du, un, et, à, être, en, pour, au

TABLE 5 – Mots de schéma endogènes

En appliquant les patrons endogènes combinés avec les mots de schéma, nous arrivons à extraire les syntagmes susceptibles d'être des candidats termes polylexicaux, tenant compte de la règle du premier tiers (Top-30 %) de la liste. Cependant, il serait intéressant d'inclure une validation définitive par les ressources morphologiques endogènes.

Les syntagmes retenus ainsi varient en fonction du corpus et de la langue. Vu qu'en français la morphologie flexionnelle est plus prononcée et que la différence de longueur entre les mots vides et informatifs est plus évidente qu'en anglais, la méthode fournit de meilleurs résultats pour le français. Afin d'augmenter la précision des patrons morphologiques endogènes, nous avons ajouté une vérification complémentaire : l'un des mots doit obligatoirement être un candidat terme monolexical.

Il faut noter également que ce ne sont pas toujours les patrons les plus fréquents qui donnent le plus de résultats pertinents, car certains patrons contiennent les mots non-informatifs extérieurs à la liste de mots de schéma retenue. La variation de

7. *n* est un mot non-informatif
8. Pour simplifier, nous avons choisi ce terme pour regrouper les mots de schéma et les mots fortement liés (Enguehard, 1993).

résultats pour un même patron peut être assez intéressante. Certains patrons sont très productifs, d'autres le sont moins. Plus un patron est long, moins il fournit de résultats.

4 Évaluation des résultats

La table 6 contient l'évaluation [9] des composantes de la méthode sur les 100 premières occurrences des listes retenues. Un candidat terme monolexical a été validé soit s'il appartient de manière non-ambiguë au domaine en question, soit s'il peut éventuellement participer à la formation d'un candidat terme polylexical. Pour les candidats termes polylexicaux, nous nous sommes limités au premier critère.

Nous avons utilisé la plate-forme TERMOSTAT [10] pour obtenir des résultats de référence pour l'extraction des termes monolexicaux. L'utilisation de cette ressource pour évaluer les candidats polylexicaux semble injuste, car TERMOSTAT se limite à des syntagmes nominaux et fournit, évidemment, une précision très élevée.

Corpus	Langue	Précision Mono				Poly
		Baseline	Fréq	TFxIDF	RME [11]	PME [12]
Chimie	FR	54 %	**81 %**	**75 %**	**87 %**	69 %
	EN	**72 %**	44 %	64 %	69 %	85 %
Télécommunication	FR	77 %	**82 %**	76 %	53 %	78 %
	EN	**88 %**	58 %	77 %	53 %	69 %

TABLE 6 – Précision dans les Top-100 candidats

4.1 Candidats termes monolexicaux

Pour le corpus français en chimie le système a identifié 8 292 candidats termes monolexicaux qui résultent de la combinaison de la fréquence absolue maximale, du TFxIDF et des ressources morphologiques endogènes (table 6).

Nous pouvons constater que la fréquence absolue des mots informatifs fournit des bons résultats pour le français (81,5 % de précision en moyenne), mais se montre beaucoup moins efficace en anglais (51 % en moyenne). Dans les résultats en anglais, nous avons des bons candidats, comme *temperature, concentration, synthesis, electrochemical*, mais aussi des mots outils comme *same, first, thus, etc*. Cela s'explique par la baisse de l'efficacité de la méthode de J. Vergne sur les langues où la différence de longueur entre les mots informatifs et vides n'est pas importante.

La mesure TFxIDF est relativement plus stable (75,5 % pour le français et 70,5 % pour l'anglais en moyenne). Il faut remarquer que TFxIDF met en évidence un grand nombre de formules chimiques comme *C6F5* qui sont difficiles à reconnaître si l'on n'est pas un spécialiste du domaine.

La combinaison des deux méthodes permet de retrouver une partie de la terminologie du domaine, mais il faut tenir compte du bruit présent dans les résultats. Certes, nous pouvons appliquer les ressources morphologiques endogènes en tant que filtre complémentaire, mais de cette manière une grande partie des résultats corrects sera perdue. Par exemple, nous pouvons perdre les candidats termes, comme *eau* ou *ADN* pour la chimie et *Mac, chunk, AMRT, etc.* pour la télécommunication. Cela aura un effet négatif sur les candidats termes polylexicaux qui contiennent ces candidats termes et qui ne passeront pas la validation par la condition d'avoir un terme monolexical confirmé. Cependant, cela reste à vérifier au cours de futurs essais.

Le filtre constitué des ressources morphologiques endogènes défini dans la section 3.3.2 a donné des résultats positifs pour les deux langues, mais s'est montré nettement plus performant dans le domaine de la chimie, ce qui doit s'expliquer par la nomenclature forte dans ce domaine. L'évaluation des résultats est faite sur la liste ordonnée suivant le nombre décroissant des tri-grammes retenus, ce qui explique la longueur des mots de tête de la liste (table 7).

9. L'évaluation a été faite manuellement par l'auteur de l'article. En cas de doute, le candidat était annoté comme invalide. À terme, il est prévu de nous adresser à des experts pour préciser l'évaluation.
10. `http://termostat.ling.umontreal.ca`
11. Ressources morphologiques endogènes
12. Patrons morphologiques endogènes

Chimie		Télécommunication	
FR	EN	FR	EN
variationnellement	bioelectrochemistry	proportionnellement	internationalization
interribonucleotide	methylphenylboronic	internationalisation	internationalisation
metallointercalation	electroconductivity	telecommunication	implementations
proportionnellement	electropolymerization	repositionnement	differentiability
environnementaliste	electropolymerization84	significativement	operationalizes
spectrophotometric	spectroelectrochemistry	transactionnelle	cooperativeness
photosensibilisation	photoelectrochemical	defphysicallayerconfiguration	generalization
cristallographiea	electropolymerized	perfectionnement	radiocommunications
surdimensionnement	triphenylmethylborate	interconnections	considerations
electrocatalytic	methoxyphenylboronicacid	recommendation	externalization

TABLE 7 – Top-10 résultats pour les ressources morphologiques endogènes

Il faut remarquer que cette méthode permet de retrouver de bons candidats même dans les résultats identifiés après la ligne 300 :

– autoregistration, intersection, multiprotocol, videoconference, multiplexer, etc. (télécommunication, EN) ;
– monocarbonyles, transcriptionnel, nanolithographie, acidification, isocarbonyle, submicroscopique, aminopropylmercaptotriazole, etc. (chimie, FR).

4.2 Candidats termes polylexicaux

Les patrons endogènes ont permis d'extraire 8 631 candidats termes pour le corpus français en chimie. Nous avons imposé la condition que le candidat terme polylexical doive contenir au moins un candidat terme monolexical. La taille de patrons varie entre 2 et 5 éléments. La performance de la méthode est illustrée dans la table 6.

Tout comme dans le cas de patrons morphosyntaxiques traditionnels, il est difficile de limiter le résultat aux bons candidats termes. La méthode présente un grand avantage : les patrons sont extraits à partir du texte sans aucune analyse préalable. Ainsi, nous arrivons à extraire des candidats termes assez complexes, comme :

– réduction abiotique du sulfate
– décroître de façon monotone
– nettoyage de l ' électrode
– système de conversion de énergie
– polarisation anodique et cathodique
– complexe tétranucléaire

Nous pensons que ces candidats sont complexes pour les systèmes classiques parce que l'étiqueteur morphosyntaxique [13] se perd dans une terminologie inconnue de son vocabulaire, ce qui met en question toute l'analyse par des patrons prédéfinis. Nous illustrons cela sur les deux occurrences du candidat terme *complexes tétranucléaires* dans le corpus en chimie qui sont étiquetées avec des erreurs différentes :

– des PRP :det du
– complexes ADJ complexe
– tétranucléaires NOM tétranucléaires

OU

– complexes NAM Complexes
– tétranucléaires ADJ tétranucléaires

En même temps, la méthode de patrons endogènes a une faiblesse. Comme le patron ne part pas d'étiquette précise, il est assez difficile de classer les résultats. Par exemple, les patrons contenant le tri-gramme -ent renvoient aussi bien aux substantifs qu'aux adverbes (*traitement du eau, méthode être fortement, etc.*).

La classification est un point d'autant plus difficile qu'il existe plusieurs critères de départ :

1. patron
2. fréquence
3. terme(s) monolexical(aux) présents

13. Dans cet exemple, il s'agit de l'étiquetage morphosyntaxique du corpus en chimie (FR) par TreeTagger (Schmid, 1994)

Corpus en chimie			Corpus en télécommunication		
Patron	Collocation	Fréq	Patron	Collocation	Fréq
[nce, que]	insuffisance technique	3	[ure, n, ert, n, née]	procédure de transfert de donnée	6
	séquence nucléotidique	3	[eur, ire]	valeur binaire	2
	résistance mécanique	3		leur propriétaire	5
	séquence peptidique	6		utilisateur stationnaire	2
	puissance massique	10		erreur binaire	11
	distance interatomique	46	[ire, n, mps, n, ion]	réduire le temps de création	3

TABLE 8 – Quelques exemples de résultats par patron sur les deux corpus en français

Classer les résultats par patron semble assez risqué, car nous avons déjà évoqué la variabilité de parties du discours pour le même n-gramme. En même temps, cela a un aspect intéressant qui consiste à voir la productivité du patron (table 8). La vue par patrons pourra servir à constituer un anti-dictionnaire de patrons pour les futures analyses. La fréquence est peut-être le critère le plus commode pour présenter le résultat (table 9).

Candidat	Fréq	Patron	Candidat	Fréq	Patron
acide nitrique	152	[ide, que]	spectrométrie de masse	89	[rie, n, sse]
mettre en évidence	127	[tre, n, nce]	atome de carbone	88	[ome, n, one]
métal de transition	124	[tal, n, ion]	énergie de adsorption	82	[gie, n, ion]
être également	124	[tre, ent]	résultat et discussion	78	[tat, n, ion]
complexe de ruthénium	123	[exe, n, ium]	corrosion du cuivre	76	[ion, n, vre]
acide phosphorique	123	[ide, que]	composant électrochimique	75	[ant, que]
température ambiant	120	[ure, ant]	génie électrique	72	[nie, que]
transfert de charge	118	[ert, n, rge]	milieu acide	71	[ieu, ide]
centre métallique	98	[tre, que]	prendre en compte	70	[dre, n, pte]
être possible	90	[tre, ble]	passage de drake	70	[age, n, ake]

TABLE 9 – Top-20 candidats polylexicaux pour le corpus en chimie (FR)

La fréquence permet de voir les associations polylexicales les plus utilisées dans le corpus. Ces associations peuvent être soit des candidats termes polylexicaux (*atome de carbone, acide phosphorique, etc.*), soit tout simplement des éléments de la langue générale largement utilisés dans les textes du même style (*résultat et discussion, mettre en évidence, être également*). De ce point de vue, il sera intéressant de vérifier la possibilité d'appliquer la méthode à l'identification du lexique transdisciplinaire.

Cependant, le tri par fréquence a ses inconvénients. La tête de la liste est majoritairement occupée par les patrons à deux ou à trois éléments. Les candidats termes à quatre ou à cinq éléments apparaissent moins fréquemment, ils ont donc peu de chances d'entrer dans la première partie de la liste. Par exemple, le candidat *optimisation du paramètre de maille* est à la fin de la liste avec la fréquence égale à deux.

La troisième possibilité est de prendre en compte les candidats termes monolexicaux. Cette option a un certain avantage : elle permet de trier les candidats termes du point de vue de l'ensemble. Le lemmatiseur de TreeTagger se trompe des fois sur les lemmes des mots de schéma (*de le* ou *du*), ce qui démultiplie les associations et corrompt les fréquences.

Cette classification par les mots en commun est également intéressante car elle donne une idée du comportement du candidat terme monolexical dans le corpus spécialisé (termes associés, variantes, exemples, etc.). Par exemple, le programme a identifié plusieurs candidats termes polylexicaux contenant le candidat terme monolexical *chunk* (table 10).

Nous pouvons observer que les candidats *chunk de donnée* et *chunks de donnée* sont séparés car le mot *chunk* est un emprunt et TreeTagger n'arrive pas à le lemmatiser. Cela joue sur la fréquence : si les deux candidats étaient réunis, leur fréquence serait égale à 57.

La paire *nouveau chunk de contrôle* et *chunk de contrôle* représente un autre cas : un candidat terme contient l'autre. Dans ce cas précis, il n'y a pas d'intérêt de retenir le plus grand, mais il existe d'autres cas où il est intéressant de garder les deux candidats, par exemple *partition du générateur photovoltaïque* et *générateur photovoltaïque*.

Collocation	Fréq	Patron
chunk de donnée	25	[unk, n, née]
format du chunk cookie-ack	2	[mat, n, unk, ack]
format du chunk heartbeat-ack	2	[mat, n, unk, ack]
format du chunk init-ack	2	[mat, n, unk, ack]
format du chunk asconf-ack	2	[mat, n, unk, ack]
nouveau chunk de contrôle	6	[eau, unk, n, ôle]
chunks de donnée	32	[nks, n, née]
chunk de contrôle	10	[unk, n, ôle]

TABLE 10 – Candidats polylexicaux contenant le candidat terme monolexical chunk

Il est alors intéressant de proposer les candidats termes polylexicaux lors de la validation d'un candidat terme monolexical, car cela fournit à l'utilisateur plus d'informations sur ce dernier. Les résultats obtenus à cette étape du travail permettent de planifier quelques améliorations de l'algorithme général du système, notamment l'élagage de certains types de termes polylexicaux et la présentation des résultats.

5 Conclusion et perspectives

La méthode proposée permet d'extraire les candidats termes mono et polylexicaux à partir d'un corpus de spécialité. Le principal avantage de la méthode est son applicabilité à une large palette de langues et de corpus. Elle ne nécessite aucune ressource morphologique ou morphosyntaxique extérieure. Ainsi, cette méthode peut être appliquée sur un corpus de tout domaine de spécialité et toutes les ressources sont alors générées par le programme.

La méthode peut être utilisée pour d'autres langues que le français et l'anglais à condition qu'un tokeniseur et un lemmatiseur soient disponibles pour la langue ciblée. Même si la tâche de lemmatisation nécessite un étiquetage morphosyntaxique, notre méthode ne l'utilise pas. De cette manière, la méthode exploite l'outil d'étiquetage morphosyntaxique différemment des autres méthodes de l'ET.

La méthode combine des méthodes statistiques (fréquence absolue et TFxIDF) et des approches mixtes (annotation par mots informatifs et vides, apprentissage des mots de schéma endogènes). La particularité de la méthode est d'utiliser des n-grammes de caractères pour remplacer des listes de formants classique et des étiquettes morphosyntaxiques.

Les résultats obtenus sur le prototype permettent d'envisager des améliorations de la méthode. Notamment, il est intéressant d'automatiser le choix des coefficients et de la taille de n-grammes en fonction de la langue, car cela pourrait donner des meilleurs résultats pour l'anglais ou pour d'autres langues.

La deuxième piste consistera à tester la possibilité de corréler les n-grammes avec les syllabes pour obtenir les ressources morhologiques et morphosyntaxiques endogènes. Cela pourrait également jouer sur l'efficacité de la méthode appliquée sur d'autres langues.

Le système sera testé sur des corpus multidomaines composés par genre afin de voir si la méthode pourrait servir à l'identification du lexique et des expressions transdisciplinaires.

Une validation des résultats par des experts de domaines est prévue dans le cadre de notre projet de recherche.

Références

ANTHONY, L. (2005). AntConc : design and development of a freeware corpus analysis toolkit for the technical writing classroom. *In International Professional Communication Conference Proceedings*, pages 729–737.

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In Advances in Natural Language Processing*, pages 380–387. Springer.

BERNHARD, D. (2006). *Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales*. Thèse de doctorat, Université Joseph Fourier – Grenoble I.

BOURIGAULT, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN*, pages 75–84, Nancy.

BOURIGAULT, D. et FABRE, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25:131–151.

BOURIGAULT, D., GONZALEZ-MULLIER, I. et GROS, C. (1996). LEXTER, a Natural Language Processing Tool for Terminology Extraction. *7th EURALEX International Congress*.

BOURIGAULT, D. et JACQUEMIN, C. (1999). Term extraction+ term clustering : An integrated platform for computer-aided terminology. *In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 15–22.

DAILLE, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In The Balancing Act : Combining Symbolic and Statistical Approaches to Language, Workshop at the 32nd Annual Meeting of the ACL (ACL'94)*, Las Cruces, New Mexico, USA.

DAILLE, B. (2003). Conceptual structuring through term variations. *In* BOND, F., KORHONEN, A., MACCARTHY, D. et VILLACICENCIO, A., éditeurs : *ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16.

DROUIN, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, XII:45–64.

DROUIN, P. et LANGLAIS, P. (2006). Évaluation du potentiel terminologique de candidats termes. *In Actes de JADT*, volume 2006, page 8.

ENGUEHARD, C. (1993). Acquisition de terminologie à partir de gros corpus. *Informatique & Langue Naturelle*, pages 373–384.

ESTOPÀ, R., VIVALDI, J. et CABRÉ, T. (2000). Extraction of monolexical terminological units : requirement analysis. *In Workshop Proceedings Second International Conference on Language Resources and Evaluation. Terminology Resources and Computation*, volume 56, pages 51–56, Athens, Greece.

FRANTZI, K., ANANIADOU, S. et MIMA, H. (2000). Automatic Recognition of Multi-Word Terms : the. *International Journal on Digital Libraries*, 3(2):115–130.

JACQUEMIN, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Thèse de doctorat, Institut de Recherche en Informatique de Nantes.

JACQUEY, E., TUTIN, A., KISTER, L., JACQUES, M.-p., HATIER, S. et OLLINGER, S. (2013). Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines. *In Terminologie et Intelligence Artificielle (TIA)*, Paris.

MCNAMEE, P. (2008). N-gram Tokenization for Indian Language Text Retrieval. *In Working Notes of the Forum for Information Retrieval Evaluation*.

MCNAMEE, P., NICHOLAS, C. et MAYFIELD, J. (2009). Addressing morphological variation in alphabetic languages. *In 32nd Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 75–82.

MORIN, E. et DAILLE, B. (2006). Comparabilité de corpus et fouille terminologique multilingue. *TAL*, 47:113–136.

MORIN, E. et DAILLE, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. *In Conférence conjointe JEP-TALN-RECITAL 2012*, pages 141–154.

OROBINSKA, O., LYON, E. et CHAUCHAT, J.-h. (2013). Enrichissement d'une ontologie de domaine par extension des relations taxonomiques à partir d'un corpus spécialisé. *In Terminologie et Intelligence Artificielle (TIA)*, volume 704, Paris.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester, {UK}.

TUTIN, A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, XII:5–14.

VERGNE, J. (2003). Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN*, 2:139–148.

VERGNE, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. *In Actes de JADT*, Louvain.

VERGNE, J. (2005). Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. *In Actes de CIDE*, pages 155–168.

WITSCHEL, H. F. (2005). Terminology Extraction and Automatic Indexing. *In Terminology and Knowledge Engineering (TKE)*, pages 1–12.

Une description des structures de la durée en Langue des Signes Française à partir d'une grammaire formelle

Mohamed Hadjadj
LIMSI, Rue John Von Neumann, 91400 Orsay, France

mohamed.hadjadj@limsi.fr

Résumé. Dans cet article, nous abordons la problématique du fonctionnement de la temporalité en langue des signes française (LSF). Nous allons étudier plus particulièrement quelques structures portant sur la durée. Nous présenterons dans un premier temps les descriptions existantes du système aspecto-temporel de la LSF et les difficultés que nous trouvons pour modéliser ces travaux. Le but de cet article est de proposer une grammaire formelle qui prenne en compte le fonctionnement de la LSF et qui puisse faire l'objet d'un traitement de modélisation. Notre démarche consiste à étudier un corpus LSF pour établir des liens de fonction à forme afin d'obtenir des règles de grammaire qu'on peut générer dans un projet de synthèse à l'aide d'un signeur avatar.

Abstract. Temporality constitutes a major issue in filed of modeling french signed language (LSF). In fact, it is very difficult to model actual discriptions of the aspect-temporal systems of LSF. In this paper we present the bases of a novel formal grammar that permits the modeling of the LSF. This paper presents a study to construct this grammar. We analysed a French SL corpus to create formel rool between the signed gesture and its signification. Our objective is to obtain rules of grammar that can generate a synthesis project using a signer avatar.

Mots-clés : grammaire, LSF, temporalité, modélisation LSF.

Keywords: grammar, LSF, temporality, modeling LSF.

1 Introduction

L'interdiction des langues des signes (LS) durant des années a eu, sans aucun doute, un impact crucial sur la nature et le nombre d'études linguistiques portant sur ces langues. C'est à (Stokoe, 1960) que revient le mérite d'attester le statut linguistique de l'ASL et donc des autres langues des signes. L'objectif de ses études était de rapprocher le fonctionnement des langues des signes de celui des langues vocales en décrivant un fonctionnement spécifique induit par la mise en œuvre de quelques paramètres manuels. Il les résume dans l'emplacement où le signe est réalisé, la forme de la main pendant la réalisation du signe et le mouvement qu'elle décrit. Par la suite, de nombreux chercheurs vont développer son modèle en ayant toujours comme références des modèles phonologiques pour les langues vocales (Battison, 1974), (Klima, Bellugi, 1976).

Cependant ces études proposent des modèles qui ne se rendent pas nécessairement compte des spécificités qu'on observe dans les langues des signes. La description des unités qui font l'objet de la construction phonologique, traits ou phonèmes, reste ambigüe. A cela, on ajoute l'ignorance de la dimension iconique, propriété de ressemblance entre le signifiant et le signifié, qu'on constate dans la forme de certains signes en LS. Les chercheurs de cette école considèrent que l'aspect iconique de ces signes intervient à un autre niveau que le niveau phonologique.

Ces limites ont donné naissance à une nouvelle approche pour décrire les langues des signes. Elle est représentée essentiellement en France, au début des années 90, par les travaux de Cristian Cuxac. Dans son modèle, Cuxac (1996, 2000) met l'iconicité au centre de la LSF. Ses travaux et ceux de ses successeurs ont su mettre en question plusieurs éléments dans l'analyse linguistique, dite classique, des langues des signes.

Cette démarche permet de mieux apercevoir le fonctionnement de la langue des signes française, mais ces travaux ne présentent pas une grammaire qui peut faire l'objet d'un travail de modélisation. Dans cet article, nous proposons une grammaire formelle qui prend en compte le fonctionnement de la langue des signes française. Nous l'appliquerons, dans cette étude, sur quelques structures de la durée en LSF.

2 La temporalité en LSF

De nombreuses études en linguistique ont pu montrer que les modèles qui s'appuient plutôt sur l'aspect grammatical du temps ne sont pas applicables sur toutes les langues, certaines langues comme le chinois (Whorf, 1968), ou st'àimcets (Demirdache H & Uribe- Etxebarria 2002) ne grammaticalisent pas le temps. La LSF, une langue très particulière, une langue visio-gestuelle, par rapport aux langues vocales, pourrait avoir son propre système temporel qui ne peut pas être décrit selon l'aspect grammatical. Le linguiste François A avance *« La description linguistique sur des bases empiriques, de langues encore inexplorées permet de mettre à jour des regroupements sémantiques inédits, des catégorisations et des stratégies d'encodage qui n'avaient pas nécessairement été observées jusqu'alors ».* (François, 2001).

Les travaux de Cuxac (Cuxac, 1996, 2000), basés sur le principe de l'iconicité[1], trouvent que le système temporel en LSF s'appuie essentiellement sur deux niveaux d'iconicités, l'iconicité diagrammatique, une sorte de schéma actantiel dans l'espace et l'iconicité du mouvement, fondée sur le principe des relations aspectuelles (Sallandre, 2003). A partir des notions: corps, espace, temps, le locuteur a un schéma représentatif dans l'espace de signation, une fois le schéma inséré, il peut exprimer toute relation temporelle à l'aide des trois axes de l'espace.

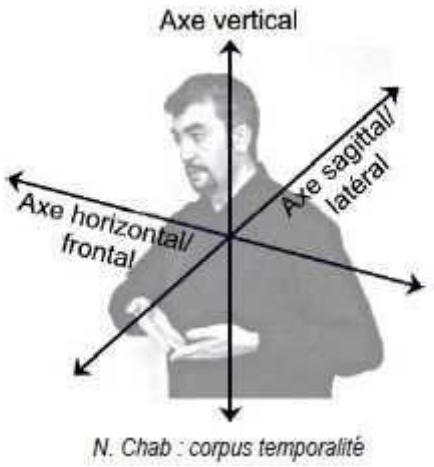


Figure1 : les 3 axes temporels

3 La notion d'aspectualité en LSF

Dans de nombreuses études, l'aspectualité en LSF est une simple juxtaposition de certains paramètres à la forme de base. L'aspectualité est résumée dans le type du mouvement, tel que le mouvement cyclique de la main pour exprimer l'aspect duratif, ou la mimique faciale pour décrire, à titre d'exemple, « le continu » (Cuxac 2000). D'autre part (Fridman, 1975) et (Deuchar, 1985) trouvent que certains signes lexicaux renvoient à une valeur aspectuo-temporelle.

Il faut également noter que les études portant sur la temporalité et l'aspectualité en LSF se basaient souvent sur des exemples étudiés séparément de leur contexte. Ainsi, dans de nombreuses recherches, une description fine des différents articulateurs non manuels ayant un rôle dans l'élaboration du sens global d'une structure n'est pas une analyse indispensable. Cette démarche ne peut décrire des phénomènes linguistiques complexes de la LSF.

[1] Dans son modèle théorique, Cuxac distingue deux visées sémiologiques: le dire en donnant à voir (les structures de Grande Iconicité) et dire sans montrer (les signes lexicaux).

4 Développement d'une grammaire formelle

4.1 La démarche suivie

Dans le cadre de notre étude, nous allons développer une grammaire formelle pour décrire le fonctionnement de quelques structures portant sur la durée en LSF. Dans notre démarche, nous nous appuyons sur une hypothèse moins restrictive que des modèles existants. Notre méthodologie consiste à établir des liens de fonction à forme. Il s'agit d'effectuer des allers-retours entre la forme, le mouvement d'un articulateur ou plusieurs, et la fonction, l'interprétation de cette forme. L'objectif est de distinguer un groupe invariant qui unit les différentes occurrences d'une forme ou d'une fonction. Pour cela, nous considérons que tous les articulateurs, manuels et non manuels, comme potentiellement pertinents sans attribution de fonction a priori. En effet, cela nous permet de préciser la forme comme étant une combinaison de différents articulateurs manuels et non manuels en précisant les différentes synchronisations nécessaires qui peuvent avoir lieu entre ces articulateurs. Cette méthodologie peut être un bon moyen pour décrire l'équilibre complexe qu'on observe dans la LSF. Ainsi, cela nous permet d'avoir une génération de la langue qui s'approche de la langue naturelle.

4.2 Le corpus

Le corpus sur lequel nous nous sommes appuyés, dans le cadre de notre étude, est conçu, en grande partie, pour étudier certaines relations temporelles. Il s'agit d'une sélection de 40 brèves journalistiques à partir de 980 brèves de l'année 2006 du site internet websourd[2], traduites ensuite par les traducteurs du même site internet pour créer un corpus parallèle (Filhol, 2013). Aussi, nous avons analysé un autre corpus, de la même nature que le premier, pour étudier d'autres types de structures du duratif. L'ensemble des deux corpus est constitué de 120 brèves journalistiques.

4.3 Présentation du schéma d'annotation

Nous présentons ci-dessous les différents articulateurs annotés avec leurs attributs. Nous avons fait le choix de ne pas étudier tous les articulateurs dans un premier temps, cependant nous avons ajouté d'autres articulateurs pour bien décrire certaines fonctions étudiées.

Les paupières (Paup)

Nous distinguons dans la piste paupières trois attributs :

- Ouvert (o): le moment d'ouverture des yeux.

- Plissé (pl): le moment où les yeux sont mi-fermés.

- Fermé (f): le moment où les paupières sont complètement serrées. Nous avons aussi annoté les clignements (cl) des yeux.

Les sourcils (Srcl)

Dans la piste sourcils, nous avons pris en compte que trois attributs :

- Standard (st) : cet attribut correspond au niveau des sourcils par rapport à la partie inférieure des paupières dans un état de pause.

- Vers le haut (v.h) : dans les cas où la distance entre les sourcils et la partie inférieure des paupières est supérieure à l'attribut standard.

[2] www.websourds.org

- Vers le bas (v.b) : dans les cas où la partie inférieure des paupières est plus proche aux sourcils par rapport à l'attribut standard.

Mouvement du buste (Mvb)

Il s'agit des mouvements où la tête reste droite, uniquement le buste doit bouger. Nous distinguons deux attributs : vers la droite (v.d) et vers la gauche (v.g).

Mouvement de la tête(Mvt)

Nous avons annoté le mouvement de la tête sur trois axes de l'espace ou ce qui correspond en LSF aux signes lexicaux: «oui, non et peut-être »

La direction du regard (DR)

Nous nous sommes contentés, dans cette piste, de différencier deux attributs majeurs : le regard vers l'interlocuteur, dans le cas de notre corpus la camera, et l'espace de signation, désignant dans la littérature l'espace devant le locuteur que celui-ci utilise pour structurer son discours, les autres regards, plus rares, ne sont pas annotés.

La description manuelle

La description manuelle est assez détaillée, 2000 signes sont déjà décrits dans le modèle Zebedee (Filhol, 2010). Cependant nous annotons le début et la fin du geste manuel pour décrire la synchronisation entre les différents articulateurs. Ainsi, nous avons annotés séparément les deux mains, dominante (m.do) et dominée (m.dé).

5 Résultats

Après avoir sélectionné les différents exemples du corpus portant sur l'expression de la durée en LSF, deux structures se distinguent, des brèves qui représentent un événement qui dure dans le temps et des brèves avec deux évènements, séparés par une durée donnée.

5.1 Un évènement qui dure dans le temps

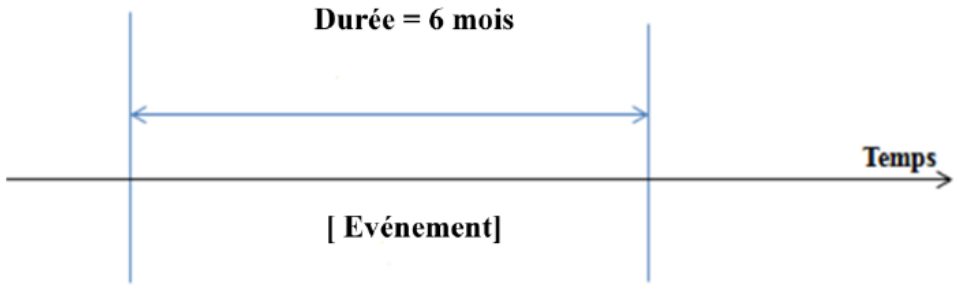


Figure 2 : structure d'un évènement qui dure dans le temps

Une fois que nous avons annoté les différentes structures d'un évènement qui dure dans le temps, deux formes totalement différentes se dégagent.

(A) (B)

Figure 3 : l'expression de la durée

En mettant le lien entre ces deux figures et leurs fonctions, nous avons constaté que le choix de l'une ou de l'autre pour exprimer la durée s'effectue selon sa longueur. Si la période est inférieure à dix jours, les locuteurs utilisent la figure (A), dans une période supérieure à dix jours, ils signent la figure (B). Nous présentons ci-dessous les groupes invariants de chaque structure. La flèche sur les schémas représente le temps, les articulateurs avec la précision de leurs attributs sont mis en gras, les signes lexicaux en italique et enfin les évènements qui précèdent ou succèdent la durée sont mis en majuscule. Les abréviations utilisées font référence au schéma d'annotation.

Durée d'un évènement d'une longueur inférieure à 10 jours.

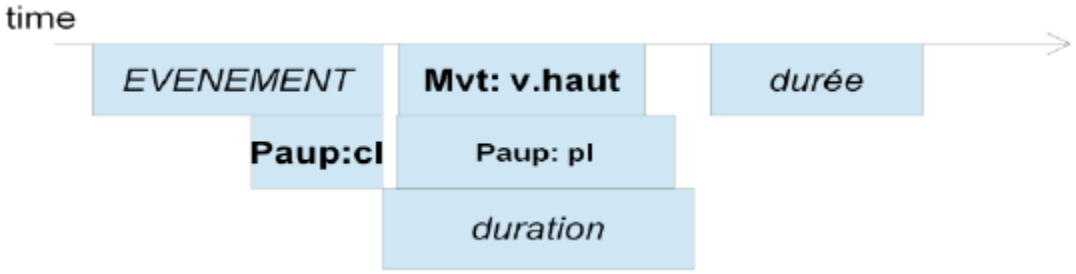


Figure 4 : le groupe invariant d'un évènement d'une longueur inférieure à 10 jours

Durée d'un évènement d'une longueur supérieure à 10 jours.

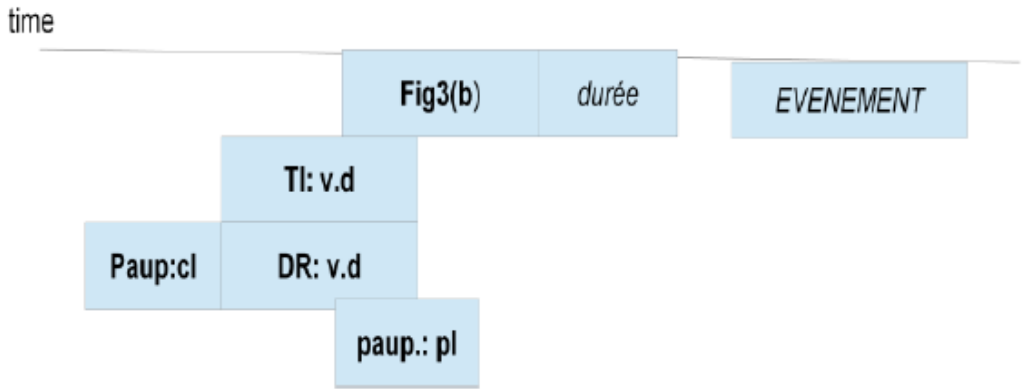


Figure5 : le groupe invariant d'un évènement d'une longueur supérieure à 10 jours.

Dans les deux groupes invariants nous remarquons un plissement des yeux pour marquer la durée. Ce point vient confirmer les résultats de (CHETELAT-PELE, 2010). Ainsi, il s'agit de combinaisons complexes entre plusieurs articulateurs et non pas juste une simple juxtaposition.

Un événement entre deux bornes temporelles

Dans les exemples d'un événement qui dure pendant une période précise, limitée par deux bornes temporelles, les locuteurs utilisent une forme bien différente de celle qu'on trouve dans des manuels de la LSF. En plus de l'ignorance des gestes non manuels et de la synchronisation entre les différents articulateurs, au niveau du geste manuel, ni l'emplacement de la main ni sa direction ne correspond à la description qu'on trouve dans des manuels, le geste est effectué plutôt sur l'axe horizontal que sur l'axe sagittal.

Figure 6 : L'expression de la durée entre deux bornes temporelles

Nous décrivons le groupe invariant de cette structure ci-dessous :

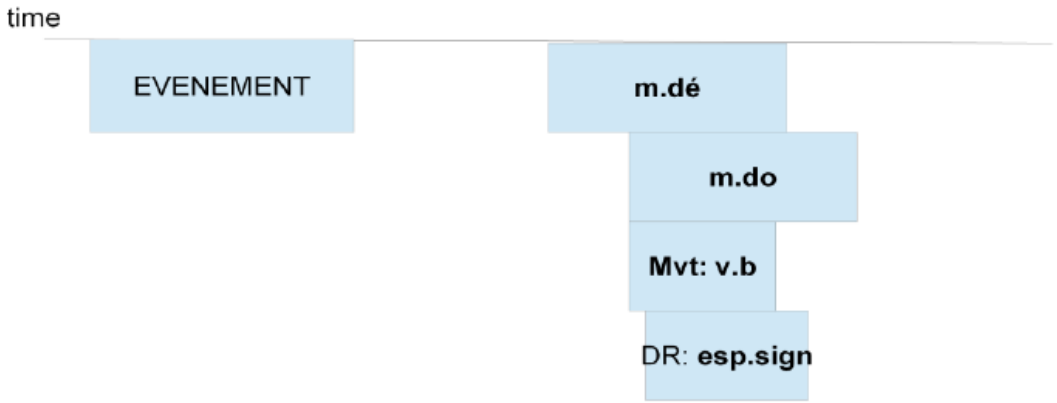


Figure 7 : le groupe invariant de L'expression de la durée entre deux bornes temporelles

Durée d'un évènement partant du présent à une borne temporelle

Ainsi, dans les structures d'un évènement entre deux bornes temporelles, nous avons trouvé une autre forme qui se distingue, les procès qui partent du présent à une borne temporelle sont signés différemment. Les résultats trouvés viennent confirmer les travaux déjà effectués dans la littérature, notamment les travaux de (Cuxac 2000). Le corps du locuteur, sur l'axe sagittal, illustre le présent, une main sera maintenue au niveau du corps, le regard suivra la deuxième main qui va vers l'avant pour désigner la deuxième borne temporelle.

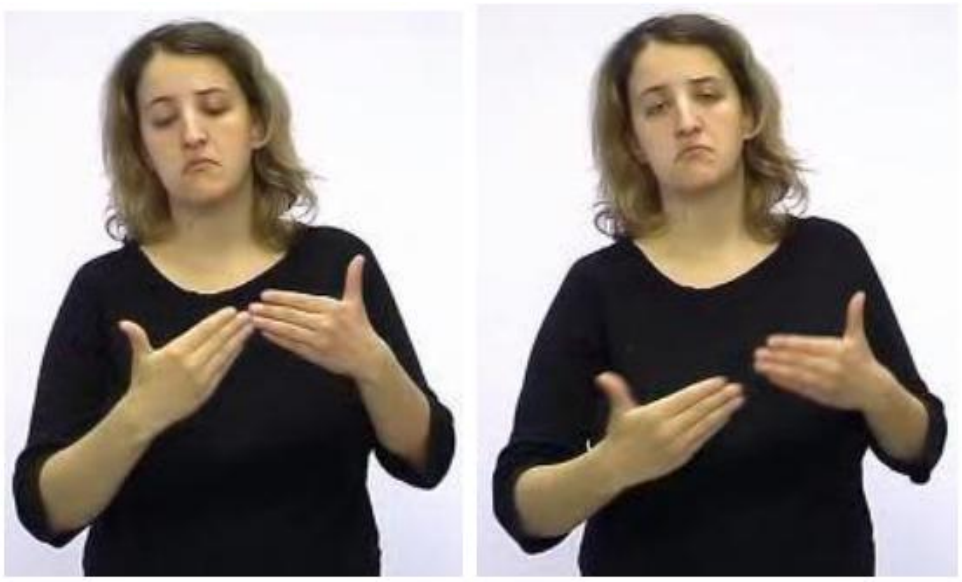


Figure 8 : l'expression de la durée entre deux bornes temporelles, partant du présent.

Nous décrivons le groupe invariant de cette structure ci-dessous :

Figure 9 : le groupe invariant de l'expression de la durée entre deux bornes temporelle, partant du présent.

5.2 Deux évènements liés par une relation de précédence, éventuellement séparés par une durée.

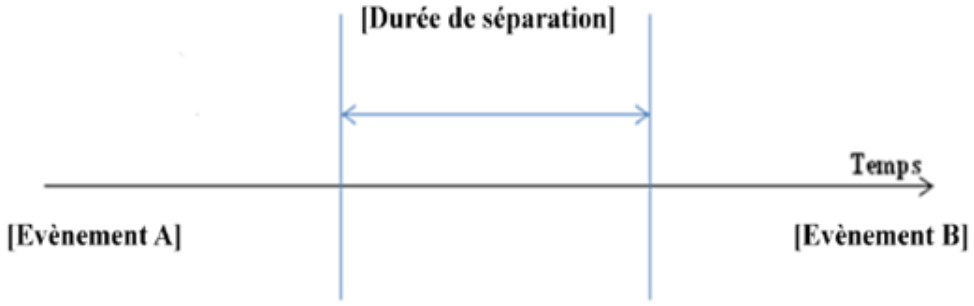


Figure 10 : structure de deux évènements séparés par une durée

En analysant les structures de deux évènements avec une durée de séparation, nous avons trouvé que la longueur de la durée est aussi importante dans le choix de la forme. Pour signer deux évènements avec une durée de séparation inférieure à 10 jours, les locuteurs commencent par l'évènement qui se déroule le premier chronologiquement, la durée de séparation est signée entre les deux évènements, elle est exprimée par un ensemble de synchronisation entre les différents articulateurs décrit dans la figure ci-dessous.

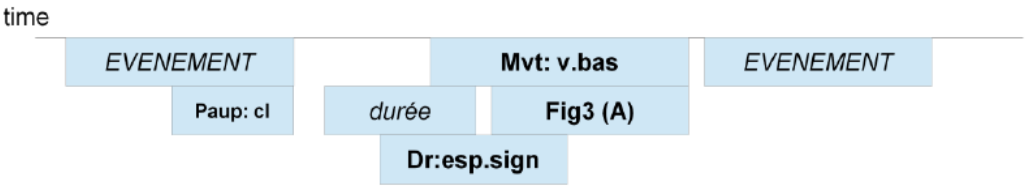


Figure 11 : groupe invariant de deux évènements avec une durée de séparation inférieure à 10 jours

En ce qui concerne les relations de précédence entre deux évènements avec une durée de séparation supérieure à 10 jours, les locuteurs signent les évènements selon leur ordre chronologique. La durée est exprimée de la même manière que dans les structures d'un évènement qui dure plus que 10 jours**.**

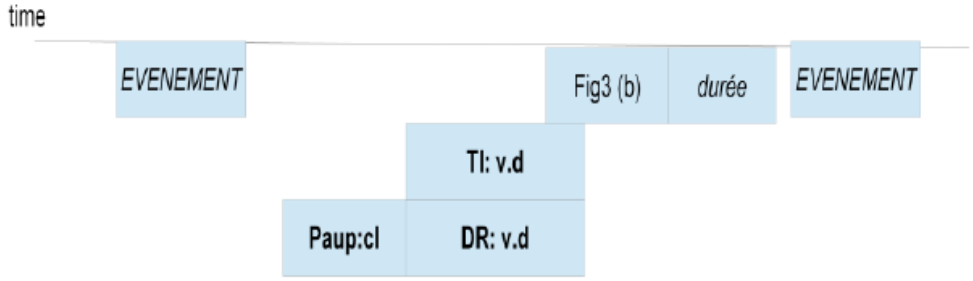


Figure 12 : groupe invariant de deux évènements avec une durée de séparation supérieure à 10 jours

6 L'imbrication des règles

Le but de notre étude est de développer une grammaire basée sur un ensemble fini de règles de dérivation qui permet d'engendrer de façon systématique toutes les phrases en LSF. Décrire la grammaire de LSF de cette manière nous permet d'imbriquer les différentes règles dans un projet de génération. Si nous prenons à titre d'exemple la brève suivante : *« Dix ans après l'évacuation musclée de l'église Saint-Bernard, le 23 août 1996 à Paris, les sans papiers et leurs soutiens ne veulent pas être « dans la commémoration » mais dans le « combat », comme l'illustre le mouvement autour des expulsés du squat de Cachan. »*

Nous constatons que la brève se constitue de deux évènements séparés par un période de dix ans, une structure déjà définie dans la section 5.2 (une succession d'évènements avec une durée de séparation supérieure à 10 jours.) Nous constatons aussi que le premier évènement est un évènement daté, dans une étude antérieure (Filhol, 2013), nous avons constaté qu'en LSF, on commence par la date avant de signer l'évènement. La figure ci-dessous représente l'architecture de la description formelle de la brève étudiée.

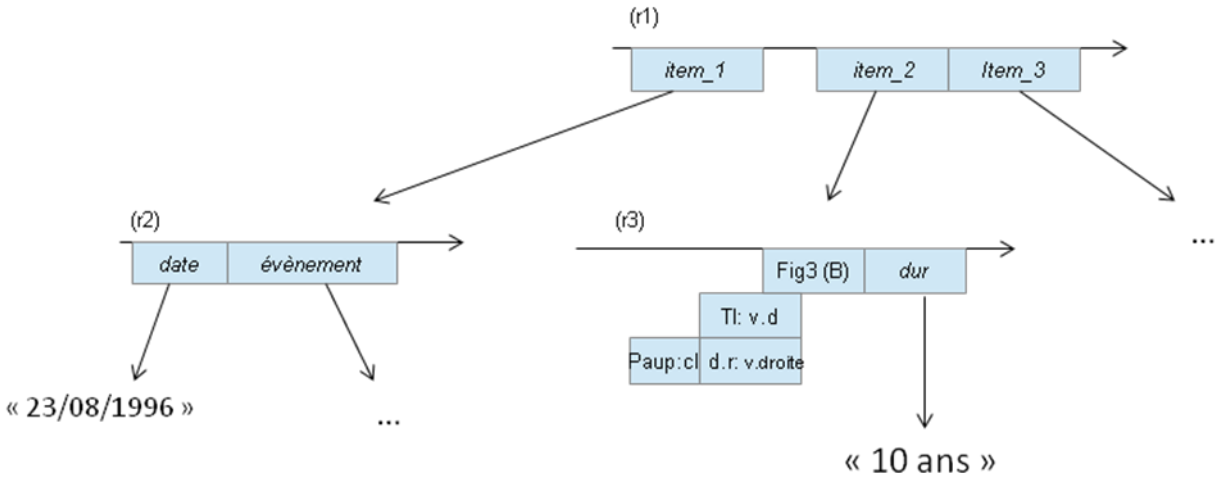


Figure 13 : l'imbrication des règles

Dans l'arbre ci-dessus, nous partons d'une règle générale (R1) pour définir une succession d'évènements avec une durée de séparation. Ainsi, à l'intérieur de (R1), on peut imbriquer d'autres règles pour un évènement daté (R2) et l'expression d'une durée de séparation (R3).

(R1)

item_1 = (R2)

Date = "23/08/1996"

Event = "évacuation musclée de l'église St-Bernard"

item_2 = (R3) (dur = "10 ans")

item_3 = [...]

7 Conclusion

La prise en compte de tous les articulateurs, manuels et non manuels, dans notre analyse forme-fonction, nous a permis de bien définir les formes de certaines fonctions portant sur le duratif. Une telle démarche nous permet de développer une grammaire formelle qui prend en compte les particularités de la LSF. Une description formelle d'une langue encore peu inexplorée doit être fondée sur des bases empiriques, une application des grammaires conçues pour les langues écrites sur les LS ne peut décrire leur fonctionnement bien particulier.

8 Références

BATTISON R. (1974). Phonological deletion in American Sign Language. SLS, 5, 1–19.

BELLUGI U., KLIMA E. (1976). Two faces of sign: Iconic and abstract. In S. Harnad, D. Hoest & I. Lancaster (eds.), Origins and evolution of language and speech. New York, New York Academy of Sciences. pp. 514-538.

CHETELAT-PELE E. (2010). Les gestes non manuels en langue des signes française. Annotation, analyse et formalisation : Application aux mouvements des sourcils et clignements des yeux. Thèse de doctorat, LIMSI.

COHEN D. (1989). L'Aspect verbal, Presses Universitaires de France. Paris.

CUXAC C. (1996). Fonctions et structures de l'iconicité des langues des signes. Analyse descriptive d'un idiolecte parisien de la langue des signes française. Université René Descartes - Paris V. Thèse de Doctorat d'Etat.

CUXAC C. (2000). La Langue des Signes Française (LSF) : Les voies de l'iconicité. Fait de Langue, 15-16 Ophrys.

DEMIRDACHE H., URIBE- ETXEBARRIA M. (2002). La grammaire des prédicats spatiotemporels : temps, aspect et adverbe de temps. Dans Temps et Aspect : de la morphologie à l'interpretation Laca, Brenda : Presses Universitaires de Vincenes

DEUCHAR M. (1985). The implications of sign language research for linguistic theory. Proceedings of the Third International Symposium on Sign Language Research, Silver Spring, Maryland: Linstok Press and Rome: Istituto di Psicología, 239-246.

FILHOL M. (2010). Search through lexical sign bases with a constraint-based model, Theoretical issues on Sign Language research (TISLR 10), Purdue, USA.

FILHOL M., HADJADJ M., TESTU B. (2013). *A rule triggering system for automatic text-to-Sign translation*, International workshop on Sign Language translation and avatar technology *(SLTAT)*, Chicago, USA.

FRANCOIS A. (2001). Gabarit de procès et opérations aspectuelles en Matlov (Océanie) in Actances 11, juin 2001, Rivaldi *(GDR 749 du CNRS).*

FRIEDMANN L. A. (1975). Space, time & person reference in ASL. *In Language 51, 940-961.*

FUSELLIER-SOUZA I., LEIX J. (2005). L'expression de la temporalité en Langue des Signes Française (LSF). Actes du colloque Conceptualisation et Surdité, *dans La nouvelle revue AIS.* Editions du CNEFEI, Suresnes, pp. 207-230

SALLANDRE M.-A. (2003). Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité. *Thèse de Doctorat,* Univ. Paris 8.

STOKOE W.C. (1960). Sign Language Structure. Studies in Linguistics. Occasional Papers n° 8. Buffalo, NY : *University of Buffalo Press.*

TOURNADRE N. (2004). Typologie des aspects verbaux et intégration à une théorie du TAM, *BSL*. Peeters.

WORF B. (1968). Linguistique et anthropologie, Denoël-Gonthier, Paris.

Interaction homme-machine en domaine large à l'aide du langage naturel : une amorce par mise en correspondance

Vincent Letard[1,2]
(1) LIMSI, CNRS, rue John von Neumann, 91405 Orsay cedex
(2) Université Paris-Sud, 91400 Orsay
letard@limsi.fr

Résumé. Cet article présente le problème de l'association entre énoncés en langage naturel exprimant des instructions opérationnelles et leurs expressions équivalentes et langage formel. Nous l'appliquons au cas du français et du langage R. Développer un assistant opérationnel apprenant, qui constitue notre objectif à long terme, requiert des moyens pour l'entraîner et l'évaluer, c'est-à-dire un système initial capable d'interagir avec l'utilisateur. Après avoir introduit la ligne directrice de ce travail, nous proposons un modèle pour représenter le problème et discutons de l'adéquation des méthodes par mise en correspondance, ou *mapping*, à notre tâche. Pour finir, nous montrons que, malgré des scores modestes, une approche simple semble suffisante pour amorcer un tel système interactif apprenant.

Abstract. We consider the problem of mapping natural language written utterances expressing operational instructions to formal language expressions, applied to French and the R programming language. Designing a learning operational assistant, which is our long term goal, requires the means to train and evaluate it, that is, a baseline system able to interact with the user. After presenting the guidelines of our work, we propose a model to represent the problem and discuss the fit of direct mapping methods to our task. Finally, we show that, while not resulting in excellent scores, a simple approach seems to be sufficient to bootstrap an interactive learning system.

Mots-clés : assistants interactifs, apprentissage artificiel, systèmes de question-réponse.

Keywords: interactive assistants, machine learning, question answering systems.

1 Introduction

Les avancées techniques et théoriques permettent d'effectuer des opérations de plus en plus puissantes et efficaces avec l'aide des ordinateurs. Pour autant, travailler avec la machine ne devient pas nécessairement plus simple. Exploiter la richesse de l'interaction homme-machine (Allen *et al.*, 2007; Volkova *et al.*, 2013) devrait permettre d'améliorer l'efficacité d'une tâche effectuée par l'humain à l'aide d'un ordinateur.

Notre objectif à long terme est de concevoir un assistant opérationnel dialogique apprenant par l'interaction à fournir une commande correcte en langage formel (LF) à partir d'un énoncé en langage naturel (LN). L'intérêt d'un tel système se trouve à la fois dans l'utilisation de l'outil informatique par un novice en programmation qui peut néanmoins exprimer son objectif en LN, et auprès d'utilisateurs plus avancés d'un langage de programmation, mais occasionnels ou bien pas toujours au plus haut niveau de complexité. Ainsi, l'utilisateur avancé pourrait enseigner au système les commandes complexes dont il se sert de loin en loin et les réutiliser facilement, et l'utilisateur novice pourrait profiter des connaissances dont dispose déjà le système (enseignées par d'autres utilisateurs au moyen de *crowdsourcing*[1] par exemple). Nous nous sommes intéressés pour ce travail au langage R. Il s'agit d'un langage de programmation pour l'analyse statistique et le traitement de données. Il est riche en fonctionnalités et souvent utilisé sous forme de scripts, qui permettent d'en oublier la syntaxe.

Avant tout, la conception d'un tel système apprenant requiert la collecte de données, et les premières tentatives ont souligné l'importance de l'utilisabilité pour le processus d'apprentissage. En effet, un système apprenant par l'interaction avec ses

[1]Le crowdsourcing est une méthode de collecte des données nécessaires à un processus avec l'aide d'un grand nombre de personnes. Ce processus peut centraliser ces données pour une utilisation locale, ou bien les réutiliser au profit de l'ensemble des personnes participantes.

	Énoncés en LN	Commandes (en R)
1	Charge les données depuis **"res.csv"**	`var1 <- read.csv("res.csv")`
2	Trace un histogramme de la colonne **2** de **tab**	`plot(hist(tab[[2]]))`
3	Dessine la répartition de la colonne **3** de **tab**	`plot(hist(tab[[3]]))`
4	Somme les colonnes **3** et **4** de **tab**	`var2 <- tab[[3]] + tab[[4]]`
5	Somme les colonnes **3** et **4** de **tab**	`var3 <- sum(c(tab[[3]], tab[[4]]))`

TAB. 1: Un échantillon d'associations entre énoncés en LN et commandes en LF. Ces exemples précisent la commande attendue pour chaque énoncé. Les éléments en gras sont liés avec les paramètres des commandes, cf. section 4.1. Les variables temporaires présentes dans la colonne de droite sont introduites par le système afin de permettre à l'utilisateur de faire référence à des résultats intermédiaires pour des traitement ultérieurs. Par ailleurs, on peut noter qu'un énoncé peut correspondre à plusieurs commandes différentes et inversement.

utilisateurs se doit de conserver leur intérêt sous peine d'obsolescence due à la perte de sa source d'information. Il nous faut donc fournir au système des capacités et connaissances initiales afin de permettre son développement incrémental avec l'aide des utilisateurs. Sans ces derniers, collecter des données se révèlerait bien plus fastidieux. Nous estimons que le prérequis minimal pour rendre le système utilisable est qu'il apporte de l'aide à l'utilisateur dans au moins 50% des cas. Ce premier seuil est volontairement bas car la visée d'un premier système est avant tout d'amorcer la collecte de données ; de meilleures performances devraient être atteintes par la suite grâce à ses nouvelles connaissances. Nous formulons en outre l'hypothèse que des méthodes simples de mise en correspondance entre LN et LF peuvent atteindre ce score.

Notre approche est basée sur une association directe paramétrée entre les énoncés en LN et les commandes en R. On utilise une base de connaissances K composée d'associations paramétrées, telles que dans le tableau 1, pour sélectionner la meilleure commande à associer à l'énoncé-requête. Les énoncés $e^* \in K$, les plus proches de l'énoncé-requête selon une mesure de similarité σ, sont choisis. Les commandes associées $C(e^*)$ dans K sont adaptées aux paramètres de l'énoncé-requête et une commande est retournée. Par exemple, étant donné l'énoncé-requête e_{req} : "Charge le fichier data.csv", le système range les énoncés de K par similarités décroissantes avec e^*. Pour le contenu de la base exemple dans le tableau 1, l'énoncé 1 doit intuitivement être classé premier, et le système doit retourner la commande : "`var1 <- read.csv("data.csv")`". Notons que plusieurs commandes peuvent être proposées en une fois afin d'offrir un choix d'alternatives à l'utilisateur.

Nous utilisons les mesures de similarité basées sur la distance de Jaccard, le tf-idf, et le score BLEU, et considérons différentes stratégies pour le choix des réponses retournées par le système. Les mesures de similarité évaluées se sont révélées être suffisemment complémentaires pour permettre l'utilisation de méthodes de combinaison, telles que le vote ou la classification automatique, pour améliorer *a posteriori* l'efficacité de la recherche.

La section 2 présente les domaines autour desquels s'articule notre problématique, ainsi que quelques travaux proches, nous posons ensuite notre formulation du problème et discutons de ses particularités en section 3. La section 4 détaille la méthode de constitution des associations et les différentes mesures de similarités utilisées, tandis que les paramètres d'analyse du LN et de l'ensemble de données utilisé sont exposés section 5. Enfin, nous donnons en section 6 les résultats d'expériences et discutons de leurs causes et implications.

2 État de l'art

2.1 Associer du langage naturel à du langage formel

Des problèmes proches ont été précédemment abordés à l'aide de différentes méthodes d'apprentissage. La transformation de requêtes en LN vers SQL est étudié par (Popescu *et al.*, 2003), le système requiert une grande précision dans ses réponses car l'objectif est de le rendre accessible au grand public et donc utilisable sur toutes sortes de bases de données. (Branavan *et al.*, 2009, 2010) utilise l'apprentissage par renforcement pour associer des instructions en anglais à des séquences de commandes en LF. Cela permet à l'association de prendre en compte les instructions haut-niveau et leurs constituantes. L'étendue des commandes élémentaires utilisables est cependant limitée aux possibilités de l'interaction graphique. Il en résulte que l'apprentissage ne peut pas produire de schémas très abstraits, du fait de la faible diversité des paramètres dans les commandes graphiques. Dans leur approche, (Kushman & Barzilay, 2013) abordent le problème de la génération d'expressions régulières correspondant à des descriptions en anglais à l'aide des grammaires combinatoires

catégorielles pour analyser le langage naturel et la représentation avec le λ-calcul pour inférer des règles de traduction du LN vers les expressions régulières. Cette approche générative par traduction permet la généralisation à partir des exemples d'apprentissage. Cependant, le pouvoir expressif des expressions régulières correspond aux grammaires de type 3 de la hierarchie de Chomsky. (Yu & Siskind, 2013) utilisent les modèles de Markov cachés établir par apprentissage une mise en correspondance entre des détections d'objets dans une séquence vidéo et des prédicats extraits de descriptions en LN. L'objectif de leur approche est différent du notre, mais le problème sous-jacent de trouver une association entre objets peut être comparé. Les objets appariés sont dans notre cas des expressions en LF plutôt que des détections dans une séquence vidéo.

2.2 Traduction automatique

La traduction automatique (Hutchins & Somers, 1992) renvoie habituellement à la transformation d'une phrase depuis un LN source en une autre portant le même sens dans un autre LN, appelé langage cible. Cette tâche est effectuée par la construction d'une représentation intermédiaire de la structure de la phrase a un niveau d'abstraction donné, puis la phase de génération encode l'objet obtenu dans le langage cible. Bien que suivant un objectif principal différent, l'une des tâches du projet XLike (Tadić *et al.*, 2012) était l'examen des possibilités de traduction d'instructions en LN (anglais) vers un LF (Cycl). Adapter une approche de ce type à un un langage formel cible opérationel (par opposition à Cycl qui est déclaratif) peut être une piste intéressante à étudier, mais il nous faut tout d'abord satisfaire l'objectif primaire de l'utilisabilité.

2.3 Recherche d'information

La question des systèmes de recherche d'information est comparable à celle de l'assistant opérationnel lors du parcours de sa base de connaissances. Les systèmes de question-réponse en particulier, révèlent des similarités avec l'assistant opérationel car les deux ont pour tâche de répondre à une expression en LN en recherchant la meilleure réponse dans l'ensemble des connaissances à leur disposition. Cependant, les systèmes de question-réponse s'appuient généralement sur la fouille de textes afin de trouver l'information correcte (Toney *et al.*, 2008). Cette méthode demande une grande quantité de données annotées (à la main ou par annotation automatique). Les tutoriels, cours ou manuels qui pourraient être utilisés à cette fin pour l'assistant opérationnel sont malheureusement trop hétérogènes et incluent des références complexes ou implicites à des connaissances générales (langage, algorithme, compilation). En un mot ils sont écrits pour l'humain, et il n'est pas envisageable de les utiliser en fouille de données sans une étude approfondie, quel que soit le mode d'annotation. D'où l'intérêt d'un assistant opérationnel apprenant capable de collecter des données standardisées et annotées à l'aide de l'utilisateur.

3 Formulation du problème

Ainsi que décrit en introduction, la base de connaissances K est représentée par un ensemble d'exemples de la relation binaire $R : LN \rightarrow LF$ qui associe un énoncé en LN à une commande en LF. Si nous considérons le cas simple dans lequel la relation est fonctionnelle et injective, chaque énoncé est associé à une unique commande. Ce n'est pas réaliste car beaucoup d'énoncés en LN portent le même sens. Le cas d'une relation non injective couvre mieux les exemples usuels : chaque commande peut être associée à un énoncé ou plus, les exemples 2 et 3 du tableau 1 illustre cette situation. Cependant, le cas le plus réaliste est celui d'une relation qui n'est ni injective ni fonctionnelle. Plusieurs énoncés peuvent être associés à la même commande, et un seul énoncé ambigu peut renvoyer à plusieurs commandes différentes (voir les exemples 4 et 5 du tableau 1). Nous devons considérer toutes ces associations lors de la comparaison de l'énoncé-requête e^* avec les énoncés de K pour sélectionner une ou plusieurs commandes à retourner.

Pour cela, plusieurs stratégies non exclusives peuvent être considérées pour déterminer ce que le système retourne à l'aide de la mesure de similarité $\sigma : LN \times LN \rightarrow \mathbb{R}$ entre deux énoncés en LN. Typiquement, il faut déterminer si le système doit répondre, et si oui, combien de commandes il doit retourner.

La première stratégie est axée sur le nombre de réponses données pour chaque énoncé-requête e_{req}. Les n premières commandes relativement au classement de leurs énoncés associés dans K sont retournées. Étant donné e_{req}, le rang r d'un énoncé $e \in K$ est donné par le nombre d'énoncés de K dont la similarité avec e_{req} est supérieure à celle de e avec e_{req}.

$$r(e|e_{req}) = |\{e' \in K : \sigma(e_{req}, e') > \sigma(e_{req}, e)\}|$$

Le second choix de stratégie est de déterminer un seuil de similarité en dessous duquel les énoncés candidats de K et leurs commandes associées sont considérés trop différents pour correspondre. Cela permet de choisir si une réponse est donnée ou non, et donc d'autoriser le silence du système, mais n'offre pas de contrôle sur le nombre de commandes retournées (bien qu'on puisse le conserver par la suite sous un seuil raisonnable). Cette stratégie retourne donc comme résultat l'ensemble des commandes dont l'énoncé associé dans K a une valeur de similarité avec e_{req} supérieur au seuil déterminé :

$$Res = \{c \in LF : (e, c) \in K, \sigma(e_{req}, e) > s\}$$

avec s le seuil de similarité sélectionné. Le tableau 2 illustre le classement des énoncés de K par similarité décroissante avec e_{req}. Notons que, bien que les énoncés de la base de connaissances aient été ordonnés, rien ne nous permet de discriminer les commandes qui leur sont associées. En conséquence, l'application de la première stratégie avec un nombre de commandes égal à 4 sélectionne les deux commandes associées à e1, et un sous-ensemble quelconque de taille 2 parmi les commandes associées à e4. Afin de rationaliser cette sélection, on peut envisager de pondérer chaque association de

e	$\sigma(e, e_{req})$	commandes associées à e dans K
e1	$0,80$	{c2, c6}
e4	$0,74$	{c4, c5, c7}
e3	$0,36$	{c3, c6}
e2	$0,36$	{c1}

TAB. 2: Exemple de classement pour un énoncé e_{req} donné.

K en fonction du nombre d'utilisations correctes et incorrectes (par exemple l'association e1 $\rightarrow$ c6 reçoit un poids de 4 car la commande c6 a été retournée pour e1 5 fois avec succès et 1 fois à tort). Ainsi, les commandes les "moins risquées" seront toujours privilégiées. Cependant, il ne s'agit que d'une optimisation de surface et elle demande des informations sur l'utilisation du système, ce dont nous ne disposons pas. D'autre part, on peut également remarquer qu'une commande peut apparaître à plusieurs endroits dans le classement des énoncés de K. c6 apparaît donc pour e1 et pour e3, avec deux valeurs de similarité différentes. Dans ce cas, seule la mieux classée sera retenue.

La formulation du problème proposée repose la fonction de similarité σ et sur les associations présentes dans la base de connaissances. Nous allons maintenant présenter notre approche pour établir ces associations et les fonctions de similarité que nous y appliquons.

4 Approche

Nous avons initialement le résultat d'une analyse syntaxique simple de l'énoncé et de la commande. La première étape à effectuer est l'acquisition des exemples et la procédure de mise à jour de la base de connaissances. Nous examinons ensuite les méthodes pour rechercher une commande à partir de la base connaissances et d'un énoncé-requête donné.

4.1 Intégration dans les connaissances

L'association correcte entre énoncés et commandes requiert au moins la prise en compte de leurs paramètres respectifs (noms de variables, valeurs numériques et chaînes de caractères entre guillemets). Les représentations génériques des énoncés et des commandes sont construites à l'aide de l'identification des paramètres dans le couple à intégrer à la base de connaissances (voir tableau 1). Ces représentations sont utilisées par la suite pour reconstruire la commande à l'aide des paramètres de l'énoncé-requête.

La base de connaissances ne contient que les formes génériques des commandes. Il s'agit du texte de la commande comportant des références non résolues pour chaque paramètre qui a été associé à un élément de l'énoncé d'apprentissage. Ces références sont résolues à la phase de recherche par association avec les tokens correspondants de l'énoncé-requête.

4.2 Retrouver les commandes

Nous avons appliqué trois mesures de similarité différentes pour la recherche de commandes afin de comparer leurs points forts et leurs points faibles : l'indice de Jaccard, une aggrégation du tf-idf[2], ainsi que le score BLEU[3]. Le choix de ces trois méthodes est dicté par la diversité des caractéristiques qu'elles mesurent, qui provient elle-même des cadres dans lesquels elles ont chacune été développées.

4.2.1 Indice de Jaccard

L'indice de Jaccard mesure la similarité entre deux ensembles à valeurs dans le même domaine. Dans notre cas, nous comparons l'ensemble des mots de l'énoncé-requête en LN et celui de l'énoncé d'apprentissage pour la commande candidate. Ils sont chacun valués dans l'ensemble des tokens possibles. L'expression de l'indice de Jaccard adaptée pour deux énoncés e_1 et e_2 est :

$$J(e_1, e_2) = \frac{|M(e_1) \cap M(e_2)|}{|M(e_1) \cup M(e_2)|}$$

où $M(e)$ désigne l'ensemble des mots de l'énoncé e utilisés. Il semble plus pertinent dans ce cas d'ignorer les mots outils. En effet, la comparaison des énoncés sous la forme d'ensembles de mots fait perdre toute information sur leur ordre dans la phrase. Sans le contexte, les mots outils n'apportent que peu d'information et introduisent plutôt un biais dans le ratio des nombres de tokens. L'indice de Jaccard est une méthode standard pour évaluer les co-occurrences des unigrammes. Elle devrait être plus efficace avec des données comprenant peu d'exemples ambigus en termes de vocabulaire.

4.2.2 tf-idf

La mesure tf-idf calcule une mesure de représentativité d'un mot dans un document, par rapport à un corpus. Elle permet donc, étant donné un mot, de classer les documents du corpus selon sa représentativité pour chacun d'eux. L'avantage de tf-idf est qu'il tient compte de la fréquence du mot dans le corpus. Ainsi, un mot apparaissant dans tous les documents, même s'il est très fréquent, ne sera représentatif d'aucun document. Il faut ici évaluer la représentativité d'une phrase plutôt que d'un seul mot. On utilise donc une aggrégation des valeurs de tf-idf pour chacun des mots composant l'énoncé-requête en LN.

$$tfidf_e(e_{req}, e_{comp}) = \frac{1}{|M\,(e_{req})|} \sum_{m \in M(e_{req})} tfidf(m, e_{comp}, E)$$

avec $E = \{enonce | (enonce, commande) \in K\}$ l'ensemble des énoncés de la base de connaissance, et où e_{req} est l'énoncé-requête et e_{comp}, l'énoncé comparé. La fonction $tfidf$ elle-même s'exprime par :

$$tfidf(m, e_{comp}, E) = \frac{|\{m_e | m_e \in e_{comp} \wedge m_e = m\}|}{|\{m_e | m_e \in e_{comp}\}|} \times log\left(\frac{|E|}{|\{e \in E | m \in e\}|}\right)$$

Cette mesure ne nécessite plus la restriction aux mots pleins car tf-idf inclut déja la normalisation par rapport à la fréquence globale (dans le corpus) des termes.

4.2.3 Le score BLEU

Le score BLEU (Papineni *et al.*, 2002) est une méthode de calcul de similarité qui a été développée pour automatiser l'évaluation de la traduction automatique. Il présente une bonne corrélation avec l'évaluation par l'humain et est donc pertinent pour la recherche de paraphrases. Cett méthode s'appuie sur la mesure de co-occurrences entre n-grammes et permet de distinguer les énoncés candidats dans lesquels l'ordre des mots est trop différent de la celui dans la référence (énoncé-requête). La valeur de précision modifiée qu'elle introduit est basée sur le rapport des co-occurrences de n-grammes entre candidat et référence, sur la taille totale du candidat, normalisée selon n.

$$P_{BLEU} = \sum_{gram_n \in e_{req}} \frac{max_{e_{comp} \in E} occ(gram_n, e_{comp})}{|\{gram_n \in e_{req}\}|}$$

[2]Term frequency-inverse document frequency
[3]Bilingual evaluation understudy

où $occ(gram_n, e) = \sum_{gram'_n \in e}[gram_n = gram'_n]$ est le nombre d'occurrences du n-gramme $gram_n$ dans l'énoncé e. On peut noter que le dénominateur, qui correspond au nombre de n-grammes de l'énoncé e_{req}, peut aussi s'écrire $|e_{req}|-(n-1)$. La méthode de calcul du score BLEU comprend également une pénalité pour la brièveté, et ainsi empêcher les longs énoncés d'être trop désavantagés. Cependant, les énoncés comparés dans notre cas ne sont pas des documents de longueur très variables mais des expressions opérationnelles de tailles et d'amplitudes beaucoup plus raisonnables. Nous avons donc laissé de côté cette normalisation.

4.3 Filtrage des éléments syntaxiques

Avant d'évaluer leur similarité, nous avons appliqué plusieurs combinaisons de filtres sur les tokens des énoncés à prendre en compte. Il est possible de conserver ou non les mots outils ou les éléments non lexicaux. Ces derniers regroupent les valeurs numériques, les chaînes de caractères entre guillemets et les noms de variables. Les noms de variables sont les mots inconnus de l'analyseur syntaxique qui correspondent à une sous-partie identique dans la commande. Si les éléments non lexicaux sont conservés, ils sont transformés en substituts standardisés. De plus, il est possible d'appliquer ou non la lemmatisation des termes lexicaux. Par exemple, en ignorant les mots outils, en conservant les éléments non lexicaux et après application de la lemmatisation, le deuxième énoncé du tableau 1 deviendrait :

*Trace un histogramme de la colonne **2** de **tab*** → TRACER HISTOGRAMME COLONNE **XXVALXX XXVARXX**

5 Paramètres expérimentaux

5.1 Analyse des énoncés

Avant toute analyse lexicale, on recherche dans les énoncés en LN des expressions arithmétiques afin de les marquer en tant que telles. Les énoncés sont ensuite analysés à l'aide de WMATCH, un analyseur syntaxique générique à base de règles, développé par Olivier Galibert (Galibert, 2009). Il s'agit d'un système modulaire, disposant d'ensembles de règles pour le français et l'anglais. À titre d'exemple, voici à quoi ressemble le résultat d'analyse pour le premier énoncé du tableau 1 :

```
<_operation>
  <_action> charge|_~V </_action>
  <_det> les </_det>
  <_subs> données|_~N </_subs>
  <_prep> depuis </_prep>
  <_mot_inconnu>
    "res.csv"
  </_mot_inconnu>
</_operation>
```

Les mots marqués comme inconnus sont considérés comme des noms de variables potentiels, à rechercher dans la commande associée. Les chaînes de caractères comme `"res.txt"` sont également marquées afin de rechercher ultérieurement des correspondances possibles avec la commande. Nous posons une hypothèse pour simplifier l'analyse : les énoncés en LN sont supposés ne contenir aucune faute d'ortographe.

D'autre part, les commandes sont normalisées par la séparation de toute paire de caractères non liés sémantiquement. Les valeurs numériques ainsi que les noms de variables/fonctions sont identifiés et marqués comme paramètres de la commande.

Ayant défini la méthode d'analyse, nous allons à présent décrire les données utilisées pour notre expérimentation.

5.2 Constitution du corpus

Le corpus d'amorçage est composé de 605 associations entre 553 énoncés uniques en français et 240 commandes uniques en R. Le faible nombre de documents (tutoriels, manuels, documentations) différents qui décrivent une grande partie des commandes R ainsi que leur hétérogénéité rendent malheureusement irréalisable l'acquisition automatique d'exemples d'apprentissage sans une étude approfondie de leurs structures. En effet, ces documents sont écrits pour des lecteurs humains, disposant de références générales sur les tâches ; de la valeur de retour des fonctions usuelles aux principes généraux de la programmation, en passant par les problématiques usuelles d'utilisation des bibliothèques d'analyse statistique. C'est pourquoi chaque paire énoncé-commande de notre corpus a été ajoutée à la main, s'assurant que chaque élément reflète complètement l'information utile de l'exemple, et qu'ils respectent les spécifications. Celles-ci sont censées être le moins restrictives possible : un énoncé en LN doit simplement être formulé comme s'il s'agissait de demander l'exécution de la tâche R associée. Cela conduit donc à une majorité de phrases à l'impératif, qui reflètent, pour les personnes expérimentées, la manière dont elles exprimeraient les tâches concernées pour des non spécialistes.

5.3 Métriques d'évaluation

Les mesures permettant une évaluation pertinente du système dépendent de son objectif. Les valeurs de précision et de rappel pour les systèmes de question-réponse sont calculées comme suit :

$$P = \frac{\text{\# réponses correctes}}{\text{\# réponses données}} \qquad R = \frac{\text{\# réponses correctes}}{\text{\# réponses correctes dans } K}$$

Ces formules peuvent être appliquées pour un système retournant un seule réponse pour chaque énoncé. En effet dans le cas contraire, la précision mesurerait la proportion de bonnes réponses totale. Or, on s'intéresse plus précisément au nombre d'énoncés test pour lesquels une commande correcte a été retournée. Nous définissons la précision par énoncé P_e et le rappel par énoncé R_e.

$$P_e = \frac{\text{\# énoncés corrects}}{\text{\# énoncés données}} \qquad R_e = \frac{\text{\# énoncés corrects}}{\text{\# énoncés corrects dans K}}$$

Le nombre d'énoncés corrects s'entend ici le nombre d'énoncés test pour lesquels le système a retourné au moins une commande correcte. On peut noter que la mesure de rappel n'est pas aussi pertinente ici qu'en recherche d'information : si l'on pose l'hypothèse que la situation montrée par les quatrième et cinquième associations du tableau 1 ne sont pas courantes[4], le nombre d'énoncés corrects pour un énoncé requête donné devrait être faible, et la plupart devraient être équivalents. Ainsi, l'objectif principal n'est pas de retourner le plus grand nombre d'énoncés corrects car un suffit. Nous écartons donc l'évaluation du système à l'aide du rappel.

Examinons à présent les scores obtenus pour la mesure de précision par énoncé selon les différents paramètres du système.

6 Résultats et analyse

6.1 Comparaison des mesures de similarité

Comme le montre le tableau 3 la mesure de similarité basée sur le tf-idf domine les deux autres mesures en comparaison individuelle, quelle que soit la combinaison de paramètres de filtrage. En effet, la forme des énoncés présents dans le corpus cause la répétition des mots du vocabulaire de l'interaction opérationnelle (verbes : "donner", "calculer", "afficher" ; noms : "vecteur", "matrice", "histogramme", ...). Cette structure peut expliquer que la fréquence inverse dans les documents (idf) soit particulièrement pertinente.

[4] L'augmentation du nombre de tâches différentes couvertes par le corpus rendra ces collisions plus fréquentes, mais cette hypothèse semble raisonnable pour une amorce.

mots outils	inclus				non inclus			
non lexicaux	inclus		non inclus		inclus		non inclus	
lemmatisation	oui	non	oui	non	oui	non	oui	non
Jaccard	38.5%	34.6%	21.2%	23.1%	36.5%	36.5%	21.2%	23.0%
tf-idf	**50%**	**50%**	**36.5%**	**40.4%**	**48.0%**	**51.9%**	**36.5%**	**40.4%**
BLEU	34.6%	34.6%	26.9%	30.8%	30.8%	32.7%	26.9%	30.8%
hasard	1.9%							

TAB. 3: Mesures de précision par énoncé (P_e), en fournissant 3 réponses pour chaque énoncé requête test. L'ensemble des associations connues contient 85% du corpus, et l'ensemble de test 10%.

La lemmatisation et l'inclusion des mots outils ne semblent pas avoir une influence évidente sur la précision par énoncé. À l'inverse, on constate une amélioration dans tout les cas avec l'inclusion des éléments non lexicaux. Ce comportement provient au moins en partie de la longueur des énoncés de K ($7,5$ mots en moyenne), dont certains sont assez courts pour ne contenir aucun token lexical significatif (*"In de A"*). L'analyse linguistique plus approfondie devrait permettre, notamment pour les mots outils, d'améliorer les performances par une exploitation plus fine de la structure des énoncés. Cependant, notre travail a pour objectif de préparer une amorce pour un système évolutif, et l'intérêt d'utiliser des règles d'analyse fixes est donc limité.

La figure 1a montre la précision obtenue avec tf-idf en fonction du nombre de commandes retournées pour chaque énoncé requête. D'après le graphique, il est intéressant de proposer au moins trois commandes à l'utilisateur afin d'obtenir une précision décente. En revanche, proposer à l'utilisateur un choix de plus de 5 éléments ne serait pas pertinent : le gain en précision par énoncé ne serait plus suffisant par rapport au temps supplémentaire nécessaire à l'utilisateur pour retrouver la bonne réponse (le cas échéant) parmi les propositions.

6.2 Autoriser le silence

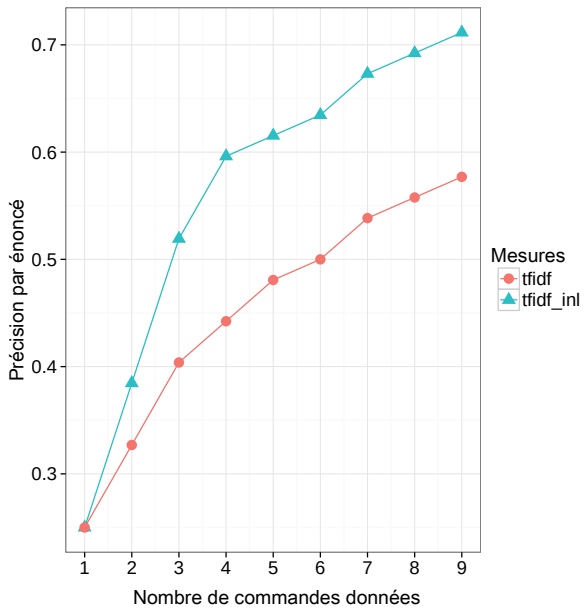
Afin d'autoriser le système à ne pas toujours donner une réponse, nous avons fixé un seuil absolu pour la valeur de similarité des commandes à retourner. Les résultats montrent, quelle que soit la mesure de similarité choisie, qu'au moins les 6 seuils les plus sélectifs donnent toujours lieu à des réponses fausses. Ce résultat peut être dû à l'existence dans l'ensemble de test, d'exemples non couverts par l'ensemble d'apprentissage. Plus vraisemblablement, il doit s'agir de commandes trop courtes (au plus un ou deux tokens lexicaux) pour être traitées correctement par la fonction de similarité.

6.3 Combinaisons

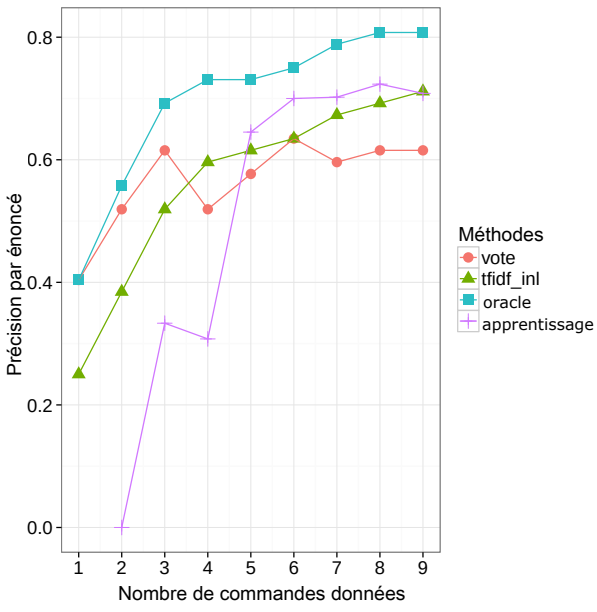
Une fois que chacune des méthodes proposées a été testée indépendamment, il peut être intéressant de tenter de les combiner. Cela permet notamment d'étudier leur complémentarité. L'oracle des 6 meilleures méthodes[5] montre une marge de progression intéressante (*cf.* figure 1b). Les résultats de la combinaison par le vote dépassent la meilleure méthode seule pour un nombre d'alternatives retournées inférieur à 4. Le vote atteint notamment 50% de bonnes réponses pour seulement 2 alternatives. La position de la courbe est moins claire pour un nombre d'alternatives plus élevé, mais notre problématique concerne l'utilisabilité du système, c'est pourquoi on s'intéresse plutôt aux gains pour un faible nombre de propositions. Cependant, il est indispensable d'effectuer les tests sur d'autres tirages de K dans le corpus afin d'obtenir une mesure de la variabilité de ces résultats.

On peut également exploiter la complémentarité des méthodes par l'entraînement d'un modèle d'apprentissage artificiel pour combiner *a posteriori* leurs résultats. La régression logistique n'est pas directement applicable à notre cas puisque les listes de commandes retournées avec les différentes mesures de similarité sont discrètes ; c'est-à-dire qu'on ne peut pas les fusionner directement à l'aide d'un paramètre réel. Nous avons donc utilisé la classification afin d'identifier les schémas pour lesquels l'une des méthodes est meilleure que les autres. Il s'agit d'entraîner un système à reconnaître les méthodes auxquelles se fier en fonction des indices dont nous disposons. Les valeurs de similarité discrétisées ont été employées en tant qu'attributs et l'ensemble des mesures ayant donné la bonne réponse en tant qu'étiquettes de référence. Ces paramètres ont été testés à l'aide des machines à vecteur de support de `libsvm` (Chang & Lin, 2011) avec un noyau

[5] L'oracle donne toujours la bonne réponse si l'une au moins des méthodes considérées la fournit.



(a) Performances du système en utilisant la similarité tf-idf avec et sans **i**nclusion des tokens **n**on **l**exicaux.



(b) Comparaison des performances pour les différentes combinaisons de méthodes. Les 6 méthodes combinées sont tf-idf, Jaccard, BLEU, et ces 3 mêmes en **i**ncluant les tokens **n**on **l**exicaux.

FIG. 1: Précision par énoncé en fonction du nombre d'alternatives pour différentes méthodes. La précision par énoncé donne le nombre d'énoncés test pour lesquels au moins une des réponses proposées est correcte ; par cette définition, sa valeur ne peut donc qu'augmenter avec le nombre de réponses apportées pour chaque énoncé.

polynomial de degré 3 et à coefficient de degré 0 nul. Les résultats sont présentés sur la figure 1b. Comme on pouvait s'y attendre, l'entraînement sur le petit corpus dont nous disposons ne produit pas d'excellents résultats. La performance de la prédiction n'atteint pas 20%, et le modèle détermine seulement la distinction entre les cas où il faut choisir la meilleure méthode et les cas où il vaut mieux ne pas répondre (dû à l'étiquette de classe "aucune"). Les résultats des tests avec le modèle appris parviennent à dépasser légèrement tf-idf à partir de 5 réponses données, mais cela ne traduit pas une amélioration : d'une part, ce score est certainement obtenu grâce à l'abstention de réponse, qui est rendue possible par la présence de l'étiquette "aucune" lors de l'apprentissage, et d'autre part un tel nombre de réponses est déjà élevé si l'on considère que chacune des requêtes de l'utilisateur conduit à un choix multiple de 5 éléments.

7 Conclusion et perspectives

Nous avons appliqué 3 méthodes de calcul de similarité à la recherche de correspondances entre les énoncés en LN et les commandes en LF. Les résultats en termes de précision par énoncé ont atteint 60% de bonnes réponses après entraînement sur un petit corpus, tout en conservant le nombre d'alternatives de réponses proposées à l'utilisateur en deçà d'un seuil raisonnable (< 5 possibilités). Ces résultats valident l'hypothèse posée : les méthodes de mise en correspondance simple permettent d'atteindre un score de 50%.

Beaucoup de méthodes approfondies ou d'approches parallèles peuvent être considérées pour optimiser le score obtenu, comme l'ajout ou le développement de mesures de similarité plus adaptées (Achananuparp *et al.*, 2008), la combinaison de l'apprentissage artificiel et du vote, ou encore l'entraînement d'un modèle pour ordonner ou réordonner les listes d'énoncés similaires. Cependant, même si elles sont utilisables pour amorcer le système, ces méthodes généralisent peu et demandent de grands volumes de données pour être efficaces. De plus, la non fonctionnalité de la relation entre énoncés et commandes introduit des ambiguïtés impossible à résoudre à partir de la seule base de connaissances.

Grâce au système initial que nous avons élaboré, il maintenant possible de rassembler automatiquement plus de données en développant un assistant opérationnel apprenant. Ce dernier peut être vu comme une plateforme de *crowdsourcing* mais aussi, grâce aux performances acceptables que nous avons obtenues, comme un agent intelligent d'aide à la programmation. Nous développons actuellement une interface en ligne dans l'optique d'assurer ces deux objectifs. L'interaction

homme machine située devrait permettre la résolution en temps réel des ambiguïtés rencontrées, grâce à l'aide de l'utilisateur ou aux informations contextuelles du dialogue.

Néanmoins, l'application de l'interaction dialogique réelle pour l'utilisation de ce système pose de nouveaux problèmes liés à la complexité des commandes. Étant donné que toute commande sera apprise par le dialogue, il sera nécessaire de limiter leur longueur pour conserver un système utilisable. Une piste de travail est donc l'extension du modèle d'association, avec notamment la possibilité de référer à des sous-parties déjà connues, afin de permettre l'enseignement incrémental de commandes composées au système.

Remerciements

Je souhaite remercier Sophie Rosset et Gabriel Illouz pour leurs relectures patientes et leurs encouragements, sans oublier Olivier Galibert pour ses excellentes remarques.

Références

ACHANANUPARP P., HU X. & SHEN X. (2008). The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, p. 305–316. Springer.

ALLEN J., CHAMBERS N., FERGUSON G., GALESCU L., JUNG H., SWIFT M. & TAYSOM W. (2007). Plow : A collaborative task learning agent. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, p. 1514 : Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999.

BRANAVAN S., CHEN H., ZETTLEMOYER L. S. & BARZILAY R. (2009). Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 1-Volume 1*, p. 82–90 : Association for Computational Linguistics.

BRANAVAN S., ZETTLEMOYER L. S. & BARZILAY R. (2010). Reading between the lines : Learning to map high-level instructions to commands. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1268–1277 : Association for Computational Linguistics.

CHANG C.-C. & LIN C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3), 27.

GALIBERT O. (2009). *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris Sud-Paris XI.

HUTCHINS W. J. & SOMERS H. L. (1992). *An introduction to machine translation*. Academic Press London.

KUSHMAN N. & BARZILAY R. (2013). Using semantic unification to generate regular expressions from natural language. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* : North American Chapter of the Association for Computational Linguistics (NAACL).

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

POPESCU A.-M., ETZIONI O. & KAUTZ H. (2003). Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th international conference on Intelligent user interfaces*, p. 149–157 : ACM.

TADIĆ M., BEKAVAC B., AGIĆ Z., SREBAČIĆ M., BEROVIĆ D. & MERKLER D. (2012). *Early machine translation based semantic annotation prototype*. Rapport interne, XLike project, www.xlike.org.

TONEY D., ROSSET S., MAX A., GALIBERT O. & BILINSKI E. (2008). An evaluation of spoken and textual interaction in the ritel interactive question answering system. In *LREC*.

VOLKOVA S., CHOUDHURY P., QUIRK C., DOLAN B., REDMOND W. & ZETTLEMOYER L. (2013). Lightly supervised learning of procedural dialog systems.

YU H. & SISKIND J. M. (2013). Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, p. 53–63.

Regroupement de structures de dérivations lexicales par raisonnement analogique

Sandrine Ollinger
CNRS, ATILF, UMR 7118 Nancy, F-54063, France
Université de Lorraine, ATILF, UMR 7118 Nancy, F-54063, France
sandrine.ollinger@atilf.fr

Résumé. Cet article propose une méthode de regroupement de structures de dérivations lexicales par raisonnement analogique. Nous présentons les caractéristiques générales d'un graphe lexical issu du Réseau Lexical du Français, dont nous exploitons par la suite les composantes faiblement connexes. Ces composantes sont regroupées en trois étapes : par isomorphisme, par similarité de relations, puis par similarité d'attributs. Les résultats du dernier regroupement sont analysés en détail.

Abstract. This paper presents a method for merging structures of lexical derivations by analogical reasoning. Following the presentation of general features of a lexical graph from the French Lexical Network, we focus on the weak connected components of this graph. This components are grouped together in three steps : by isomorphism, by relational similarity and finally by attributional similarity. The results of the last merging are analyzed in detail.

Mots-clés : graphe lexical, composantes connexes, analogie, raisonnement analogique, dérivation lexicale.

Keywords: lexical graph, analogy, connected components, analogical reasoning, lexical derivation.

1 Introduction

Cet article rend compte d'une expérimentation réalisée dans le cadre du projet RELIEF (REssource Lexicale Informatisée d'Envergure sur le Français), dont le but principal est le développement d'une modélisation informatisée à large couverture du lexique français : le Réseau Lexical du Français, ou RL-fr. Il s'inscrit dans une volonté d'assister les lexicographes dans la suite de leur travail en proposant des méthodes permettant le développement d'outils d'aide à la rédaction lexicographique [1] et l'enrichissement automatique de la ressource, sous couvert de validation manuelle.

Nous émettons l'hypothèse que le travail lexicographique s'effectue pour une part importante par raisonnement analogique et que le lexique d'une langue regorge de sous-parties analogues [2]. Une telle hypothèse sous-entend que le lexique s'organise en sous-groupes d'unités entretenant des relations privilégiées et que la structuration de ces relations se répète à l'intérieur du lexique.

Nous commencerons ici par présenter la structuration du RL-fr en graphe lexical. Nous ferons ensuite une brève présentation du raisonnement analogique, avant de nous concentrer sur l'application d'un tel raisonnement aux sous-groupes d'unités facilement isolables que sont les composantes faiblement connexes du graphe. Nous discuterons ensuite de la pertinence de l'exploitation des structures détectées pour l'enrichissement du RL-fr.

2 Le Réseau Lexical du Français

Le RL-fr s'apparente à la famille des réseaux lexicaux du type des WordNet (Fellbaum, 1998), de BabelNet (Navigli & Ponzetto, 2010) et de JeuxDeMots (Lafourcade & Joubert, 2010). Il s'en distingue par une visée de description lexico-

1. Ces outils pourront être développés sous la forme de fonctionnalités de l'éditeur lexicographique MvsDicet (Gader *et al.*, 2012).
2. Cette hypothèse est à rapprocher des considérations de (Grimes, 1990) sur les fonctions lexicales inverses et les régularités à motifs dans le lexique, ainsi que des travaux sur l'inférence de relations lexicales dans le réseau lexical JeuxDeMots (Zarrouk *et al.*, 2014).

graphique dans la lignée des dictionnaires virtuels (Atkins, 1996; Spohr, 2012) et des travaux en Lexicologie Explicative et Combinatoire (Mel'čuk *et al.*, 1995)[3]. Sa réalisation, manuelle, a débuté en 2011, au laboratoire ATILF. Sa structuration suit le modèle de système lexical introduit par (Polguère, 2009). Il s'agit d'un graphe orienté encapsulé dans une base de données contenant des entités et des relations de natures variées. Nous choisissons ici de nous concentrer sur le graphe lexical extrait de cette base. Après une brève présentation de ses différents éléments, nous montrerons en quoi il se rapproche d'un graphe «petit monde» et pourquoi une telle particularité nous intéresse.

2.1 Éléments du réseau

Le RL-fr fournit une description détaillée des unités lexicales du français. Conformément au cadre de la Lexicologie Explicative et Combinatoire, une unité lexicale s'entend ici comme une *lexie*, ayant un sens, une forme phonique/graphique et un ensemble de traits de combinatoire (Mel'čuk *et al.*, 1995, p.16). L'approche classique d'une entrée de dictionnaire regroupant différents sens d'une même unité est ici abandonnée au profit d'une approche consacrant une entrée indépendante à chaque sens. Le regroupement des lexies partageant une forme phonique/graphique et liés sémantiquement reste toute fois accessible par le biais de la notion de *vocable*. Un vocable est dit *monosémique* s'il ne correspond qu'à une seule lexie, *polysémique* dans le cas inverse. Les lexies partageant une forme phonique/graphique, mais aucun lien sémantique sont traitées comme des *homonymes* et réparties dans des vocables distincts.

Le graphe lexical issu de la base de données du RL-fr, désormais G_{RLfr}, correspond à l'ensemble fini des unités lexicales du RL-fr, muni de l'ensemble des relations directes entre ces unités.

Il contient trois types de sommets :

- des unités lexicales monolexématiques, ou *lexèmes* : VACHE **I.1** [*Dans le pré, des vaches broutent de l'herbe.*] ;
- des unités polylexématiques, ou *phrasèmes* :
  - des *locutions* : ⌜ PLANCHER DES VACHES ⌝ ;
  - des expressions phraséologiques non lexicalisées, telles que les *clichés linguistiques* : *Comment ça va ?* .

Cette granularité le distingue notamment des graphes lexicaux issus de dictionnaires papier exploités par (Gaume, 2004; Gaillard *et al.*, 2011a), dont les sommets sont des formes phoniques/graphiques. De plus, chaque sommet du G_{RLfr} est associé à une description lexicographique formelle. Celle des lexèmes et des locutions contient un ensemble de caractéristiques grammaticales, une étiquette sémantique (paraphrase minimale), une forme propositionnelle (structure prédicative), une combinatoire lexicale et des exemples d'emplois. Celle des phrasèmes contient, en plus, l'ensemble des lexies qu'ils incluent formellement. La description des expressions phraséologiques non lexicalisées est simplifiée. Elle ne comporte ni combinatoire lexicale, ni étiquette sémantique, ni forme propositionnelle.

Les relations entre ces unités, qui correspondent aux arcs de G_{RLfr}, sont également de trois types :

- des liens de *fonctions lexicales* (Mel'čuk *et al.*, 1995, p.125-152), désormais FL, qui rendent compte de la combinatoire lexicale des unités ;
- des liens de *copolysémie*, désormais CP, qui rendent compte des liens sémantiques entre les unités d'un vocable polysémique ;
- des liens d'*inclusion formelle*, qui rendent compte des différents lexèmes présents dans la forme d'un phrasème.

Les liens de FL sont les plus nombreux. Ils se répartissent en deux grandes classes : ceux mettant en jeu des FL servant à encoder des relations paradigmatiques – comme la synonymie – et ceux mettant en jeu des FL servant à encoder des relations syntagmatiques – comme la cooccurrence entre un nom et ses verbes supports. Selon (Mel'čuk *et al.*, 1995, p.126), «une **fonction lexicale** [=FL] est une fonction au sens mathématique». Elle se note traditionnellement :

F(*lexie 1*) = {*lexie 2*, *lexie 3*, ...}

L'ensemble de lexies {*lexie 2*, *lexie 3*, ...} est alors appelé *valeur d'application* de la FL **F** à son *argument*, la *lexie 1*.

Les liens de FL, pour leur part, mettent en relation les lexies deux à deux. Ainsi, si, comme en (1), la valeur d'application de la FL **Magn** (FL syntagmatique encodant la relation entre une lexie et ses cooccurrents d'intensification) contient une seule lexie, il existe un seul lien. En revanche si, comme en (2), elle comprend plus d'une lexie, il existe autant de liens que de lexies contenues dans cet ensemble.

(1) **Magn**(*coma*) = {*profond*$_{Adj}$ **II**}

3. Une étude comparative du RL-fr et du Wordnet de Princeton est disponible dans les actes de la conférence GWC 2014 (Gader *et al.*, 2014).

(2) **Magn**(*aboyer* I) = {*furieusement* I ; *férocement*}

De plus, chaque classe de FL est subdivisée en familles, correspondant à des types de relations. Ainsi, la famille **Syn** regroupe l'ensemble des FL relevant de la synonymie. Elle comporte notamment les FL suivantes :

– synonymie exacte : **Syn**(*pull*) = {*pull-over*} ;
– synonymie plus riche : $\mathbf{Syn}_{\supset}$(*fixer*) = {*clouer* I} ;
– synonymie plus riche relative au sexe : $\mathbf{Syn}_{\supset}^{\mathbf{sex}}$(*sénateur*) = {*sénatrice*} ;
– synonymie à intersection de sens : $\mathbf{Syn}_{\cap}$(*pull*) = {*sweat, sweat-shirt*}.

Nous parlerons des liens *sortants* d'une lexie pour désigner l'ensemble des liens correspondant à des applications de FL dont elle est l'argument. À l'inverse, nous parlerons de liens *entrants* pour désigner l'ensemble des liens correspondant à des applications de FL dont elle est un élément de la valeur.

2.2 Analyse topologique formelle

Nous émettons l'hypothèse que G_{RLfr} s'organise en sous-groupes d'unités entretenant des relations privilégiées. Une telle structure se rapproche de celle des graphes de données réelles observés dans de nombreux domaines (Watts & Strogatz, 1998; Newman, 2003; Gaume, 2004). Afin de caractériser G_{RLfr} et de déterminer si sa structure est celle d'un tel graphe, dit graphe petit monde, une analyse topologique, appelée *pedigree de graphe*, a été réalisée [4], visible dans le tableau 1.

sommets	21 992	coefficient d'agrégation	0,1327
arcs	42 626	Distribution des degrés entrants	
degré sortant moyen	1,9383	a	-2,3977
boucles	36	r^2	0,9397
arcs multiples	577	Plus grande composante connexe	
arcs symétriques	19 906	sommets	15 302
sommets isolés	3 226	arcs	38 274
composantes connexes	4 311	L	13,0402

TABLE 1: Pedigree du RL-fr

2.2.1 Caractéristiques formelles

La partie gauche du tableau 1 présente une première partie du pedigree de G_{RLfr}. Il s'agit d'un multigraphe orienté, qui comporte 21 992 sommets et 42 626 arcs. Chacun de ses sommets est, en moyenne, la source de près de deux relations. En tant que multigraphe [5], il comporte des boucles et des arcs multiples [6]. Les boucles, peu nombreuses, correspondent à des phénomènes lexicaux particuliers, tels que celui observable pour la lexie POIDS I, qui désigne à la fois une caractéristique physique et le deuxième actant de celle-ci. Les arcs multiples sont un peu plus nombreux. Ils correspondent, également, à des phénomènes lexicaux particuliers, tels que celui qui est observable entre les lexies ABOYER I et JAPPER, où il existe à la fois une relation de quasi-synonymie et une relation d'atténuation.

Les arcs symétriques sont beaucoup plus nombreux [7]. Il s'agit majoritairement d'arcs correspondant à des FL de la famille des synonymes et de dérivations syntaxiques. C'est le cas, par exemple, pour les lexies DANSER et DANSE I, pour lesquelles les relations de nominalisation et de verbalisation suivantes sont encodées [8] : $\mathbf{S}_0$(*danser*) = {*danse* I} et $\mathbf{V}_0$(*danse* I) = {*danser*}.

La proportion de sommets isolés (14,69%) doit être considérée dans le temps. Le tableau 2 montre comment elle a diminué au cours des six derniers mois, tandis que le nombre global de sommets a augmenté. Il montre également que la

4. Nous avons utilisé à cette fin le script *pedigree.py*, développé par Emmanuel Navarro (Gaillard *et al.*, 2011b).
5. À la suite de (Tabourier, 2010), nous ignorons ici la distinction entre multigraphes et pseudographes.
6. Attention, le nombre d'arcs multiple fourni par *pedigree.py* est calculé de façon séquentielle. L'ensemble des arcs du graphe est parcouru et c'est uniquement lorsqu'un arc correspond à un couple de sommets déjà reliés que le compteur d'arcs multiples est incrémenté.
7. Nous appelons arcs symétriques les arcs $a \rightarrow b$ pour lesquels il existe un arc $b \rightarrow a$.
8. Notez que les relations entre les lexies sont considérées en synchronie et qu'il n'est pas question, ici, d'encoder une dérivation morphologique orientée. Les lexies entrant en relation de dérivation syntaxique ne sont d'ailleurs pas nécessairement morphologiquement liées.

connectivité du RL-fr croît plus rapidement que sa nomenclature.

	28/10/13	10/03/14	Évolution
sommets isolés	3 540	3 226	-9%
sommets	20 793	21 992	+6%
arcs	34 922	42 626	+22%
composantes connexes	4 832	4 311	-11%

TABLE 2: Évolution du RL-fr.

La décomposition du graphe en composantes connexes réalisée ici consiste à le partitionner en sous-ensembles maximaux de sommets tous reliés entre eux, sans prendre en considération l'orientation des arcs. Il s'agit d'une décomposition en composantes faiblement connexes. Chaque sommet isolé est considéré comme une composante. Comme le montre le tableau 2, le nombre de composantes tend à diminuer et nous estimons que le RL-fr deviendra entièrement connexe avant d'avoir atteint sa maturité. Cependant, ces composantes constituent des sous-groupes de lexies facilement isolables et nous pensons que leur observation constitue une première étape intéressante dans la recherche et l'analyse de sous-ensembles de connexions lexicales privilégiées et récurrentes.

2.2.2 Graphe petit monde ?

Les graphes petits mondes se distinguent par la concomitance des quatre caractéristiques suivantes :

1. une faible densité, c.-à-d. un petit nombre d'arcs relativement au nombre de sommets ;
2. un coefficient d'agrégation élevé, c.-à-d. une forte probabilité que deux sommets voisins d'un même sommet soient eux-mêmes voisins ;
3. une distribution des degrés sortants et entrants (distribution des probabilités du nombre d'arcs associés à un sommet) qui suit une loi de puissance ;
4. une faible moyenne des plus courts chemins entre deux sommets quelconques du graphe.

Pour déterminer la densité d'un graphe, il faut s'intéresser aux nombres d'arcs (m) et de sommets (n) qui le constituent. Si G_{RLfr} était un graphe simple (sans boucle ni arcs multiples), son nombre maximal d'arcs vaudrait $n \times (n-1)$, soit environ 484 x 10^6. Nous pouvons donc affirmer que sa densité est faible. De plus, selon (Gaume, 2004), le nombre d'arcs observés dans les graphes petits mondes est généralement inférieur à $nlog(n)$. Pour un graphe de 21 992 sommets, il ne doit donc pas excéder 95 495 arcs, soit plus du double du nombre présent dans G_{RLfr} (42 626).

Pour déterminer si G_{RLfr} possède les trois autres caractéristiques des graphes petits mondes, nous nous intéressons à la seconde partie de son pedigree, à droite du tableau 1.

Le coefficient d'agrégation doit être considéré par rapport à celui d'un graphe aléatoire classique[9] de même densité (Newman, 2003), soit 0,00018. Nous pouvons donc affirmer que G_{RLfr} présente un coefficient d'agrégation élevé.

La distribution des degrés et la moyenne des plus courts chemins permettent de se faire une idée sur l'organisation des agrégats au sein du graphe. La distribution des degrés entrants est ici fortement corrélée (0,9397) à une loi de puissance de coefficient -2,3977. Cela signifie que la probabilité pour un sommet quelconque d'avoir beaucoup de voisins est faible et que celle d'en avoir peu est forte. Dans le cas de G_{RLfr}, les lexies fortement connectées sont des lexies carrefour, telles que FAIRE **II.1** [*Il fait du ping-pong.*], jouant un rôle central dans l'organisation du lexique et les lexies très faiblement connectées sont des lexies rares, telles que MONOGRAMME [*Cette assiette est signée PAK, monogramme de Pieter Adriaensz Kocks.*].

(Bollobás & Riordan, 2004) ont montré que la longueur moyenne des plus courts chemins (L) des graphes petits mondes n'excède pas $logn/loglogn$. Une telle valeur signifie qu'il est possible de passer rapidement d'un sommet du graphe à n'importe quel autre. G_{RLfr} n'étant pas connexe, une telle mesure est problématique (Newman, 2003). Une alternative possible consiste à effectuer cette mesure sur la plus grande partie connexe du graphe. Dans le cas de G_{RLfr}, cette composante comporte 15 302 sommets, son L ne devrait donc pas excéder 6,8091. Il est pourtant de près du double (13,0402). Cependant, si nous considérons cette valeur dans le temps, à l'aide du tableau 3, nous constatons que le nombre de sommets de la plus grande composante connexe du RL-fr a augmenté de 17% au cours des six derniers mois, alors que la longueur moyenne de ses plus courts chemins a diminué de 20%.

9. Pour un graphe aléatoire classique, la valeur de C est estimée à $2m/n^2$.

	28/10/13	10/03/14	Evolution
n de la plus grande partie connexe	13 082	15 302	+17%
L de la plus grande partie connexe	16,3112	13,0402	-20%

TABLE 3: Évolution de la L_lcc du RL-fr.

En conclusion, nous pouvons dire que G_{RLfr} est proche d'un graphe petit monde. La moyenne des plus courts chemins de sa plus grande composante connexe est beaucoup plus grande qu'attendu, mais diminue au fur et à mesure du développement de la ressource. L'hypothèse d'un lexique s'organisant en sous-groupes d'unités entretenant des relations privilégiées est, pour sa part, d'ores et déjà confirmée. La jeunesse du RL-fr ne nous permet pas de déterminer avec exactitude si la modélisation du lexique qu'il propose aboutira à un graphe petit monde. Mais elle nous permet d'exploiter son absence de connectivité pour tester l'hypothèse de structures locales analogues et mettre en place un protocole de détection automatique de celles-ci.

3 À la recherche de configurations de dérivations lexicales

Comme nous l'avons vu en 2.2, G_{RLfr} s'organise en agrégats lexicaux [10]. Nous émettons l'hypothèse que la nature de ces agrégats varie en fonction de leur taille. Ainsi, les agrégats de grandes tailles correspondraient à des champs sémantiques, tandis que les agrégats plus denses et plus petits correspondraient à des connexions lexicales particulières.

3.1 Configurations de dérivations lexicales

Le travail présenté ici porte sur la seconde catégorie d'agrégats. Nous pensons que des connexions lexicales particulières se répètent à l'intérieur du graphe et qu'il est possible d'en élaborer des modèles. De tels modèles, que nous nommons *configurations de dérivations lexicales*, pourraient être exploités pour enrichir le RL-fr et intégrer de nouvelles fonctionnalités à l'éditeur lexicographique utilisé pour son développement. Ils seraient constitués d'un ensemble de relations orientées entre lexies, ou plus exactement entre profils de lexies détaillant les caractéristiques nécessaires au déclenchement d'une configuration. Deux axes d'enrichissement automatique seraient alors envisageables : la génération de liens entre lexies correspondant aux profils et l'enrichissement des descriptions incomplètes de lexies d'ores et déjà interconnectées.

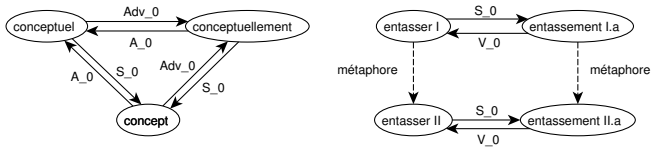


FIGURE 1: Exemples de connexions lexicales.

La figure 1 présente deux exemples d'agrégats lexicaux. Les relations en œuvre dans le premier sont la nominalisation [$\mathbf{S}_0$], l'adverbialisation [$\mathbf{Adv}_0$] et l'adjectivation [$\mathbf{A}_0$]. Une telle structure de relations ne concerne pas que les lexies CONCEPT, CONCEPTUEL et CONCEPTUELLEMENT. Nous pouvons facilement prédire qu'il s'agit d'une configuration de dérivations lexicales, qu'il serait intéressant de pouvoir propager dans le réseau.

Le second exemple illustre un autre type de configuration. Elle met en œuvre, d'une part, une symétrie de dérivations syntaxiques entre deux couples de lexies (nominalisation [$\mathbf{S}_0$] et verbalisation [$\mathbf{V}_0$]) et, d'autre part, une symétrie de dérivation métaphorique entre ces mêmes lexies réorganisées en couples différents. Ici aussi, nous pouvons facilement prédire qu'il s'agit d'une structure récurrente dans le RL-fr, pour laquelle il serait intéressant de définir des profils de lexies.

Nous pensons que de nombreuses autres configurations de dérivations lexicales existent, mettant en jeu des ensembles de

10. Ces agrégats sont à rapprocher des notions de communautés (Borgatti *et al.*, 1990; Navarro *et al.*, 2010) et de motifs locaux (Milo *et al.*, 2004; Wernicke, 2006) que nous ne détaillerons pas ici.

plus de deux lexies. Dans un premier temps, nous avons choisi de chercher à identifier celles en jeu dans les composantes connexes de G_{RLfr}. Ces composantes présentent l'avantage d'être facilement accessibles. Nous leur consacrons donc la présente expérience, qui pose les bases d'une procédure d'identification de configurations de dérivations lexicales par regroupement de microstructures analogues.

3.2 Regroupement de composantes connexes analogues

Comme nous l'avons vu en 2.2, G_{RLfr} est partitionnable en 4 311 composantes connexes, désormais CC. Afin d'identifier parmi elles des configurations de dérivations lexicales, nous avons choisi de procéder par regroupements successifs, visant l'automatisation d'un raisonnement analogique.

3.2.1 Raisonnement analogique

À la suite de (Gentner, 1983; Medin *et al.*, 1990), nous considérons le raisonnement analogique comme un appariement structurel. Les lexies s'apparentent alors à des objets disposant d'un certain nombre d'attributs, éléments de leur description lexicographique, et entretenant des relations, représentées par les arcs de G_{RLfr}. Dans une telle approche, une analogie s'établit entre une CC source et une CC cible. La « bonne qualité » d'une analogie implique que les relations présentes dans la CC source soient mises en correspondance avec les relations de la CC cible. La projection des attributs est, elle, de moindre importance.

Nous empruntons à (Turney, 2006) les notions de similarités de relations et d'attributs, ainsi que de mesures de celles-ci. Rapportée aux données que nous exploitons, la similarité de relations entre deux CC, CC_1 et CC_2, dépend du degré de correspondance entre les relations qu'elles mettent en jeu. La mesure de cette similarité est une fonction qui associe les deux CC à un nombre réel, $sim_r(CC_1, CC_2) \in \Re$. La similarité d'attributs, pour sa part, s'établit entre deux lexies L_1, L_2 et dépend du degré de correspondance entre leurs descriptions lexicographiques. La mesure de cette similarité est une fonction qui associe les deux lexies à un nombre réel, $sim_a(L_1, L_2) \in \Re$.

Tout comme le fait (Lepage, 2003), nous avons choisi de restreindre l'ensemble des valeurs possibles de sim_r et sim_a en les ramenant à des nombres réels compris entre 0 et 1 ; 0 équivalent à l'absence de similarité, 1 à une similarité complète.

À partir de ces considérations, l'identification de CC analogues, susceptibles de faire émerger des configurations de dérivations lexicales, s'est déroulé en trois étapes de regroupement : par isomorphisme (3.2.2), par similarité de relations (3.2.3), puis par similarité d'attributs (3.2.4).

3.2.2 Regroupement par isomorphisme

La première étape a consisté à regrouper les CC par structures mathématiques [11]. Chaque CC a alors été considérée comme un graphe indépendant et comparée aux autres CC en vue d'établir des ensembles isomorphes [12].

Deux graphes sont isomorphes s'ils comportent le même nombre de sommets, le même nombre d'arcs et que leurs arcs se répartissent entre les sommets de manière identique. Ainsi, dans la figure 2, seuls les deux premiers graphes le sont.

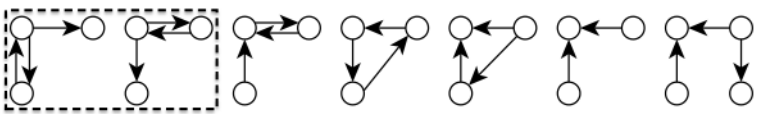


FIGURE 2: Exemple d'isomorphisme de graphes.

Lors de ce traitement, 3 226 lexies isolées et 517 CC ne contenant que deux sommets ont été exclus. 140 CC ne partageant leur structure avec aucune autre ont également été écartées. 428 CC ont été conservées, réparties en 36 groupes.

11. Nous avons choisi de laisser de côté la question de l'existence de sous-structures analogues à l'intérieur d'une ou plusieurs CC et de nous concentrer sur les CC directement manipulables.

12. Nous avons utilisé à cette fin la librairie python *igraph* et sa fonctionnalité *isomorphic*. G_{RLfr} et ses CC étant orientés, cette fonctionnalité a eu recours à l'algorithme VF2 (Cordella *et al.*, 2001).

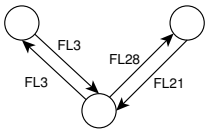
3.2.3 Regroupement par similarités de relations

La deuxième étape a consisté à subdiviser les groupes de CC isomorphes obtenues précédemment en fonction des relations qu'elles mettaient en œuvre.

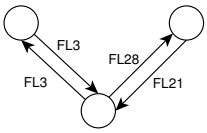
À cette fin, nous avons étiqueté l'ensemble des arcs de la manière suivante :

– les liens de FL sont étiquetés à l'aide du préfixe **FL**, suivi de l'identifiant unique de la FL dans la base de données du RL-fr : un lien de nominalisation $\mathbf{S}_0$ devient **FL21** ;
– les liens de CP sont étiquetés à l'aide du préfixe **CP**, suivi de l'identifiant unique du type de co-polysémie dans la base de données du RL-fr : un lien de métaphore devient **CP1** ;
– les liens d'inclusion formelle sont étiquetés **PH**.

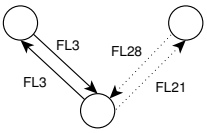
Nous avons ensuite comparé les ensembles d'étiquettes d'arcs des CC. Si les ensembles étaient identiques, nous avons estimé être dans une situation de similarité de relations complète, équivalent à un $sim_r = 1$. Ceci n'est cependant pas nécessairement exact, car l'orientation des relations n'est pas prise en compte dans ce traitement. La figure 3 montre ainsi trois CC regroupées par similarité de relations, alors que seulement deux d'entre-elles le sont réellement.



(a) CC2718



(b) CC2058



(c) CC1496

FIGURE 3: Exemple de regroupement par similarité de relations.

Les relations en jeu dans ces CC sont la synonymie exacte (**FL3**), la nominalisation (**FL21**) et l'adjectivation (**FL28**). Nous pouvons donc en conclure que les CC 2718 et 2058 sont chacune composées de deux noms en relation de synonymie exacte et d'un adjectif en lien de dérivation syntaxique avec l'un d'entre eux, tandis que la CC 1496 est composée de deux adjectifs et d'un nom en lien de dérivation syntaxique avec l'un d'entre eux. Ces conclusions sont confirmées par l'observation des lexies effectivement concernées : TOMBEAU, TOMBE et TOMBAL pour la CC 2718 ; PEUPLE**1**, ETHNIE et ETHNIQUE pour la CC 2058 ; ISRAÉLITE$_{Adj}$, JUÏF$_{Adj}$ **1** et JUDAÏSME pour la CC 1496.

À l'issue de cette étape, 192 CC ont été conservées, réparties en 47 groupes. Les CC isolées ont été exclues.

3.2.4 Regroupement par similarité d'attributs

La dernière étape de notre automatisation du raisonnement analogique a consisté à subdiviser les groupes de CC en similarité de relations complète en fonction des lexies qu'elles mettaient en jeu.

Comme nous l'avons souligné, l'orientation des relations n'a pas été prise en compte dans le précédent regroupement. Nous pensons qu'une ultime étape, de regroupement par similarité d'attributs des lexies, permet de remédier à ce manquement. Ainsi, en comparant l'ensemble des lexies de la CC 2718 à celles de la CC 2058, trois couples de lexies en situation de similarité d'attributs complète seraient obtenus : $sim_a(\text{TOMBEAU, PEUPLE1}) = 1$, $sim_a(\text{TOMBE, ETHNIE}) = 1$, $sim_a(\text{TOMBAL, ETHNIQUE}) = 1$. En revanche, en comparant l'ensemble des lexies de la CC 2718 à celle de la CC 1496, seuls deux couples de lexies le seraient : $sim_a(\text{TOMBE, JUDAISME}) = 1$, $sim_a(\text{TOMBAL, JUIF}_{Adj}\,1) = 1$. Les CC 2718 et 2058 seraient alors regroupées, en tant que CC analogues, relevant d'une même configuration de dérivations lexicales, tandis que la CC 1496 se retrouverait isolée.

L'ensemble des éléments de description lexicographique disponible pour chaque lexie n'est pas pertinent pour effectuer de tels regroupements. Aussi, nous avons choisi de nous concentrer sur deux types d'éléments de description à notre disposition : les caractéristiques grammaticales et la combinatoire lexicale.

Les caractéristiques grammaticales encodées dans le RL-fr sont variées. Il s'agit de caractéristiques fondamentales, de marques d'usage langagier, stylistique et rhétorique, de caractéristiques formelles, de positions syntaxiques et d'informations de linéarisation. Chacune de ces informations ne semble pas pertinente pour déterminer si les connexions lexicales entre lexies sont analogues. Par exemple, la différence entre caractéristiques fondamentales de genre pour deux lexies

nominales n'est significative que dans des cas particuliers, comme la dérivation entre un nom de fonction masculin et son équivalent féminin. Nous avons donc souhaité concentrer notre attention sur les parties du discours. La granularité de ces dernières (50 parties du discours de surface et 9 parties du discours profondes [13]) risquait cependant de nous amener à considérer comme différentes des CC que nous aurions souhaité conserver regroupées. Nous avons donc eu recours à un artifice élaboré en collaboration avec les lexicographes : un ensemble de 13 méta-parties du discours, auxquelles a été rapportée chacune des 59 parties du discours existantes.

La combinatoire lexicale, quant à elle, nous a semblé être un élément essentiel pour déterminer la présence d'analogie entre CC. En effet, elle fournit des informations sur le rôle que joue chaque lexie dans l'organisation générale du lexique. Une lexie verbale comme FAIRE **I**, par exemple, joue un rôle carrefour. Elle n'est associée qu'à un seul lien de FL sortant – encodant une relation de synonymie – mais a 109 liens de FL entrants – dont 82 rendant compte de son utilisation en tant que verbe support et 18 en tant que verbe de réalisation. Elle se distingue en cela d'une lexie verbale plus classique comme KIDNAPPER **I**, qui est associée à 14 liens de FL sortants – encodant des relations de synonymie, de nominalisation et de dérivation sémantique nominale actancielle – et à quatre liens de FL entrants – encodant des relations de synonymie, de verbalisation et de causation. La comparaison de ces deux lexies conduit à penser que trois propriétés méritent d'être considérées comme pertinentes : la nature des liens de FL sortants, la nature des liens de FL entrants et le rapport arithmétique entre le nombre de liens entrants et le nombre de liens sortants. Cependant, la granularité des FL (673 FL distinctes en jeu dans notre graphe) risquait, ici aussi, de nous amener à différencier des CC qui ne devraient pas l'être. Nous avons donc choisi d'établir la comparaison de nature des liens de FL au niveau des familles de FL. De plus, les familles permettant de rendre compte de relations de synonymie (**Syn**), d'antonymie (**Anti**) et de contrastivité (**Contr**) ont été exclues de l'ensemble des points de comparaison. En effet, nous estimons que les liens relevant de ces familles amèneraient à des distinctions inappropriées. Ainsi, deux CC en similarité de relations complète mettant en jeu l'une la lexie FRÉQUEMMENT et l'autre la lexie EXTRÊMEMENT seraient considérées comme différentes, car la lexie FRÉQUEMMENT entretient une relation de synonymie avec SOUVENT, tandis que la lexie EXTRÊMEMENT ne compte aucun synonyme.

Nous avons finalement associé à chaque lexie un ensemble d'attributs constitué de la manière suivante :

- un attribut rendant compte de sa méta-partie du discours ;
- autant d'attributs FLout que de familles de FL en jeu dans l'ensemble de ses liens de FL sortants ;
- autant d'attributs FLin que de familles de FL en jeu dans l'ensemble de ses liens de FL entrants ;
- un attribut rendant compte du rapport arithmétique entre le nombre de liens entrants et le nombre de liens sortants, valant **out+** en cas de supériorité numérique des liens sortants, **in+** en cas de supériorité numérique des liens entrants ou **in=out** en cas d'égalité numérique.

Nous avons ensuite mesuré la similarité d'attributs des lexies en jeu dans chacun des 47 groupes de CC en situation de similarité de relations complète et les avons subdivisés de la manière suivante :

Soit un groupe composé des deux CC CC_1 et CC_2, comportant chacune trois lexies. Chacun des ensembles d'attributs des lexies de CC_1 a été comparé à chacun des ensembles d'attributs des lexies de CC_2. 9 mesures de sim_a ont donc été réalisées, pour lesquelles $sim_a(Lexie_1, Lexie_2) = \frac{nbr\ d'attributs\ communs}{nbr\ d'attributs\ Lexie_1 + nbr\ d'attributs\ Lexie_2}$. Si exactement trois $sim_a = 1$ ont été trouvées, les CC ont été considérées comme analogues et maintenues dans un seul groupe. Si plus de trois $sim_a = 1$ ont été trouvées, le groupe a été maintenu, mais la question de l'analogie des CC est restée en suspens. Enfin, si moins de trois $sim_a = 1$ ont été trouvées, les CC ont été considérées comme non analogues et chacune s'est retrouvée isolée.

Pour les groupes de plus de deux CC le regroupement effectué peut offrir plusieurs possibilités. En effet, une même CC peut partager un nombre différent de $sima = 1$ avec chacune des CC de son groupe. Nous avons alors décidé de regrouper les CC par nombre maximal de $sim_a = 1$. Une fois les CC partageant le plus de $sim_a = 1$ regroupées, le nombre de $sim_a = 1$ partagées par les CC restantes est considéré, etc.

4 Analyse des résultats

À l'issue de la dernière étape, 92 CC ont été réparties en 24 groupes de CC analogues et 80 CC ont été réparties en 20 groupes sans que la question de leur analogie soit tranchée. Les 20 CC restantes ont été isolées. Nous avons observé en dé-

13. Les parties du discours profondes se distinguent des parties du discours de surface. Ainsi, un nom commun (partie du discours de surface) comme la lexie BŒUF **IV** [*Ryan Gosling lui fait un effet bœuf.*] a un emploi appositif, il a donc la valence passive d'un adjectif (partie du discours profonde). Pour une introduction détaillée de ces notions, nous vous invitons à consulter (Mel'čuk, 2006).

tail les CC ainsi regroupées et isolées. L'objectif de cette analyse était à la fois de vérifier la pertinence des regroupements et de se faire une idée sur l'exploitation possible de ces résultats.

4.1 Groupes de composantes analogues

Chacun des 24 groupes constitués automatiquement contient bien des CC analogues. Elles mettent toutes en jeu des ensembles de trois lexies, reliées par des liens de FL. Cinq groupes ont la particularité de rassembler des CC comportant deux lexies de même méta-partie du discours. Cependant, la comparaison des attributs des lexies de ces CC deux à deux aboutit à seulement trois $sim_a = 1$, correspondant aux lexies occupant une place identique dans la structure des CC. Ainsi, les CC de la figure 4 présentent une structure de dérivations lexicales par verbalisation (**FL23**), nominalisation (**FL21**) et dérivation sémantique nominale du premier actant (**FL31**). Elles mettent chacune en œuvre un verbe et deux noms. Dans ce cas précis, les trois similarités d'attributs complètes comptabilisées concernent les couples (SUCCÉDER **I**, FOUILLER), (SUCCESSION **I**, FOUILLE) et (SUCCESSEUR, FOUILLEUR).

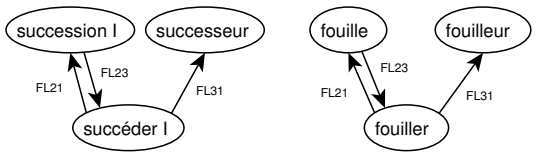


FIGURE 4: Exemple de CC analogues comportant deux lexies de même méta-pdd.

Quelles que soient les CC regroupées, les lexies qu'elles contiennent sont peu décrites dans le RL-fr. Il ne nous semble donc pas pertinent de les exploiter pour définir des profils de lexies susceptibles de déclencher des relations particulières. En revanche, certains groupes de CC présentent des structures imbricables. Cette particularité nous amène à nous interroger sur la granularité des configurations de dérivations lexicales. Faut-il chercher à établir les modèles les plus denses possible et exploiter les imbrications de CC analogues pour enrichir automatiquement les CC comportant le moins de liens de FL ? Le tableau 4 présente les imbrications observées [14] et les suggestions d'ajout de liens qui en découlent.

Parmi les huit groupes de CC comportant des lexies ayant pour méta-partie du discours adjectif, nom et verbe, deux imbrications apparaissent. Parmi les neuf groupes de CC comportant des lexies ayant pour méta-partie du discours adjectif, nom et adverbe, les imbrications sont plus nombreuses. Elles se divisent en deux chaînes distinctes.

Nous avons consulté l'équipe de lexicographes pour savoir si les CC les moins denses étaient toujours valides une fois enrichies. Un seul des cas que nous leur avons présentés a été rejeté. Il s'agit du résultat de la première chaîne d'imbrications concernant des groupes lexies « adjectif, nom et adverbe ». L'ajout d'une relation de dérivation sémantique adjectivale du premier actant (**FL104**) est considéré comme une erreur. Cette observation nous met en garde contre la propagation automatique de mauvais liens.

4.2 Groupes de composantes à l'analogie incertaine

L'observation des 20 groupes de CC dont la question de l'analogie était restée en suspens nous amènent à constater qu'il s'agit de groupes de CC analogues. Deux d'entre eux concernent des relations d'insertions formelles. Ils sortent donc du cadre de la dérivation lexicale qui nous intéresse. Il est cependant intéressant de constater que chacun de ces groupes correspond à une structure syntaxique de locution nominale particulière : **N + de + N** (LEVÉE DE BOUCLIERS) pour l'un, **Adj + N** (TIERS ÉTAT) pour l'autre.

L'ensemble des autres groupes comporte des CC qui mettent en œuvre des relations de synonymie ou d'antonymie. Deux d'entre eux ont même la particularité de rassembler des CC ne comportant que des liens de synonymie. L'ensemble des lexies qui les composent sont en similarité d'attributs complète, $nbr\,de\,sim_a = 1 : (nbr\,lexies)^2$. Ces groupes ne sont pas exploitables pour l'enrichissement automatique du RL-fr. Les CC d'un seul groupe comportent des liens d'antonymie.

14. Les FL en jeu dans ces groupes de CC sont : la dérivation sémantique adjectivale du premier et du deuxième actant (**FL104** et **FL111**), la dérivation sémantique adjectivale potentielle du deuxième actant (**FL107**), la nominalisation simple et prédicative (**FL21** et **FL366**), la verbalisation (**FL23**), l'adjectivisation (**FL28**) et l'adverbalisation (**FL28**).

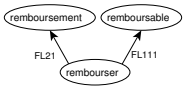
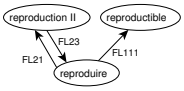
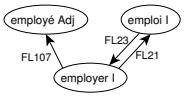
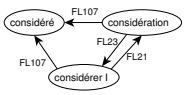
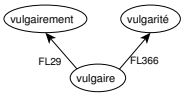
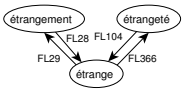
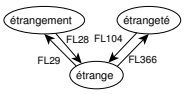
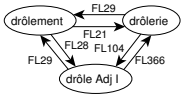
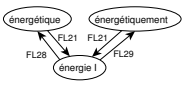
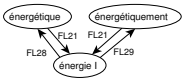
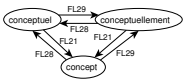
méta-pdd	liens communs	groupe 1	groupe 2	ajouts
adjectif, nom, verbe	FL111, FL21			FL23
	FL107, FL21, FL23			FL107
adjectif, nom, adverbe	FL29, FL366			FL28
	FL28, FL29, FL366			FL104
	FL104, FL28, FL29, FL366			FL21, FL29
	FL28, FL29			2 x FL21
	2 x FL21, FL28, FL29			FL28, FL29

TABLE 4: Imbrication de groupes de composantes analogues.

Il s'agit de CC de quatre lexies, entre lesquelles nous observons huit situations de similarité d'attributs complète. Dans 14 des groupes restants, la dérivation entre un nom masculin et son équivalent féminin [15] est présente dans l'ensemble des CC. Quel que soit le nombre de lexies en jeu dans ces CC, nous observons que le nombre de similarités d'attributs complètes vaut toujours deux de plus, $nbr\,de\,sim_a = 1 : nbr\,lexies + 2$. Seulement deux cas d'imbrication de structures sont constatés dans ces groupes. Le dernier groupe rassemble des CC comportant, entre autres, des relations de synonymie exacte. Ici aussi, le nombre de similarités d'attributs complètes vaut deux de plus que le nombre de lexies interconnectées.

À l'issue de cette analyse, nous pensons que le critère selon lequel deux CC regroupées par similarité de relations sont analogues si leurs lexies sont en situation de similarité d'attributs strictement deux à deux doit être affiné en fonction des relations mises en œuvre.

4.3 Composantes isolées

À l'issue de la dernière étape du traitement, 20 CC se sont retrouvées isolées, alors qu'elles étaient en situation de similarité de relations complète avec au moins une autre CC. Ces CC ne comportent que des liens de FL. En 3.2.3, nous avons émis l'hypothèse que l'absence de prise en compte de l'orientation des relations pouvait être à l'origine de mauvais regroupements. L'analyse des CC isolées nous permet de vérifier cette hypothèse et d'observer d'autres cas de figure.

Afin d'effectuer cette analyse, nous nous sommes intéressée aux couples constitués d'une CC isolée et de chacune des autres CC avec lesquelles elle était précédemment regroupée. Le tableau 5 montre le résultat de cette analyse. L'ensemble des cas de figure rencontrés peut être subdivisé à partir de deux critères : les méta-parties du discours des lexies en jeu dans les CC et la répartition des liens entre ces lexies.

15. Cette dérivation sémantique est encodée à l'aide des FL **Syn**$_{\subset}^{\textbf{sex}}$ et **Syn**$_{\supset}^{\textbf{sex}}$. Sur cette question précise, nous vous invitons à consulter (Delaite & Polguère, 2013).

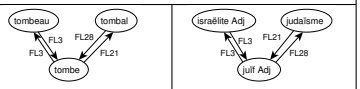
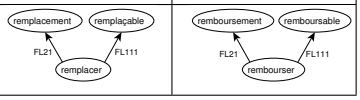
méta-pdd	liens	CC isolées	couples	$sim_a = 1$	exemple
$\neq$	$\neq$	1	2	2/3	
$\neq$	$=$	14	31	0/3	
$=$	$\neq$	4	19	1/3	
?	$=$	1	2	2/3	

TABLE 5: Répartition des groupes de composantes non analogues.

La première catégorie ainsi obtenue correspond à l'exemple illustré par la figure 3 de la section 3.2.3, elle ne permet d'envisager aucun enrichissement automatique. La deuxième n'en permet pas davantage. Elle comporte toutefois un cas intéressant de CC contenant une lexie mal catégorisée [16]. Cette observation a été transmise aux lexicographes et la description de la lexie corrigée. La troisième catégorie concerne, dans deux cas sur trois, des groupes de dix CC, dont une seulement n'est pas analogue aux autres. Cette catégorie nous semble exploitable pour générer automatiquement des liens manquants. Ainsi, dans le cas utilisé comme exemple, un lien de FL23 pourrait être ajouté à chacune des CC. Il correspondrait à la verbalisation du nom dans le cas de la CC non analogue aux neuf autres de son groupe – $\mathbf{V}_0$(*simplification*) = {*simplifier*} – et à la verbalisation des adjectifs dans les neuf autres cas – $\mathbf{V}_0$(*reproductible*) = {*reproduire*}. La dernière catégorie permet la détection d'une anomalie dans le réseau. En effet, elle est due à l'absence de partie du discours dans la description de la lexie REMPLACEMENT.

5 Conclusion

Les résultats de notre expérimentation confirment l'hypothèse d'un lexique s'organisant en sous-groupes d'unités correspondant à des structures de relations récurrentes, identifiables par automatisation du raisonnement analogique.

Les lexies présentes dans les composantes faiblement connexes exploitées ici sont cependant trop peu décrites pour permettre d'établir des configurations de dérivations lexicales comportant à la fois un ensemble de relations et des profils de lexies. De plus, nous n'avons pu observer aucune configuration mettant en jeu des relations de co-polysémie. Pour remédier à ces manquements, nous envisageons de nous désintéresser des composantes connexes au profit des sous-structures moins aisément isolables que sont les motifs locaux (Milo *et al.*, 2004; Wernicke, 2006). De surcroît, une méthode reste à développer en sortie de notre procédure, pour abstraire des configurations de dérivations lexicales des regroupements de structures analogues effectués.

Parallèlement à cela, une évaluation des configurations de dérivations lexicales par les lexicographes doit être mise en place. Elle devra permettre de déterminer des critères de bonne granularité des configurations et d'évaluer les risques de propagation d'erreurs liés à leur exploitation.

Références

ATKINS S. B. T. (1996). Bilingual dictionaries : Past, present and future. In M. GELLERSTAM, J. JÄRBORG, S.-G. MALMGREN, K. NORÉN, L. ROGSTRÖM & C. R. PAPMEHL, Eds., *Proceedings of the 7th EURALEX International Congress*, p. 515–546, Göteborg, Sweden : Novum Grafiska AB.

BOLLOBÁS B. & RIORDAN O. (2004). The diameter of a scale-free random graph. *Combinatorica*, **24**(1), 5–34.

BORGATTI S. P., EVERETT M. G. & SHIREY P. R. (1990). LS sets, lambda sets and other cohesive subsets. *Social Networks*, **12**(4), 337–357.

16. Il s'agit de la lexie verbale DONNER **I.3**, pour laquelle les caractéristiques grammaticales «nom commun» et «masc» étaient encodées.

CORDELLA L. P., FOGGIA P., SANSONE C. & VENTO M. (2001). An improved algorithm for matching large graphs. In *In : 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen*, p. 149–159.

DELAITE C. & POLGUÈRE A. (2013). Sex-Based Nominal Pairs in the French Lexical Network : It's Not What You Think. In *Proceedings of the 6th International Conference on Meaning-Text Theory (MTT'13)*, p. 29–40, Prague, Tchèque, République.

FELLBAUM C. (1998). *WordNet : an electronic lexical database*. Language, Speech and Communication. MIT Press.

GADER N., LUX-POGODALLA V., POLGUÈRE A. *et al.* (2012). Hand-crafting a lexical network with a knowledge-based graph editor. In *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex˜ III)*, p. 109–125.

GADER N., OLLINGER S. & POLGUÈRE A. (2014). One Lexicon, Two Structures : So What Gives ? In *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*, p. 163–171, Tartu, Estonie : Global WordNet Association.

GAILLARD B., GAUME B. & NAVARRO E. (2011a). Invariants and variability of synonymy networks : Self mediated agreement by confluence. In *Proceedings of TextGraphs-6 : Graph-based Methods for Natural Language Processing*, TextGraphs-6, p. 15–23, Stroudsburg, PA, USA : Association for Computational Linguistics.

GAILLARD B., GAUME B. & NAVARRO E. (2011b). Invariants and variability of synonymy networks : Self mediated agreement by confluence. In *Proc. of TextGraphs-6 : Graph-based Methods for NLP*, p. 15–23, Portland : ACL.

GAUME B. (2004). Balades aléatoires dans les petits mondes lexicaux. *Information interaction intelligence*, **4**(2), 39–96.

GENTNER D. (1983). Structure-mapping : A theoretical framework for analogy. *Cognitive Science*, **7**(2), 155–170.

GRIMES J. E. (1990). Inverse lexical functions. In *Meaning-text theory : Linguistics, lexicography, and implications*, p. 350–364. University of Ottawa Press, James Steele edition.

LAFOURCADE M. & JOUBERT A. (2010). Computing trees of named word usages from a crowdsourced lexical network. *Investigationes Linguisticae*, **21**, 39–56.

LEPAGE Y. (2003). *De l'analogie rendant compte de la commutation en linguistique*. PhD thesis, Université Joseph-Fourier - Grenoble I.

MEDIN D. L., GOLDSTONE R. L. & GENTNER D. (1990). Similarity involving attributes and relations : Judgments of similarity and difference are not inverses. *Psychological Science*, **1**(1), 64–69.

MEL'ČUK I. (2006). Parties du discours et locutions. *Bulletin de la Société de linguistique de Paris*, **101**(1), 29–65.

MEL'ČUK I., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Paris/Louvain-la-Neuve : Duculot.

MILO R., ITZKOVITZ S., KASHTAN N., LEVITT R., SHEN-ORR S., AYZENSHTAT I., SHEFFER M. & ALON U. (2004). Superfamilies of evolved and designed networks. *Science*, **303**(5663), 1538–1542.

NAVARRO E., CAZABET R., CAZABET R. & CAZABET R. (2010). Détection de communautés, étude comparative sur graphes réels.

NAVIGLI R. & PONZETTO S. P. (2010). Babelnet : Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, p. 216–225, Stroudsburg, PA, USA : Association for Computational Linguistics.

NEWMAN M. E. J. (2003). The structure and function of complex networks. *SIAM REVIEW*, **45**, 167–256.

POLGUÈRE A. (2009). Lexical systems : graph models of natural language lexicons. *Language Resources and Evaluation*, **43**(1), 41–55.

SPOHR D. (2012). *Towards a Multifunctional Lexical Resource*, volume 141 of *Lexicographica. Series Maior*. De Gruyter.

TABOURIER L. (2010). *Méthode de comparaison des topologies de graphes complexes : applications aux réseaux sociaux*. PhD thesis, Paris 6.

TURNEY P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, **32**(3), 379–416.

WATTS D. J. & STROGATZ S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–442.

WERNICKE S. (2006). Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **3**(4), 347–359.

ZARROUK M., LAFOURCADE M. & JOUBERT A. (2014). About inferences in a crowdsourced lexical-semantic network. In *proc of 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden.

Méthodes par extraction pour le résumé automatique de conversations parlées provenant de centres d'appels

Jérémy Trione[1]
(1) Aix-Marseille Université, CNRS, LIF UMR 7279, 13000, Marseille, France
Jeremy.trione@lif.univ-mrs.fr

Résumé. Dans ce papier nous traitons des résumés automatiques de conversations parlées spontanées. Pour cela nous utilisons des conversations provenant de cas réels d'appels téléphoniques de centre d'appels issues du corpus DECODA. Nous testons des méthodes extractives classiques utilisées en résumé de texte (MMR) ainsi que des méthodes basées sur des heuristiques du dialogue dans le cadre des centres d'appels. Il s'agit de la sélection du tour de parole le plus long dans le premier quart de la conversation, dans l'ensemble de la conversation et dans le dernier quart de la conversation. L'ensemble est évalué avec la métrique ROUGE. Les résultats obtenus soulignent les limites de ces approches « classiques » et confirment la nécessité d'envisager des méthodes abstractives intégrant des informations de structures sur les conversations. En effet, ces premiers résultats montrent que les méthodes heuristiques basées sur la structure produisent des résultats comparables, voir meilleurs que des méthodes telles que MMR.

Abstract. In this paper we speak about automatic spoken conversation summaries. We use conversation from some real cases call from a call center extracted from the DECODA corpus. We test some extractive summary methods used in text summary (MMR) and some dialogue heuristics methods. It's mainly to select the longest speaker turn in different part of the dialogue, the first quarter, the whole dialogue, and the last quarter of the dialogue. All the results are evaluated thanks the ROUGE software. The results show the limits of these classical approaches and suggest that we need some abstractive methods including some structural features of the conversation. In fact, these results show that the structural heuristics based methods are even or better than the classic method like MMR.

Mots-clés : Résumé de conversations parlées, résumé par extraction, ROUGE, corpus DECODA, MMR.
Keywords: spoken conversation summarization, extractive summary, ROUGE, DECODA corpus, MMR.

1 Introduction

Dans les centres d'appels, l'étude des traces (ou « logs ») d'interaction entre conseillers et clients permet d'évaluer le travail des conseillers téléphoniques afin d'optimiser les relations avec les usagers, ou encore faciliter la recherche d'informations sur l'ensemble des appels pour détecter d'éventuels problèmes et extraire des statistiques sur les besoins des usagers du service. Aujourd'hui seule une infime partie des données recueillies dans les centres d'appels est utilisée pour les tâches citées au-dessus (moins de 1%). Le traitement automatique de ces conversations et notamment la génération de résumés pourrait alors permettre de généraliser les traitements et ainsi améliorer les services proposés. Dans la suite de cet article nous nous intéresserons donc à la génération de résumés de conversations téléphoniques provenant de centres d'appels.

Le corpus RATP-DECODA[1] contient des transcriptions de conversations entre des usagers et des conseillers de la RATP. Il contient également des résumés de ces conversations, sous la forme de « synopsis » de quelques lignes, établies par des experts de ce service à des fins d'analyse. L'étude présentée dans cet article consiste à comparer ces synopsis à des résumés produits par des méthodes classiques de résumés automatiques de texte par extraction. En analysant les limites

[1] http://sldr.org/sldr000847/fr

de ces systèmes, nous justifions la nécessité d'enrichir les systèmes automatiques avec des informations de structure sur les conversations dans la perspective produire des résumés par abstraction.

2 Etat de l'art

De façon générale on peut identifier deux types de résumés, les résumés extractifs et les résumés abstractifs. Le résumé extractif est aujourd'hui l'approche la plus répandue, elle consiste à pondérer les phrases selon leur représentativité (Li,et al, 2006). Le résumé abstractif consiste à extraire des informations ou des notions du document à résumer afin de rédiger un tout nouveau texte. Ces informations peuvent être par exemple des concepts d'opinions (Ganesan et al, 2010).

La plupart des travaux réalisés sont des résumés extractifs, certains traitent des journaux télévisés (Ribeiro et Martins de Matos, 2007), d'autres concernent des résumés de réunions comme les travaux de G. Murray (2008), en se basant sur le corpus des réunions de l'ICSI[2] (Janin. et al, 2003) et AMI[3] (Carletta et al, 2006). Il base ses premières expériences sur des systèmes de résumés de textes classiques (en utilisant une version modifiée de MMR), ainsi que sur des approches basées sur la structure des réunions. De la même façon des travaux ont été réalisés sur le résumé d'e-mails comme G. Murray (2008) ou encore J. Lin (2007) principalement sur le corpus d'e-mails d'Enron. Ici deux approches sont étudiées, le résumé de l'ensemble des e-mails d'une conversation, et le résumé des e-mails individuellement. J. Lin et al (2008) utilisent un modèle probabiliste basé sur la méthode de *sentence compression* (K. Knight et D. Marcu, 2000). D'autres travaux réalisés par Xiaodan Zhu et Gerald Penn (2006) portent sur le résumé de conversations spontanées, ils utilisent les données de SWITCHBOARD, qui contiennent des conversations annotées manuellement. Ils utilisent des méthodes de résumés classiques (i.e. MMR) ainsi que des heuristiques de localisation, de prosodie ou de disfluence. Enfin des recherches concernent aussi le résumé de cours d'enseignement oraux (Togashi, 2006) en utilisant aussi des transcriptions manuelles et automatiques et en prenant en compte d'autres concepts audio comme la prosodie pour la génération de résumés. Dans cette étude aussi il est question d'utiliser des techniques de résumés de textes classiques, ainsi que des informations structurelles mais aussi des informations sur l'audio (prosodie et disfluences). Nous noterons aussi que l'ensemble de ces études a été évalué grâce à la métrique ROUGE.

3 Corpus

Dans le cadre de notre étude notre corpus sera composé de 200 conversations téléphoniques issues d'un centre d'appels de la RATP. Ces appels proviennent du corpus du projet DECODA[4] (Bechet, et al. 2012). Chaque conversation est disponible en version audio et textuelle, les transcriptions utilisées ont été réalisées manuellement.

Ces conversations ont été recueillies dans un centre de la RATP sur une période d'une journée. Etant donné qu'elles ont été enregistrées à partir d'un centre d'appels de transport, elles traitent de tous sujets se rapportant de près ou de loin au transport. Cela va de la demande d'itinéraire, aux oublis d'objets sur le réseau, en passant par des plaintes de fonctionnement. Ci-dessous le tableau regroupe les dix sujets d'appel les plus courants :

Raison de l'appel	%
Info trafic	22.5
Itinéraire	17.2
Objets trouvés/perdus	15.9
Souscription aux forfaits	11.4
Horaires	4.6
Billets	4.5
Appels spécialisés	4.5
Aucun sujet particulier	3.6
Nouvel enregistrement	3.4
Information tarifaire	3.0

TABLE 1 : Top 10 des sujets d'appels sur le corpus DECODA.

2 International Computer Science Institute

3 https://www.idiap.ch/dataset/ami/

4 DEpouillement automatique de Conversation provenant de centre D'Appels : http://decoda.univ-avignon.fr/

Notons que la répartition de ces sujets reste la même sur les 200 conversations choisies au préalable.

La durée de l'appel varie entre 55 secondes pour les plus courts et 16 minutes pour les plus longs. Le tableau ci-dessous regroupe plus de détails en ce qui concerne la taille du corpus en termes de mots.

	En moyenne par conversation	Sur l'ensemble du corpus
Nombre de mots	414.1	82819
Nombre de phrases	66.8	13351

TABLE 2 : Détails structurels sur le corpus de DECODA

Afin d'enrichir ce corpus pour notre étude, chaque conversation s'est vu attribuer un minimum de deux résumés réalisés par deux annotateurs différents. Le premier est un expert du domaine ayant réalisé un premier jeu de synopsis afin de se repérer dans les conversations traitées. Le second annotateur est un étudiant n'ayant aucune notion particulière dans la réalisation de résumé. Dans un premier temps aucune contrainte n'a été donnée aux annotateurs, que ce soit des contraintes de temps, de langue ou autre. Le but était d'observer dans quelle mesure les synopsis de chacun pouvaient différer et si des schémas se retrouvaient entre les individus.

La principale contrainte qui nous concernait était la taille, après l'étude des 400 premiers synopsis recueillis nous avons obtenu une taille moyenne pour les résumés de 6% à 7% de la taille de la conversation originale (cette taille est basée sur le nombre de mots de la conversation et du synopsis). De cette seule contrainte découle plusieurs autres. Par exemple le langage utilisé ne devra pas forcément suivre une syntaxe très poussée, les phrases simples et courtes seront alors privilégiées par les annotateurs, de la même façon la quantité de détails rapportés sera elle aussi limitée par la taille du document, effectivement seule l'information principale devra remonter sous peine de dépasser la limite de taille imposée précédemment.

Ci-dessous est présenté un exemple de conversation que l'on peut trouver dans le corpus RATP-DECODA.

Usager : allô bonjour monsieur monsieur je m'excuse de vous déranger je vous appelle de la Haute-Loire pourriez-vous m'indiquer s'il vous plaît le bus qui correspond de la Gare de Lyon à la Gare heu Montparnasse ?

Conseiller : alors vous avez le 91 Madame

Usager : c'est le 91 ?

Conseiller: oui

Conseiller : Gare de Lyon Gare Montparnasse ce sera le 91

Usager : d'accord Monsieur

Conseiller : oui

Usager : et c'est une ligne directe donc ?

Conseiller : c'est une ligne directe tout à fait

Usager : vous êtes très gentil monsieur je vous remercie

Conseiller : je vous en prie

Usager : bonne journée au revoir

Conseiller : au revoir Monsieur bonne journée

Usager : au revoir

Conseiller : merci au revoir

TABLE 3 : Exemple de conversation du corpus DECODA

4 Résumés de conversations

En partant de ce qui existe dans la littérature pour les textes, et en se basant sur quelques heuristiques du dialogue concernant la position de l'information au sein d'une conversation issue d'un centre d'appels, nous appliquons des méthodes classiques utilisées en résumé automatique, l'ensemble de ces méthodes constitue ce que nous appellerons *baseline*. Les conversations utilisées sont issues du corpus du projet DECODA. Ce sont des conversations spontanées, provenant de cas concrets et réels (c'est-à-dire non joués par des acteurs) entre des utilisateurs de la RATP[5] et les conseillers téléphoniques.

Définissons dans un premier temps ce qu'est une conversation dans cette étude ainsi que le résumé qui peut lui être associé.

4.1 Conversation

La définition la plus classique d'une conversation, serait : un échange d'informations entre au moins deux individus portant sur un ou plusieurs sujets précis. Dans notre cas la partie qui va nous intéresser concerne l'échange d'informations. En effet dans le cadre de la rédaction d'un résumé, la détection des informations importantes et intéressantes est primordiale.

D'un point de vu structurel, le découpage d'une conversation ne peut pas se faire grâce à des phrases, puisque la notion même de phrase est difficilement définissable au sein d'une conversation, ceci étant dû à l'absence de ponctuation concrète. Pour nous donner alors une unité de découpage, nous utiliserons le tour de parole. On appelle tour de parole le laps de temps pendant lequel un interlocuteur s'exprime. Chaque tour de parole sera alors susceptible de contenir une certaine quantité d'informations relatives à la conversation.

À la différence d'un texte classique (c'est-à-dire rédigé et réfléchi) une conversation est un échange spontané entre individus, de ce fait les informations au sein de celle-ci peuvent être altérées par de nombreux phénomènes directement liés à la spontanéité. Ce caractère spontané de la conversation introduit du bruit dans les données, il s'agit par exemple de nombreuses répétitions, des changements brusques d'idées, des erreurs de langue et autres. À cela nous pouvons aussi ajouter le fait que nous travaillons essentiellement sur des conversations téléphoniques, la compréhension devient encore plus bruitée par des problèmes liés à la qualité sonore. Afin de pallier ces problèmes, nous utilisons essentiellement les transcriptions manuelles des conversations pour la suite de ce papier.

4.2 Synopsis

En ce qui concerne le résumé, une définition simple serait : Forme abrégée du contenu d'un texte, d'un document, d'un film. Pour une conversation cette forme abrégée doit contenir l'ensemble des informations clés qui ont été abordées au cours de celle-ci. Nous appellerons cette forme « synopsis » pour la suite de l'étude. Chaque synopsis doit être capable de retranscrire les informations véhiculées dans la conversation en un nombre de phrases (ou mots) réduit. Il est important de préciser que la forme de nos synopsis est textuelle et abstractive, c'est-à-dire que la forme résumée de nos conversations ne sera pas une nouvelle conversation plus courte ou une sélection des tours de parole les plus pertinents, mais un court texte rappelant les idées abordées à l'instar des synopsis de films.

Le principal problème de ces synopsis est directement lié à leur nature. En effet un synopsis est le résultat produit par une personne qui souhaite résumer une conversation, mais cette même conversation pourrait très bien être résumée d'une façon totalement différente par un second individu. Afin de limiter ce caractère subjectif dans notre étude, nous nous basons pour notre évaluation sur plusieurs (au moins deux) résumés de références pour une même conversation. À cela s'ajoute la création d'un guide d'annotation en synopsis pour tout annotateur désirant enrichir le corpus, afin que tous les synopsis produits suivent la même orientation.

À ce caractère de subjectivité lié à l'individu, on peut ajouter des variantes dans l'orientation d'un résumé. Dans le cadre des centres d'appels on peut identifier deux catégories de synopsis : des synopsis basés sur le contenu sémantique, c'est-à-dire sur le sujet réel de la conversation, et des synopsis basés sur les interactions entre l'usager et le conseiller, cela correspondrait par exemple à privilégier la durée et l'efficacité du conseiller par rapport au problème même de l'usager. Dans notre étude nous nous intéressons principalement aux synopsis basés sur le contenu sémantique de la conversation.

[5] Régie Autonome des Transport Parisien (http://www.ratp.fr)

4.3 Exemples

Pour illustrer nos propos voici deux exemples concrets de synopsis rédigés par nos annotateurs.

	Annotateur 1	Annotateur 2
Conversation 1	quel bus pour gare de Lyon vers Montparnasse	Demande de renseignement sur une ligne de bus pour aller de de Gare de Lyon à Gare Montparnasse.
Conversation 2	horaires RER E de Meaux à la Gare de l'Est	Demande d'horaires de train de la gare de Maux à la gare de l'Est à une heure donnée

TABLE 4 : Exemples de synopsis

Comme on peut facilement le voir, les synopsis de l'annotateur 2 sont syntaxiquement mieux construits que ceux de l'annotateur 1, mais en termes d'informations, les deux synopsis sont très similaires.

5 Méthodes de résumés

Dans cette partie nous présentons les méthodes classiques de résumés extractifs et heuristiques de sélection selon la position que nous utilisons dans notre étude.

5.1 Maximal Marginal Relevance (MMR)

MMR (Corbonnell et Goldstein, 1998) est largement utilisé dans le résumé de conversation du fait de sa simplicité et de son efficacité. Il sélectionne les tours de paroles les plus riches de sens d'un texte tout en évitant la redondance des informations. Dans le cadre du résumé extractif le score attribué à un tour de parole S est calculé comme suit :

$$MMR(S_i) = \lambda \times Sim_1(S_i, D) - (1 - \lambda) \times Sim_2(S_i, Summ)$$

Où D représente le vecteur document, Summ représente les tours de paroles qui ont été extraits pour le résumé, et λ est une constante utilisée pour ajuster la relation entre la pertinence et la redondance. Les deux fonctions de similarité (Sim_1 et Sim_2) représentent respectivement la similarité entre un tour de parole par rapport à l'ensemble du document et par rapport au résumé courant. Les tours de paroles avec le plus haut score MMR sont sélectionnés itérativement pour générer le résumé, jusqu'à ce que la limite de taille soit atteinte. Ici on sélectionne des tours de paroles jusqu'à ce que le nombre de mots excède 6% de la totalité des mots qui composent le texte de base.

5.2 Heuristiques basées sur le dialogue.

Pour établir notre *baseline* nous avons utilisé quelques heuristiques basées sur les conversations. Nos conversations sont essentiellement les appels provenant de centres d'appels. De ce fait on peut facilement penser qu'il existe une certaine structure qui se répète dans ces appels. On notera que pour la sélection des tours de parole selon leur position, tous les tours de parole sont pris en compte, autant les tours de parole de l'usager que ceux du conseillé.

Le premier constat et le plus évident est que l'usager appelle le centre pour obtenir des informations sur un sujet bien précis. Il est donc normal de penser que l'information essentielle de l'appel se trouve en début de conversation dans les tous premiers tours de parole, juste après les échanges de politesse conventionnels (« bonjour »). D'autre part l'opérateur se doit d'écouter l'appelant pour savoir quelle est sa requête. À partir de ces deux constats nous établissons notre première heuristique : Le résumé (que l'on notera LB pour la suite de ce papier) sera constitué de l'unique tour de parole dont la taille en mot est maximale parmi tous les tours de parole du premier quart de la conversation.

Dans la même optique que précédemment nous prenons cette fois comme résumé (que l'on notera LA pour la suite) l'unique tour de parole dont la taille est maximale sur l'ensemble de la conversation. Celui-ci peut symboliser soit une explication détaillée de la requête par l'usager, soit une explication de la réponse à cette requête par l'opérateur. Dans les deux cas nous espérons aussi révéler une prise d'informations importante dans la conversation.

Afin d'observer si la fin de la conversation contient ou non des informations utiles, nous utiliserons comme résumé (que l'on notera LE) l'unique tour de parole dont la taille est maximale parmi les tours de parole contenus dans le dernier quart

de la conversation. La fin d'une conversation pourrait très bien représenter un rappel global de ce qu'il s'est dit tout au long de l'appel, mais pourrait aussi contenir la solution apportée ou non par l'opérateur à l'usager.

Enfin nous testons un système complètement aléatoire pour se donner une idée de son efficacité dans cette situation. Pour cela nous relevons des tours de parole aléatoirement tant que la contrainte de taille n'a pas été atteinte. Nous noterons ce résumé RS.

6 Résultats et interprétations

Dans cette partie nous présentons d'une part la méthode d'évaluation utilisée, et d'autre part les résultats obtenus à partir de cette dernière.

6.1 Evaluation

L'évaluation de résumés par des humains est une tâche longue et coûteuse. De ce fait pour évaluer nos systèmes nous nous sommes tournés vers une évaluation automatique. Pour cela nous utiliserons l'évaluation ROUGE. Celle-ci est basée sur les occurrences de mots entre les résumés produits automatiquement et les résumés humains dits « idéaux ». ROUGE compare alors le texte produit par les systèmes avec un ensemble de résumés humains sur le même document original. ROUGE-1 à ROUGE-4 sont de simples mesures d'occurrences de n-grammes, qui détectent les mêmes segments entre les résumés produits et ceux de références. ROUGE-L mesure les séquences communes entre les deux types de résumés.

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Où n correspond au nombre de mots contenus dans la séquence de mots à rechercher lors de l'évaluation, $Count_{match}(gram_n)$ représente le nombre maximum de mots qui interviennent à la fois dans le résumé à évaluer et dans les échantillons de résumés de référence.

Lin (Lin, 2003) a montré que ROUGE-1 et ROUGE-2 constituaient un bon indicateur du jugement humain.

6.2 Résultat et interprétation

Nous avons donc lancé nos évaluations avec les résumés générés à partir d'heuristiques basées sur les conversations : le plus long tour de parole en début de conversation (LB), le plus long tour de parole sur toute la conversation (LA) et le plus long tour de parole en fin de conversation (LE), ainsi que la méthode d'extraction de tours de parole aléatoires (RS). Nous utilisons aussi à titre de comparaison un résumé constitué de l'ensemble de la conversation (FT). Les résultats sont regroupés dans le tableau ci-dessous :

	ROUGE-1	ROUGE-2	ROUGE-L
LB	0.183	0.613	0.145
LA	0.175	0.051	0.132
MMR	0.150	0.051	0.119
FT	0.092	0.035	0.070
LE	0.055	0.009	0.043
RS	0.049	0.011	0.041

TABLE 5 : Résultats de l'évaluation

On notera tout d'abord que l'évaluation des synopsis extractifs générés se fait avec des synopsis abstractifs. Cela aurait pour effet de réduire le score de chaque méthode, mais étant donné que toutes les méthodes évaluées ici sont basées sur de l'extraction de tours de parole, la variation des scores sera la même pour toutes. Cependant ces résultats pourraient être difficilement comparables avec d'autres résultats impliquant l'utilisation de ces méthodes dans un autre cadre que celui décrit dans ce papier.

Cela mis de côté on s'aperçoit que les meilleurs résultats sont obtenus avec l'extraction du tour de parole le plus long en début de conversation (LB), cela met bien valeur le fait que dans les conversations du corpus de RATP-DECODA

l'information essentielle est souvent contenue dans les tout premiers échanges entre l'usager et l'opérateur. De la même façon il n'y a que très peu d'informations en fin de conversation comme nous le montre le résultat de l'évaluation LE.

L'extraction du plus long tour de parole en début de conversation et sur toute la conversation donne des résultats assez proches. Cela peut s'expliquer par la nature la conversation. En effet lors d'un appel, le consommateur va souvent expliquer son problème en début d'appel sans que l'opérateur ne l'interrompe. En revanche au cours de la conversation, c'est-à-dire pendant la résolution dudit problème, il y aura souvent de nombreux échanges rapides entre les deux interlocuteurs.

En ce qui concerne les résultats donnés par l'algorithme de MMR, ce-dernier peut facilement être biaisé avec ce genre de données. La nature spontanée et naturelle de la conversation introduit de nombreux bruits, dont la répétition fréquente d'un même terme ou d'une même idée. Au niveau de l'appel cela peut se retranscrire par une période d'incompréhension entre les deux parties, menant alors à de nombreuses répétitions d'informations peu importantes. MMR va alors se concentrer sur ces répétitions et ne renvoyer que des tours de parole de faible importance au niveau information. Cependant les résultats restent tout de même relativement proches des heuristiques basées sur le dialogue les plus efficaces.

6.3 Exemples de résumés produits

Ci-dessous est présenté un exemple de résumé produit en utilisant le MMR et les heuristiques simples de la conversation.

	ROUGE-1	Synopsis
Humain 1	/	perdu téléphone mais pas retrouvé
Humain 2	/	Demande de renseignements sur la perte d'un téléphone portable Alcatel Blanc sur la ligne de bus 20. Non retrouvé.
LB	0.205	allô oui bonjour monsieur je téléphonais pour savoir si par hasard vous auriez trouvé un téléphone portable euh hier dans l'autobus euh la ligne 20 en fin en fin d'après midi disons en soirée plutôt
LA	0.205	allô oui bonjour monsieur je téléphonais pour savoir si par hasard vous auriez trouvé un téléphone portable euh hier dans l'autobus euh la ligne 20 en fin en fin d'après midi disons en soirée plutôt
MMR	0.178	allô oui bonjour monsieur je téléphonais pour savoir si par hasard vous auriez trouvé un téléphone portable euh hier dans l'autobus euh la ligne 20 en fin en fin d'après midi disons en soirée plutôt. ligne numéro 20, un téléphone portable.
LE	0	bonne journée au revoir

TABLE 6 : Différents synopsis obtenus sur un exemple concret.

Comme on peut le voir, le MMR et LB sont les deux méthodes qui semblent le plus efficace en terme de récupération d'informations, avec un point supplémentaire pour le MMR, mais ceci est dû au fait qu'il récupère deux tours de parole pour son synopsis alors que LB, LA et LE n'en récupère qu'une seule. De la même façon LE n'est clairement pas significatif.

7 Conclusion et perspectives

Dans notre étude nous n'avons testé et évalué que des systèmes basés sur des méthodes extractives classiques des résumés de textes afin de se donner une idée de l'efficacité de ces derniers. Cependant nos résumés de références sont abstractifs, ce qui fausse nos résultats dans la mesure où les résultats obtenus ne sont comparables qu'entre eux. Notre système d'évaluation ROUGE atteint aussi ses limites en ne capturant pas la variabilité lexicale, de ce fait il faudra penser à faire une évaluation manuelle.

Cependant on s'aperçoit tout de même d'après les résultats obtenus que les méthodes les plus efficaces sont directement liées à la structure de l'appel. La méthode de résumé par MMR qui fonctionne bien dans les cas des résumés de texte, est ici battue par les méthodes basées sur les heuristiques du dialogue (LB et LA). De ce fait pour la suite de nos études nous nous orienterons plus dans la direction de la structure des appels reçus dans les centres téléphoniques, en essayant de trouver par exemple une segmentation de la conversation en unités logiques.

Remerciements

Ces travaux de recherche ont été financés en partie par l'Union Européenne à travers le projet SENSEI[6] (FP7/2007-2013 - n° 610916 – SENSEI).

Références

JANIN, A., BARON, D., EDWARDS, J., ELLIS, D., GELBART, D., MORGAN, N., ... & WOOTERS, C. (2003). The ICSI meeting corpus. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 1. I-364. IEEE.

GODFREY, J. J., HOLLIMAN, E. C., & MCDANIEL, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (Vol. 1, pp. 517-520). IEEE.

CARLETTA, J., ASHBY, S., BOURBAN, S., FLYNN, M., GUILLEMOT, M., HAIN, T., ... & WELLNER, P. (2006). The AMI meeting corpus: A pre-announcement. *Machine learning for multimodal interaction*. 28-39. Springer Berlin Heidelberg.

RIBEIRO R., DE MATOS DM. (2007). Extractive Summarization of Broadcast News: Comparing Strategies for European Portuguese. *Text, Speech and Dialogue*, 115-122.

MURRAY G., CARENINI G. (2008).Summarizing Spoken and Written Conversations. *EMNLP*, 773-782.

LIN J., ZAJIC D. M., DORR B. J. (2007). Single-document and multi-document summarization techniques for email threads using sentence compression, IPM 44, 1600-1610.

KNIGHT K., DANIEL M. (2000) Statistics-based summarization-step one: Sentence compression, *AAAI/IAAI*, 703-710.

ZHU X., PENN G (2006). Summarization of Spontaneous Conversations, *INTERSPEECH*, 1531-1534.

TOGASHI S., YAMAGUSHI M, NAKAGAWA S. (2006). Summarization of spoken lectures based on linguistic surface and prosodic information, *STL*, 34-37.

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R., ARBILLOT E. (2012). DECODA: a call-center human-human spoken conversation corpus. *LREC*.

GANESAN K., ZHAI CX, HAN J. (2010). . Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. *International Conference on Computational Linguistics 23*, 340-348.

LI W., WU M., LU Q, XU W., YUAN C (2006). Extractive Summarization using Inter- and Intra- Event Relevance. *ACL* 44, 369-376.

LIN C.-Y., HOVY E. H. (2003). Autonamic evaluation of summaries using n-gram co-occurrence statistics, *HLT-NAACL*.

CARBONNELL J., GOLDSTEIN J. (1998), The use of MMR, the diversity-based reranking for reordering documents and producing summaries, *SIGIR*.

[6] http://www.sensei-conversation.eu/

Identification de Contextes Riches en Connaissances en corpus comparable

Firas Hmida
LINA UMR 6241,Université de Nantes
firas.hmida@univ-nantes.fr

Résumé. Dans les études s'intéressant à la traduction assistée par ordinateur (TAO), l'un des objectifs consiste à repérer les passages qui focalisent l'attention des traducteurs humains. Ces passages sont considérés comme étant des contextes « riches en connaissances », car ils aident à la traduction. Certains contextes textuels ne donnent qu'une simple attestation d'un terme recherché, même s'ils le renferment. Ils ne fournissent pas d'informations qui permettent de saisir sa signification. D'autres, en revanche, contiennent des fragments de définitions, mentionnent une variante terminologique ou utilisent d'autres notions facilitant la compréhension du terme. Ce travail s'intéresse aux « contextes définitoires » qui sont utiles à l'acquisition de connaissances à partir de textes, en particulier dans la perspective de traduction terminologique assistée par ordinateur à partir de corpus comparables. En effet, l'association d'un exemple à un terme permet d'en appréhender le sens exact. Nous proposons, tout d'abord, trois hypothèses définissant la notion d'exemple définitoire. Ensuite nous évaluons sa validité grâce une méthode s'appuyant sur les Contextes Riches en Connaissances (CRC) ainsi que les relations hiérarchiques reliant les termes entre eux.

Abstract. Some contexts provide only a simple explanation of a given term even if they contain it. However, others contain fragments of definitions, mention a terminological variant or use other concepts to make it easy the term understanding. In this work we focus on « definitory contexts » that would be valuable to a human for knowledge acquisition from texts, mainly in order to assist in terminological translation from comparable corpora. Indeed, provide the term with an example, makes it possible to understand its exact meaning. First, we specify three hypothesis defining the concept of a defnitory example. Then we evaluate its validity through a method based on the knowledge-Rich Contexts (KRCs) and hierarchical relationships between terms.

Mots-clés : Contextes Riches en Connaissances, CRC, identification de définitions, identification d'exemples, énoncé définitoire, terminologie, traduction terminologique.

Keywords: Knowledge-Rich Contexts, KRCs, mining definitions, minging examples, terminology, terminological translation.

1 Introduction

Ces dernières années, de nombreux travaux se sont tournés vers l'exploitation des corpus comparables. Ces corpus sont définis par Bowker & Pearson (2002) comme étant : des corpus contenant des documents multilingues qui ne sont pas des traductions mais qui partagent certaines caractéristiques telles que la période et le thème. Les principaux travaux en extraction de terminologies bilingues à partir de ces corpus se basent sur l'hypothèse qu'un mot et sa traduction apparaissent souvent dans les mêmes environnements lexicaux (Firth, 1957). Les approches standard (Fung & McKeown, 1997; Rapp, 1999) et par similarité inter-langue (Déjean & Gaussier, 2002; Daille & Morin, 2005), dédiés à l'extraction de lexiques bilingues à partir de corpus comparables, reposent plus particulièrement sur ce principe. En effet, elles permettent, à partir d'un terme à traduire (dans une langue source) d'obtenir une liste ordonnée de traductions candidates (dans une langue cible). Ces traductions sont souvent obtenues en comparant le contexte traduit, en langue cible, du terme source avec l'ensemble des contextes des termes de la langue cible. Les traductions candidates se présentent sous la forme d'une liste plate qui ne fournit pas d'information contextuelle structurée permettant de saisir le contexte d'utilisation du terme visé. Par exemple la méthode standard propose *axillary dissection*, *axillary node dissection* où *dissection* comme traductions candidates (en Anglais) correspondant au terme *curage axillaire* (en Français).
En termes de performance, ces approches de traduction semblent avoir atteint leurs limites, et les recherches les plus récentes se focalisent plutôt sur l'évaluation de ces approches (Laroche & Langlais, 2010). Si l'accès à la terminologie bilingue s'avère indispensable au processus de traduction, le potentiel des méthodes de traduction adoptées doit être amélioré au moyen d'une contextualisation pertinente des termes. En effet, il faut être capable d'appréhender le sens exact

d'un terme et de l'employer correctement. Par exemple, le terme *clavier* désigne dans le domaine musical un instrument de musique, tandis qu'il correspond en informatique à un périphérique d'entrée.
Nous étudions, ici, la possibilité d'associer à un terme à traduire dans une langue source (ou à un terme traduit dans une langue cible) un contexte permettant d'en appréhender le sens exact, dans le but d'aider à sa traduction. Pour cela, nous nous appuyons sur les Contexte Riche en Connaissances (CRC) qui sont introduits par Meyer (2001), dans le but d'améliorer la traduction terminologique assistée par ordinateur.
Ce travail s'inscrit dans le cadre du projet CRISTAL [1] ayant pour objectif de développer une technologie d'extraction de CRC permettant de produire de nouveaux dictionnaires. Ces derniers sont censés pouvoir lister pour chaque terme (ou sa traduction), les CRC et illustrer les connaissances qu'ils contiennent. Par conséquent, ce type de contextes auront pour effet d'attirer l'attention de l'utilisateur (éventuellement non expert en terminologie) sur des phénomènes linguistiques qu'il ne soupçonne pas, et de réduire le temps d'accès à l'information pertinente.
Après un rappel de la terminologie utilisée et de l'état de l'art en identification de contextes riches en connaissances (section 2 et 3), nous présentons notre contribution sur le repérage et l'exploitation de ces énoncés (section 4). Ensuite, nous discutons les résultats obtenus pour conclure dans la (section 5) sur des perspectives à ce travail.

2 Définitions

2.1 Contexte Riche en Connaissances (CRC)

Meyer (2001) fut la première à proposer l'appellation Contextes Riches en Connaissances (CRC), pour désigner les contextes qui permettent de repérer, grâce à des éléments lexico-syntaxiques, des relations (souvent lexicales ou lexico-syntaxiques) entre plusieurs termes. Il s'agit de portions de textes qui contiennent i) des termes d'un domaine spécialisé et ii) des marqueurs explicitant des relations entre ces termes.
Par exemple, la phrase *Les graisses dans le sang sont essentiellement le cholestérol et les triglycérides* est définie comme un contexte riche en connaissances pour le terme *graisse dans le sang*. Dans cette phrase, la présence du marqueur *être_essentiellement* explicite une relation hiérarchique entre les termes *graisse dans le sang*, *cholestérol* et *triglycéride*.
Par ailleurs, les CRC sont intéressants pour l'acquisition de relations sémantiques entre les termes et pour construire des définitions. Dans l'exemple précédent, il existe une relation de définition (par dénotation) entre *graisse dans le sang*, *cholestérol* et *triglycéride*.

2.2 Marqueur de relation

Dans la littérature, les CRC sont souvent identifiés grâce à des marqueurs de relations. Ces derniers servent à repérer et classer finement les relations terminologiques dans un corpus spécialisé. Il s'agit d'un ensemble de mots, d'expressions ou de symboles révélant de façon récurrente une relation terminologique. Par exemple, la structure *telle que* est un marqueur de relation qui exprime une relation d'hyperonymie pouvant relier deux termes comme dans la phrase qui suit : *une hormone telle que l'insuline*... où *hormone* et *insuline* sont deux termes.

2.3 Patron de connaissances (PC)

Les marqueurs de relations sont habituellement utilisés afin d'identifier les CRC. Ils sont le plus souvent modélisés et mis en œuvre grâce aux patrons de connaissances. Ces derniers sont l'une des principales stratégies utilisées dans le but d'isoler les CRC, et ainsi d'écarter les contextes jugés moins utiles. Meyer (2001) et Pearson (1998) ont montré l'intérêt des marqueurs linguistiques indiquant des relations sémantiques entre des termes pour exploiter, dans un but lexicographique, les corpus spécialisés. Cette démarche permet à un terminologue de pointer vers le sous-ensemble des phrases susceptibles de véhiculer les informations souhaitées (Barrière, 2004).
Un patron de connaissances est une expression régulière, formée de mots, de catégories grammaticales ou sémantiques et de symboles, visant à identifier des fragments de texte explicitant des formes lexicales et des catégories grammaticales. Par exemple, dans la phrase *X est un type de Y* (X et Y étant deux termes différents), la structure *est un type de* est un patron de connaissances modélisant le marqueur *être_un_type_de*. Nous appelons PC définitoire, un PC s'articulant autour d'un marqueur de définition.

1. www.projet-cristal.org

Synthèse : Les contextes riches en connaissances sont des contextes contenant un patron de connaissances associé à un marqueur de relation. Reprenons l'exemple précédent :

CRC : *les graisses dans le sang sont essentiellement le cholestérol et les triglycérides*.

PC : *X sont essentiellement Y et Z* ; X, Y et Z sont trois termes.

Marqueur de relation : *être_essentiellement*.

2.4 Terme secondaire

Selon Saggion (2004), un terme secondaire [2] peut être un nom, un verbe ou un adjectif qui cooccurre souvent avec un terme donné dans des définitions provenant de ressources externes, comme le Web par exemple. Saggion (2004) considère les termes secondaires comme un indice pour identifier les définitions. L'intuition derrière l'apparition de cette notion revient à l'observation suivante : cherchant les définitions du mot *Goth* parmi 217 phrases contenant toutes ce mot, Saggion (2004) a remarqué que le mot *subculture* apparaît régulièrement dans les définitions du mot *Goth* du Web. En examinant les 217 phrases de départ, il s'est avéré que seulement 5 d'entre elles étaient des définitions contenant toutes le mot *subculture*. Citons à titre d'exemple la phrase *The goth is a contemporary subculture found in many countries*. À travers cette observation, nous pouvons noter qu'un terme à traduire et son terme secondaire apparaissent souvent dans les définitions et rarement dans les autres contextes (non définitoires).

3 État de l'art

De nombreux travaux se sont intéressés à l'identification automatique de définitions dans différents domaines : terminologie (Gangemi *et al.*, 2003; Velardi *et al.*, 2013), lexicographie (Saggion, 2004), etc. Dans ce travail, nous abordons plutôt l'identification des définitions dans la perspective de l'aide à la traduction terminologique. Nous étudierons, tout d'abord, les approches à base de PC, puis les approches supervisées et semi-supervisées. Dans ces travaux les limites séparant les définitions des CRC n'étaient pas bien déterminées puisqu'ils ont souvent été exploités dans le but d'identifier les définitions.

3.1 Approches à base de PC

Les méthodes basées sur les patrons de connaissances ont été adoptées dans plusieurs travaux de la littérature. Auger (1997), par exemple, a consacré son travail à repérer des définitions avec des PC lexicaux. Quant à Rebeyrolle (2000), elle a utilisé des PC lexico-syntaxiques et a également tenu compte des contraintes liées à la ponctuation. Rebeyrolle (2000) a évalué sa méthode en considérant un corpus étiqueté manuellement par les définitions, comme référence. Elle a obtenu une précision comprise entre 17,95 % et 79,19 % et un rappel entre 94,75 % et 100 % selon les PC utilisés afin de repérer les définitions. Muresan & Klavans (2002) ont proposé leur outil DEFINDER également basé sur des PC (ex. *is defined as, is called*) et des marqueurs de relation comme *()* et - -. Cet outil permet tout d'abord de sélectionner des définitions candidates à partir d'articles médicaux disponibles sur le Web. Ensuite, les définitions complexes sont filtrées par un analyseur grammatical et les sorties de cet outil (DEFINDER) sont comparées à un ensemble de textes étalon. L'évaluation a donné 86,95 % de précision et 75,47 % de rappel.
Malaisé *et al.* (2004) se sont appuyés sur les travaux de Auger (1997) et de Rebeyrolle (2000) pour définir une liste de marqueurs adaptés à leurs corpus. Ils ont davantage pris en considération des marqueurs liés à la ponctuation. Malaisé *et al.* (2004) ont tenté de repérer les définitions pour en extraire ultérieurement les relations entre les termes afin d'aider à la construction d'ontologie. L'évaluation de ce point a soulevé le problème d'avoir une précision faible quand il s'agit des marqueurs linguistiques de reformulation plutôt que des marqueurs lexicaux métalinguistiques [3]. Cette remarque a également été mentionnée par Rebeyrolle (2000).

Saggion (2004) a proposé un système reposant sur l'utilisation des PC et les termes secondaires afin d'identifier les passages définitoires pour ensuite extraire des définitions. Pour cela, il disposait en amont d'une liste de 69 PC. Il a présupposé

2. « *Terms that co-occur with the definiendum (outside the target collection) in definition-bearing passages seem to play an important role for the identification of definitions in the target collection [..] Our methode considers nouns, verbs and adjective as candidate secondary terms.* »(Saggion, 2004)

3. renferment un lexique métalinguistique

que cette liste pourrait servir à identifier les passages et sélectionner les définitions indépendamment des corpus. Parmi ces 69 PC, 36 sont destinés aux questions générales (Qu'est ce que X ? ; X est à définir), et 33 pour traiter des questions spécifiques (Qui est X ? ; X est à définir). Saggion (2004) a sollicité WordNet, Britannica et le Web (ressources externes) afin de déterminer les termes secondaires. Dans WordNet seulement sont considérés les hyperonymes du mot X (à définir) ou les mots les plus fréquents dans son contexte. Tandis que dans Britannica, les termes secondaires ne sont extraits que si les phrases contenant une référence explicite de X. Dans le cas des mots venant d'autres pages Web, la phrase doit contenir un PC pour tenir compte des termes secondaires qu'elle contient. Pour identifier les passages contenant des définitions, Saggion (2004) introduit ses requêtes enrichies par les termes secondaires comme des entrées. Ensuite une phrase est retenue comme une définition si elle contient soit un PC, soit le mot à définir avec au moins trois termes secondaires. Lors de TREC 2003, le système de Saggion (2004) a obtenu un score de 0,236 (meilleur 0,555, moyen 0,192) en termes de F-mesure. Les définitions ont été évaluées par rapport à des définitions de référence.
Barrière (2004) considère les PC comme un outil "clé" permettant de repérer les CRC. Elle a organisé les PC présents dans les définitions du dictionnaire numérique American Heritage First Dictionary (AHFD) selon leurs types et la relation sémantique qu'elles expriment. Les PC sont classés en trois grandes catégories : **statiques** donnant des contextes qui ne sont pas liés à des événements, **dynamiques** contenant les relations causales et temporelles, et **événementielles** introduisant des événements (intrinsèques/extrinsèques). Ensuite, elle a analysé la généralité/spécificité des PC par rapport aux domaines et aux relations sémantiques exprimées par ces PC dans le domaine de la plongée sous-marine (1 million de mots). Barrière (2004) a remarqué que les relations sémantiques d'hyperonymie et méronymie, par exemple, sont exprimées de la même façon dans le corpus de la plongée sous-marine que dans le dictionnaire AHFD. Cependant, il existe d'autres relations telles que *risk prevention* qui sont exprimées avec de nouveaux PC. Celles-ci sont considérées comme spécifiques au domaine de la plongée sous-marine : les PC explicitant ces relations n'apparaissent que dans le corpus du domaine étudié. Afin de repérer les énoncés définitoires, les travaux présentés dans cette section se sont basés sur des marqueurs, habituellement lexico-syntaxiques, signalant des CRC. Si Barrière (2004) et Rebeyrolle (2000) ont étudié les structures linguistiques exprimant la définition dans les textes, d'autres travaux comme celui de Saggion (2004) ont choisi d'exploiter les termes secondaires comme étant un indice de définition. Cette notion de termes secondaires étant similaire à celle de la collocation (limitée aux énoncés définitoires), elle permet d'identifier le contexte définitoire « typique ».

3.2 Approches supervisées

Les méthodes utilisant les patrons de connaissances ont mis en évidence plusieurs difficultés. En effet, les définitions peuvent être exprimées de manières différentes. Ceci rend difficile l'obtention d'un ensemble de PC permettant d'identifier toutes les définitions. C'est pour cette raison que plusieurs recherches se sont orientées vers des méthodes moins dépendantes des PC. Fahmi & Bouma (2006), par exemple, ont proposé une méthode reposant sur l'apprentissage supervisé afin de repérer les définitions dans un corpus allemand du domaine médical. Ce corpus est constitué des pages de Wikipedia. Ils ont commencé par extraire toutes les phrases contenant le marqueur *to be* afin d'obtenir des définitions candidates. Parmi ces phrases, les définitions sont isolées manuellement. Ensuite, Fahmi & Bouma (2006) ont déduit les traits permettant de distinguer les définitions des autres phrases. Il s'agit de la position de la phrase, la distribution des mots et des bigrammes, ainsi que des traits syntaxiques comme le type du déterminant et la position du sujet dans la phrase. Ils ont alors intégré ces traits dans trois systèmes d'apprentissage différents : Bayésien naïf, SVM (Support Vector Machine) et MaxEnt (Maximum Entropy). Les résultats obtenus varient de 77 % à 92.3 % (le meilleur fourni par MaxEnt) en termes de précision. Quelques années plus tard, Westerhout (2009), inspirée par Fahmi & Bouma (2006), a proposé une méthode hybride dans laquelle elle a eu recours à l'apprentissage supervisé et aux patrons de connaissances. Ainsi, elle a exploité à la fois des informations linguistiques (telles que la l'étiquette grammaticale) et des informations structurelles. L'auteur a ajouté le type des noms et la structure du document aux traits de son système. Les meilleurs résultats (F-mesure=0.63) proviennent du patron *is a*.

3.3 Approches semi-supervisées

Peu de travaux, comme celui de Kilgarriff *et al.* (2008), se sont penchés sur l'identification des exemples. Kilgarriff *et al.* (2008) présente GDEX, un outil qui propose aux lexicographes plusieurs exemples permettant de définir un mot donné. Il s'est référé à Atkins & Rundell (2008) pour qualifier un bon exemple comme étant : lisible et informatif. Afin de concrétiser ces critères, il a proposé des traits privilégiés (positifs) tels que la longueur de phrase, la présence de la collocation souhaitée dans la clause principale de la phrase. Il a également considéré la présence de pronoms et d'anaphores comme des traits pénalisant (négatifs). Pour ce faire, il a effectué des jeux de test basées sur des comparaisons avec son corpus

d'apprentissage considéré comme un corpus de référence. D'après les résultats, la longueur de la phrase et la fréquence des mots sont les traits qui influencent principalement le choix des exemples. Toujours dans le but d'aider les lexicographes, Didakowski *et al.* (2012) se sont également intéressés à l'extraction des exemples. Même si leur approche était similaire à celle de Kilgarriff *et al.* (2008), les traits dans Didakowski *et al.* (2012) ont été exploités afin d'associer des scores aux phrases.
Pour faire face au problème de portabilité des PC, Reiplinger *et al.* (2012) ont proposé une méthode semi-supervisée. Ils ont appliqué des couples de termes liés par des relations sémantiques afin d'extraire des définitions à partir des articles ACL Anthology Reference Corpus (ACL ARC) (Bird *et al.*, 2008). À partir d'un ensemble limité de paires de terme-définition et des PC définis auparavant, leur système a acquis de nouvelles paires terme-définition ainsi que de nouveaux PC. Les résultats obtenus montrent que cette technique peut être appliquée pour extraire des définitions. Quant à Navigli & Velardi (2010), ils considèrent comme définition toute phrase pouvant être associée à un automate également associé à une définition et généré en amont. Cet automate correspond à un patron de connaissances s'intéressant plutôt à la structure de la phrase. Navigli & Velardi (2010) disposent d'un corpus de définitions extraites de Wikipedia et étiquetées manuellement lui permettant d'en déduire des modèles de définitions sous forme d'automates (finis déterministes).
La plupart des méthodes abordent le repérage automatique des définitions avec des structures linguistiques définies auparavant (excepté Kilgarriff *et al.* (2008) et Didakowski *et al.* (2012)). On distingue principalement deux types d'approches : linguistiques et informatiques. Barrière (2004), par exemple, s'est donnée explicitement pour objectif la description de patrons de connaissances signalant des énoncés riches en informations sémantiques. D'autres méthodes ont essayé de trouver un compromis entre les deux, c'est-à-dire soit en traduisant la structure linguistique en modèle générique (comme Navigli & Velardi (2010)), soit en utilisant des méthodes informatiques en parallèle avec des méthodes linguistiques, telles que les PC.

Nous poursuivons notre travail sur le même principe qui est d'exploiter les PC et la présence d'autres termes comme un indice de définition. Nous proposons tout d'abord de profiter des relations hiérarchiques reliant les termes entre eux, puis ensuite de sélectionner les définitions à l'aide des marqueurs de définitions.

4 Détection de CRC

Rappelons que notre objectif consiste à exploiter les corpus comparables dans le but d'aider à la traduction terminologique. Pour cela nous proposons une méthode permettant d'illustrer un terme à traduire dans une langue source (ou un terme traduit dans une langue cible) pour aider à sa traduction. En effet, l'association d'un exemple à un terme permet d'en appréhender le sens exact.

4.1 Notion d'exemple définitoire (ED)

Nous postulons tout d'abord que le contexte d'apparition d'un terme est un exemple candidat. Cependant, les contextes d'apparition d'un terme, qui peuvent être nombreux, ne sont pas tous des exemples pertinents. Voici par exemple 3 contextes du terme *diabète de type 2* :

(a) *En ce qui concerne le risque que l'enfant ait lui-même un diabète de type 2 à la cinquantaine ou fasse un diabète gestationnel s'il s'agit d'une fille, on a pendant longtemps estimé qu'il dépendait essentiellement de la transmission par la mère d'un capital génétique favorisant le diabète de type 2 (même risque dans ce cas que si le père de l'enfant à naître a un diabète de type 2) mais des études récentes semblent en faveur d'un rôle possible du niveau de glycémie pendant la grossesse lorsque le diabète n'est pas maîtrisé.*

(b) *Dans le monde, 150 millions de personnes souffrent de diabète, dont 90 % de diabète de type 2.*

(c) *Le diabète de type 2 est un diabète qui s'accompagne souvent d'un excès de poids.*

Nous allons proposer des caractéristiques permettant de considérer un contexte comme un exemple pertinent.

4.1.1 Aspect définitoire

Hypothèse 1 : *"un exemple est illustré par une définition explicite"*

Le contexte doit donner des renseignements sur le terme visé afin d'en appréhender le sens exact. Lehmann & Martin-Berthet (1998) distinguent trois types de contexte :

1. Les contextes définitoires sont des fragments textuels servant à enrichir des définitions canoniques [4] du terme.
   Par exemple *le diabète gestationnel est un diabète qui se développe pendant la grossesse* est une définition canonique représentant un contexte définitoire pour le terme *diabète gestationnel*.
2. Les contextes encyclopédiques donnent une information complémentaire sur le terme.
   La phrase *le diabète gestationnel concerne entre 1 et 4 % des grossesses* ne fournit pas une définition mais donne plutôt une information supplémentaire sur le terme concerné.
3. Les contextes linguistiques caractérisent le comportement du terme dans le discours notamment par rapport aux collocations.
   Par exemple *une glycémie, mesurée à jeun, avec une valeur normale, ne permet pas d'exclure le diabète gestationnel,* permet d'associer *glycémie* au terme *diabète gestationnel* dans un discours spécialisé.

Debora Farji-Haguet [5] (traductrice chargée de cours en traduction technique et en terminologie à l'Université de Paris 7) affirme que les traducteurs s'intéresseront plus aux définitions qu'aux autres types de contexte afin de différencier deux termes du même domaine tels que *diabète de type 1* et *diabète de type 2*.
Dans ce travail, nous nous intéressons aux aspects définitoires que peuvent révéler les contextes. D'autre part, les énoncés définitoires peuvent être repérés automatiquement dans le texte grâce à des structures signalant des segments de définitions (Rebeyrolle & Tanguy, 2000). Par exemple, parmi les contextes (a), (b) et (c), seulement (c) correspond à une structure définitoire exprimée par le verbe *est* explicitant la définition. Notre objectif vise les contextes contenant des indices de définitions.

4.1.2 Niveau phrastique

Hypothèse 2 : *"un exemple est limité à une phrase"*

Plusieurs travaux traitant des contextes tels que Meyer (2001) et Martínez *et al.* (2009) ont choisi de travailler sur des unités textuelles plus petites que les phrases. Ces contextes sous-phrastiques risquent de ne pas donner assez d'informations sur le terme en question. En outre, dans les paragraphes le risque est plus élevé d'extraire des informations imprécises concernant ce terme. Afin d'éviter cet écueil nous avons fait le choix de travailler uniquement sur des phrases entières comme étant un compromis, et ainsi considérer le contexte (a) comme un contexte non-pertinent.

4.1.3 Présence d'hyperonyme

Hypothèse 3 : *"un exemple contient le terme à illustrer et son hyperonyme"*

Selon Lehmann & Martin-Berthet (1998), un terme est souvent défini par recours à son hyperonyme. En effet, définir un terme revient à déterminer d'abord sa classe générale, puis le spécifier par rapport aux autres termes appartenant à celle-ci. On rencontre deux types de définitions : la définition intensionnelle (non intentionnelle) qui décrit le terme par son hyperonyme (habituellement dans les définitions canoniques)(ex. *un diabète gestationnel est un diabète*), contrairement à la définition par extension qui donne l'hyponyme du terme (ex. *le diabète contient le diabète de type 1, diabète de type 2...*) (Lehmann & Martin-Berthet, 1998). Dans notre travail nous nous intéressons aux contextes contenant un hyperonyme du terme en question étant donné que bon nombre de travaux considèrent la relation d'hyperonymie comme la plus fréquemment utilisée pour le définir (Green *et al.*, 2002). Même si le contexte (b) respecte cette contrainte, *diabète*, qui est l'hyperonyme de *diabète du type 2*), n'explicite pas la définition de ce dernier.

4.1.4 Synthèse

À partir des trois hypothèses précédentes, nous proposons de définir la notion d'exemple définitoire (ED) considéré comme un contexte pertinent permettant de préciser le sens du terme concerné comme étant **une phrase contenant un hyperonyme définissant le terme visé**. Par conséquent, parmi les exemples précédents ((a), (b) et (c)) seulement (c) sera

4. La définition canonique consiste à désigner d'abord le genre (la classe générale), dont relève le référent du terme à définir, puis à spécifier les différences qui le séparent des autres espèces appartenant au même genre (Lehmann & Martin-Berthet, 1998).
5. http://hosting.eila.univ-paris-diderot.fr/~juilliar/sitetermino/cours/cours_total_deb_john_2003.htm

retenu comme étant un exemple définitoire associé au terme *diabète du type 2*.

Néanmoins, les définitions du même terme peuvent varier d'un dictionnaire à l'autre. D'une part en raison du choix de l'hyperonyme et d'autre part parce que parfois l'hyperonyme direct peut être ambigu. La problématique est donc la suivante : quel hyperonyme faut-il choisir afin de mieux définir le terme ?

4.2 Identification d'exemples définitoires

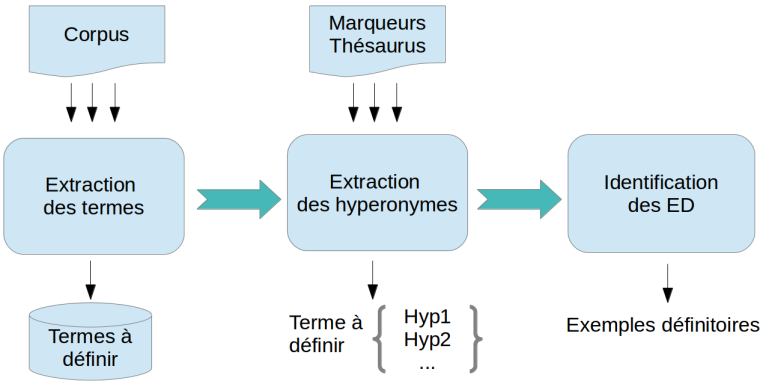


FIGURE 1 – Méthode d'identification des exemples définitoires.

Afin d'évaluer la validité des hypothèses précédentes et vérifier si notre définition de l'exemple définitoire est très restrictive, nous avons suivi la méthodologie illustrée par la figure 1. Cette méthodologie se décompose en 3 étapes :

1. **Extraction terminologique :** Dans le but d'associer à chaque terme du corpus ses exemples, nous avons d'abord extrait automatiquement la terminologie. Ensuite, nous avons filtré les termes les moins fréquents du corpus. Dans la suite, nous désignons par T la liste finale des termes filtrés.
2. **Extraction des hyperonymes :** En exploitant les relations hiérarchiques reliant les termes entre eux, nous déterminons, pour chaque terme appartenant à T, tous les hyperonymes présents dans le corpus étudié.
3. **Sélection des exemples :** Les phrases contenant un terme avec un de ses hyperonymes obtenus pendant l'étape précédente, seront considérées comme des exemples définitoires candidats. Ensuite, ces exemples candidats sont évalués manuellement dans le but d'identifier ceux qui sont des exemples définitoires.

4.3 Évaluation

4.3.1 Expérimentations

Nous avons appliqué la méthodologie décrite dans la section précédente sur un corpus français spécialisé. Il s'agit d'un corpus relatif à la thématique « diabète et alimentation » composé d'articles scientifiques contenant 206 460 mots (Goeuriot, 2009), dans lequel des définitions de 137 termes (simples et composés) ont été manuellement annotées (Nakao, 2010).

1. **Extraction terminologique :** Nous avons tout d'abord utilisé TermSuite [6] afin d'extraire respectivement les termes simples et les termes composés. Ensuite, nous avons considéré seulement les termes qui apparaissent plus de 10 fois dans le corpus pour les termes simples, et plus de 5 fois pour les termes composés. Ce choix est supposé maximiser la probabilité d'avoir des définitions parmi les phrases où occurre le terme. La terminologie obtenue contient, après ce filtrage, 677 termes composés et 809 termes simples.
2. **Extraction des hyperonymes :** Afin de pouvoir comparer les ED avec les définitions de référence (annotées dans le corpus), nous nous sommes intéressé seulement aux termes dont la définition est manuellement identifiée par Nakao (2010). Deux stratégies ont été adoptées de manière indépendante pour identifier les hyperonymes des termes

6. https ://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite

définis : l'exploitation des marqueurs d'hyperonymie et l'utilisation d'un thésaurus.
À défaut de disposer de toutes les relations hiérarchiques reliant les termes entre eux, nous avons eu recours au thésaurus MeSH [7] afin de déterminer pour chaque terme extrait et présent dans le MeSH ses différents hyperonymes (jusqu'au quatrième niveau supérieur si possible). Par exemple si on considère la fonction *Hyperonyme (cholestérol alimentaire)* indiquant l'hyperonyme direct du terme *cholestérol alimentaire*, on obtient :

Hyperonyme (cholestérol alimentaire) = cholestérol : hyperonyme de niveau 1
Hyperonyme (cholestérol) = stérol : hyperonyme de niveau 2
Hyperonyme (stérol) = lipides : hyperonyme de niveau 3

D'autre part, le thésaurus permet éventuellement d'enrichir la terminologie avec les termes qui sont présents dans le corpus mais qui n'ont pas été extraits par l'outil d'extraction, comme *stérol*. Après avoir projeté la terminologie filtrée sur le thésaurus MeSH, seulement 18 termes simples et 10 termes composés sont retenus.
Concernant les marqueurs d'hyperonymie, nous avons utilisé une liste de 33 marqueurs de relation issus de la thèse de Séguéla (2001), dont 22 sont présentés dans la table 1. Contrairement aux marqueurs utilisés dans l'état de l'art, ceux que nous avons appliqués intègrent le terme et son hyperonyme. Ceci s'explique par le fait que nous nous intéressons à associer à un terme son ED (considéré comme CRC), tandis que les travaux de la bibliographie identifient les CRC en ne prenant pas en compte la présence d'un terme précis.

X_EST_UN_Y	ON_VERBE_DEFINITOIRE_X_UN_Y
X_ETRE_UNE_SORTE_DE_Y	X_EST_UN_Y_TRES
X_EST_LE_Y_LE_PLUS	X_ETRE_LE_PLUS_DE_TOUS_LES_Y
X_ET_AUTRES_Y	X_ET_D_AUTRES_Y
Y_ET_ADVERBE_DE_SPECIFICATION_X	X_VIRGULE_LE_Y_LE_PLUS
X_VIRGULE_LE_PLUS_ADJ_DES_Y	LE_PLUS_ADJ_DES_Y_VIRGULE_SOIT_X
Y_VIRGULE_ADVERBE_DE_SPECIFICATION_X	Y_VIRGULE_X_ADVERBE_DE_SPECIFICATION
Y_VIRGULE_ADVERBE_INCLUSION_X	Y_VIRGULE_ADVERBE_EXCLUSION_X
Y_PARENTHESES_X	INCLUSION_Y_VIRGULE_X
Y_PARMI_LESQUELS_X	Y_EXEMPLIFICATION_X
Y_ENUMERATION_X	Y_DEUX_POINTS_X

TABLE 1 – Exemples de marqueurs d'hyperonymie utilisés (X étant le terme et Y son hyperonyme)

4.3.2 Résultats

Terme	ex. sans hyp		avec hyp1		avec hyp2		avec hyp3		avec hyp4	
	ED	occ.	ED	occ.	ED	occ.	ED	occ.	ED	occ.
diabète de type1	2	32	4	45	0	0	0	0	0	0
diabète de type 2	2	46	6	69	0	0	0	1	0	0
diabète gestationnel	2	37	3	16	0	0	0	0	0	0
cholestérol alimentaire	0	5	6	6	0	0	0	0	0	0
ration calorique	0	14	0	1	0	0	0	0	0	0
excès de poids	6	49	0	1	0	0	0	0	0	0
régime alimentaire	0	14	0	0	0	0	0	0	0	0
perte de poids	0	16	0	0	0	0	0	0	0	0
prise de poids	0	31	0	0	0	0	0	0	0	0
activité physique	-	-	0	0	0	0	0	0	0	0

TABLE 2 – Résultats d'identification des exemples définitoires : hyperonymes extraits à partir du MeSH : cas des termes composés

La première colonne des tables 2 et 4 contient les termes définis dans le corpus étudié, qui sont obtenus après la projection sur le thésaurus. Suite à l'utilisation des marqueurs d'hyperonymie, les termes définis, et qui sont retenus, sont illustrés par

7. http ://mesh.inserm.fr/mesh/

Terme	nbr. d'hyperonymes	nbr. d'ED
diabète de type 1	1	1
diabète de type 2	4	1
diabète non insulino-dépendant	1	1
autosurveillance glycémique	1	0
diabète gestationnel	1	0
united kingdom prospective diabetes study	1	0
diabetes control and complications trial research group	1	0

TABLE 3 – Résultats d'identification des exemples définitoires : hyperonymes extraits en utilisant les marqueurs de relation : cas des termes composés

la première colonne des tables 3 et 5. La seconde colonne (ex. sans hyp) des tables 2 et 4 représente le nombre d'exemples définitoires (comme défini en section 4.1) identifiés parmi l'ensemble des phrases contenant le terme indépendamment de ses hyperonymes i (i étant le niveau d'hyperonymie). Nous avons considéré un exemple définitoire comme valide, s'il a été annoté dans le corpus comme une définition associée au terme en question. Par exemple, le terme *diabète de type 1* apparaît dans 32 phrases dont seulement 2 sont des exemples définitoires sans présence de ses hyperonymes. Ces deux exemples définitoires sont marqués dans le corpus étudié comme des définitions du terme *diabète de type 1*. Ce dernier co-occure également avec son hyperonyme direct (i.e *diabète*) dans 4 exemples définitoires parmi 45 phrases (colonne 3 de la table 2). Par contre le couple (*diabète de type 1, Hyperonyme 3(diabète de type 1)*) n'apparaît pas dans le corpus.

Terme	ex. sans hyp		avec hyp1		avec hyp2		avec hyp3		avec hyp4	
	ED	occ.	ED	occ.	ex.	occ.	ED	occ.	ED	occ.
glycémie	4	622	1	42	0	0	0	0	0	0
agpi	2	31	0	8	0	3	0	0	0	0
dnid	0	8	0	4	0	0	0	0	0	0
saccharose	2	23	0	0	0	0	0	0	0	0
amidon	0	23	0	0	0	0	0	0	0	0
pancréas	1	76	0	0	0	0	0	0	0	0
diabète	4	649	0	0	0	3	0	0	0	0
artère	0	89	0	0	0	0	0	0	0	0
athérosclérose	0	19	0	0	0	0	0	0	0	0
insuline	6	498	0	0	0	42	0	0	0	0
hyperglycémie	0	65	0	0	0	0	0	0	0	0
acétonurie	1	10	0	0	0	0	0	0	0	0
cétonurie	1	16	0	0	0	0	0	0	0	0
hypoglycémie	0	92	0	0	0	0	0	0	0	0
glycosurie	0	9	0	0	0	0	0	0	0	0
fructosamine	1	5	0	0	0	0	0	0	0	0
cholestérol	0	107	0	0	0	0	0	0	0	0
glucose	0	103	0	0	0	3	0	0	0	0

TABLE 4 – Résultats d'identification des exemples définitoires : hyperonyme extrait à partir du MeSH : cas des termes simples

La deuxième et la troisième colonne des tables 3 et 5 contiennent le nombre d'hyperonymes associés à chaque terme, en fonction des patrons utilisés. Par exemple, dans le cas du terme *diabète de type 2*, 4 hyperonymes sont trouvés et un seul exemple définitoire a été identifié.
Rappelons que nous sommes partis de l'hypothèse qu'un exemple définitoire est une phrase contenant à la fois le terme et son hyperonyme explicitant une définition. Dans un premier temps, nous nous sommes basés seulement sur la présence du terme et son hyperonyme afin d'identifier un ED, comme présenté dans les résultats précédents. Dans un second temps, nous avons exploité, en plus des hyperonymes, les marqueurs de relation qui explicitent la définition, afin de repérer les ED. En effet, la première colonne des tables 6 et 7 présentent respectivement les termes simples et composés définis qui sont retenus après avoir utilisé les marqueurs d'hyperonymie. La troisième colonne de ces tables contient le nombre d'occurrences d'un terme avec son hyperonyme (colonne 2). Parmi ces occurrences, celle qui contient un

Terme	nbr. d'hyperonymes	nbr. d'ED
diabète	4	3
insuline	4	3
glucose	2	0

TABLE 5 – Résultats d'identification des exemples définitoires : hyperonymes extraits en utilisant les marqueurs : cas des termes simples

marqueur explicitant la définition, est considérée comme étant un ED. En ce qui concerne les termes dont les hyperonymes sont proposés par le MeSH, seulement 2 ED ont été identifiés dans le cas des termes *diabète de type 1* et *diabète de type 2*. Ces résultats ont été manuellement validés. La table 2 montre que, d'une part, s'il existe des ED dans le corpus, l'

Terme	hyperonyme	nbr. d'occ	nbr. d'ED
diabète	cause	48	2
diabète	facteur	116	4
diabète	facteur de risque vasculaire	5	2
diabète	maladie	100	1
insuline	facteur	26	1
insuline	hormone	21	2
insuline	moment	7	0
insuline	patient	52	0
glucose	facteur	11	0
glucose	phénomène	5	0

TABLE 6 – Résultats d'identification des ED : hyperonyme (extrait avec des marqueurs d'hyperonymie) + Marqueurs de relation explicitant la définition : cas des termes simples

Terme	hyperonyme	nbr. d'occ	nbr. d'ED
united kingdom prospective diabetes study	u.k.p.d.s	2	2
diabetes control and complications trial research group	d.c.c.t	5	2
diabète de type 2	argument	9	0
diabète de type 2	critère	6	0
diabète de type 2	diabète	69	2
diabète de type 2	maladie	5	3
diabète de type 1	diabète	45	2
diabète non insulino-indépendant	maladie métabolique	1	1
Autosurveillance glycémique	outil	1	1
diabète gestationnel	risque	10	2

TABLE 7 – Résultats d'identification des ED : hyperonyme (extrait avec des marqueurs d'hyperonymie) + Marqueurs de relation explicitant la définition : cas des termes composés

hyperonyme de niveau 1 s'avère plus performant que les autres hyperonymes proposés par le MeSH. D'autre part, les résultats montrent que les définitions contenant cet hyperonyme direct sont majoritaires par rapport aux définitions sans hyperonymes.
Par ailleurs les colonnes (avec. hyp) des tables 2 et 4 (ainsi que la table 5) montrent que les hyperonymes sont rares dans les contextes des termes en question. Ceci explique le nombre limité des termes présents dans les tables 3 et 5. Ainsi, nous déduisons que l'hypothèse 3 est restrictive pour identifier les exemples définitoires. Concernant l'hypothèse 2, la segmentation du corpus peut influencer les résultats. Par exemple, une phrase mal-segmentée ne peut pas être considérée comme un exemple définitoire. De plus, dans le cas d'une définition contenant deux phrases, où la première contient le terme visé et la deuxième contient son hyperonyme, l'hypothèse 2 présente également une contrainte supplémentaire empêchant de retenir une des deux phrases comme un exemple définitoire. Par exemple, aucune de ces deux phrases *ces schémas insuliniques sont également appelés basal-bolus. Bolus étant un mot latin signifiant action de jeter, coup de dé, coup de filet...* ne peut être considérée comme un exemple définitoire du terme *basal-bolus*.

5 Conclusion et perspectives

Dans ce travail nous nous sommes intéressés à l'identification des exemples définitoires dans un corpus comparable comme étant des CRC. Nous avons tout d'abord proposé trois hypothèses définissant cette notion. Nous avons considéré la relation d'hyperonymie comme un indicateur de définitions. Ensuite, nous avons proposé une méthode permettant de sélectionner ces exemples définitoires. Les résultats obtenus ont montré que les hypothèses proposées sont valides mais contraignantes étant donné que les hyperonymes sont peu fréquents dans les contextes des termes visés.
En ce qui concerne l'évaluation, les deux stratégies que nous avons adoptées pendant les premières expériences ne semblent pas être appropriées. D'une part, car la comparai- son des CRC candidats avec des définitions, qui sont souvent rares dans un petit corpus, ne permet pas d'avoir des résultats significatifs. Un CRC a été considéré comme valide, si et seulement s'il déclenchait une définition qui était préalablement annotée. Autrement dit, il s'agissait d'évaluer la richesse des CRC en termes de « définitoires », tandis qu'ils pouvaient apparaître dans des phrases autres que les définitions. D'autre part, parce que faire valider manuellement les CRC est un exercice sans doute très coûteux. Dans un premier temps, nous proposons d'exploiter séparément les hypothèses présentées, afin de rendre moins restrictives la définition de l'exemple définitoire. Dans un second temps, nous utiliserons d'autres relations telles que la causalité et la méronymie, comme indice de CRC.

Remerciement

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-12CORD-0020.

Références

ATKINS B. S. & RUNDELL M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

AUGER A. (1997). Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles. *Thèse de doctorat, Université de Neuchâtel*.

BARRIÈRE C. (2004). Knowledge-rich contexts discovery. *Proceedings of the Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)*.

F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

B. BIGI, Ed. (2014). *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, Marseille. ATALA, LPL.

BIRD S., DALE R., DORR B. J., GIBSON B. R., JOSEPH M., KAN M.-Y., LEE D., POWLEY B., RADEV D. R. & TAN Y. F. (2008). The acl anthology reference corpus : A reference dataset for bibliographic research in computational linguistics. In *LREC* : European Language Resources Association.

BOWKER L. & PEARSON J. (2002). *Working with specialized language : a practical guide to using corpora*. Routledge.

DAILLE B. & MORIN E. (2005). French-english terminology extraction from comparable corpora. In *Natural Language Processing–IJCNLP 2005*, p. 707–718. Springer.

DÉJEAN H. & GAUSSIER E. (2002). Une nouvelle approchea l'extraction de lexiques bilinguesa partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, p. 1–22.

DIDAKOWSKI J., LEMNITZER L. & GEYKEN A. (2012). Automatic example sentence extraction for a contemporary german dictionary. In *Proceedings EURALEX*, p. 343–349.

FAHMI I. & BOUMA G. (2006). Learning to identify definitions using syntactic features. In *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications*, p. 64–71.

FIRTH J. R. (1957). A synopsis of linguistic theory 1930-55. *The Philological Society*, **1952-59**, 1–32.

FUNG P. & MCKEOWN K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, p. 192–202.

GANGEMI A., NAVIGLI R. & VELARDI P. (2003). The ontowordnet project : Extension and axiomatization of conceptual relations in wordnet. In R. MEERSMAN, Z. TARI & D. C. SCHMIDT, Eds., *CoopIS/DOA/ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, p. 820–838 : Springer.

GOEURIOT L. (2009). *Découverte et caractérisation des corpus comparables spécialisés*. PhD thesis, Université de Nantes.

GREEN R., BEAN C. & MYAENG S. (2002). *The Semantics of Relationships : An Interdisciplinary Perspective*. Information science and knowledge management. Kluwer Academic Publishers.

KILGARRIFF A., HUSÁK M., MCADAM K., RUNDELL M. & RYCHLÝ P. (2008). Gdex : Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benarmara *et al.*, 2007), p. 101–110.

LAROCHE A. & LANGLAIS P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd international conference on computational linguistics*, p. 617–625 : Association for Computational Linguistics.

LEHMANN A. & MARTIN-BERTHET F. (1998). *Introduction à la lexicologie : sémantique et morphologie*. Collection Lettres supérieures. Ed. Dunod.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. *Actes de TALN*, p. 269–278.

MARTÍNEZ R., MARTÍNEZ G., DE LINGÜÍSTICA APLICADA U. P. F. I. U. & BACH C. (2009). *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. Sèrie tesis. Universitat Pompeu Fabra.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In *Recent Advances in Computational Terminology*, p. 279–302.

MURESAN S. & KLAVANS J. (2002). A method for automatically building and evaluating dictionary resources. In *LREC* : European Language Resources Association.

NAKAO Y. (2010). *Analyse contrastive français-japonais du discours en langue de spécialité-modalité et définition phrastique*. PhD thesis, Université de Nantes.

NAVIGLI R. & VELARDI P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1318–1327, Uppsala, Sweden : Association for Computational Linguistics.

PEARSON J. (1998). *Terms in context*, volume 1. John Benjamins Publishing.

RAPP R. (1999). Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 519–526 : Association for Computational Linguistics.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. PhD thesis, Toulouse 2.

REBEYROLLE J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, **25**, 153–174.

REIPLINGER M., SCHÄFER U. & WOLSKA M. (2012). Extracting glossary sentences from scholarly articles : A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, p. 55–65, Stroudsburg, PA, USA : Association for Computational Linguistics.

SAGGION H. (2004). Identifying definitions in text collections for question answering. In *LREC* : European Language Resources Association.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benarmara *et al.*, 2007), p. 401–410.

VELARDI P., FARALLI S. & NAVIGLI R. (2013). Ontolearn reloaded : A graph-based algorithm for taxonomy induction. *Computational Linguistics*, **39**(3), 665–707.

WESTERHOUT E. (2009). Extraction of definitions using grammar-enhanced machine learning. In A. LASCARIDES, C. GARDENT & J. NIVRE, Eds., *EACL (Student Research Workshop)*, p. 88–96 : The Association for Computer Linguistics.

Induction d'une grammaire de propriétés à granularité variable à partir du treebank arabe ATB

Raja Bensalem Bahloul[1, 2], Marwa Elkarwi[1, 2]
(1) Laboratoire Mir@cl, FSEGS, Université de Sfax, Sfax, Tunisie
(2) Laboratoire Parole et Langage (LPL), CNRS, Université d'Aix-Marseille, Aix-en-Provence, France
raja_ben_salem@yahoo.com, marwaelkarwi89@gmail.com

Résumé. Dans cet article, nous présentons une démarche pour l'induction d'une grammaire de propriétés (GP) arabe en utilisant le treebank ATB. Cette démarche se base sur deux principales étapes : (1) l'induction d'une grammaire hors contexte et (2) l'induction d'une GP par la génération automatique des relations qui peuvent exister entre les unités grammaticales décrites dans la CFG. Le produit obtenu constitue une ressource ouvrant de nouvelles perspectives pour la description et le traitement de la langue arabe.

Abstract. This paper presents an approach for building an Arabic property grammar using the treebank ATB. This approach consists in two main steps: (1) inducing a context-free grammar from a treebank and (2) inducing a property grammar. So, we acquire first a context-free grammar (CFG) from the source treebank and then, we induce the property grammar by generating automatically existing relations between grammatical units described in the CFG. The result is a new resource for Arabic, opening the way to new tools and descriptions.

Mots-clés : Treebanks, langue arabe, grammaire hors-contexte, grammaires de propriétés
Keywords: Treebanks, Arabic language, context-free grammar, property grammars

1 Introduction

Le formalisme de Grammaires de Propriétés (GP) est l'une des approches linguistiques qui mettent la notion de contraintes au cœur de l'analyse (Blache, 2001 ; Blache, 2005). Ce formalisme se distingue des autres approches basées sur les contraintes par sa représentation simple, directe, locale et décentralisée des informations linguistiques. En effet, à la différence des théories génératives, cette approche représente, d'une manière indépendante, tout type d'information, quelle que soit sa position, mais également des informations incomplètes, partielles et non canoniques, ce qui favorise sa flexibilité et sa robustesse. De plus, cette approche ne requiert pas la construction d'une structure locale des informations syntaxiques avant d'utiliser les contraintes qu'elle a décrites, comme le font les autres approches basées sur les contraintes (en HPSG (cf. (Pollard, 1994)) un arbre local et en CDG (cf. (Maruyama, 1990)) une relation de dépendance). Mais plutôt, elle spécifie les informations syntaxiques directement sur des catégories. Ainsi, une GP est formée par un ensemble de propriétés indépendantes l'une de l'autre. Ces propriétés expriment différentes relations (syntaxiques, sémantiques, etc.) entre les catégories qui forment la structure syntaxique. Elles peuvent être très spécifiques (concernant un ensemble limité de catégories) ou au contraire très générales.

Les qualités qui caractérisent le formalisme de GP nous ont incités à l'utiliser pour construire une nouvelle ressource robuste et riche pour la langue arabe. Le traitement de cette langue présente plusieurs défis. Ces défis ne sont pas seulement liés à certaines spécificités de l'arabe à étudier (comme l'absence des voyelles, la nature agglutinative des mots), mais aussi à des phénomènes linguistiques particuliers à traiter (comme les relatives, les anaphores et les coordinations). Une construction manuelle de cette nouvelle ressource en se basant sur un corpus regroupant toutes les règles de la grammaire arabe, est certainement difficile et couteuse. Ceci requiert en effet beaucoup de temps ainsi que la collaboration de plusieurs linguistes. Une autre manière de procéder peut être proposée : construire la GP à partir d'un corpus annoté. Les treebanks qui sont des corpus annotés manuellement formant une structure morphosyntaxique à plusieurs niveaux d'analyse (niveau mot, niveau syntagme et niveau phrase) peuvent être exploités dans ce cadre. Les treebanks arabes sont déjà rares, le premier lancé étant le treebank PATB (notamment appelé ATB) (Maamouri et Bies, 2004). Nous avons choisi de l'utiliser vu ses qualités qui s'avèrent convenables à la construction de notre GP. La première qualité caractérisant ce treebank est sa représentation à base de syntagmes, conforme à la structure syntaxique

hiérarchisée de la GP à construire. Ce choix est motivé également par la richesse, la fiabilité et la compatibilité des annotations de Part-of-Speech (POS) et de relations syntaxiques et sémantiques de l'ATB à des consensus. Ces annotations sont en fait élaborées et validées par des linguistes. De même, la grammaire de ce treebank est adaptée à l'arabe standard moderne. Il ne faut pas oublier aussi la pertinence, la variété et la grande taille qui caractérisent ses documents sources. Ces documents méritent d'être qualifiés « pertinents » grâce à leur conversion par plusieurs autres treebanks à leur représentation. Le fait de disposer d'une ressource de ce type permet de générer automatiquement et de façon très contrôlée de nouvelles ressources dans d'autres formalismes. Des ressources à large couverture sont ainsi obtenues, héritant des qualités du treebank d'origine tout en gagnant en temps de construction.

Toutefois, le fait que les catégories représentées dans le treebank se caractérisent par une forte granularité, peut affecter la taille des informations à représenter dans la GP. Il faut alors intégrer des mécanismes de contrôle la réduisant. Une autre difficulté peut être rencontrée au niveau de la génération des propriétés dans notre GP. En effet, il y a des propriétés faciles à déduire, mais il y en a d'autres nécessitant des heuristiques.

Dans cet article, le processus d'induction de notre GP se déroule sur deux phases : La première consiste à induire une grammaire hors-contexte (Context-Free Grammar, CFG) à partir de l'ATB. La seconde phase porte sur la déduction des différentes relations qui existent entre les catégories de chaque unité syntaxique à partir des règles de la CFG obtenue. La taille de la GP obtenue peut être contrôlée en variant les différents niveaux de granularité des catégories grammaticales de l'ATB. En plus, avec les types de propriétés définis dans le présent article, la démarche d'induction de GP que nous avons adoptée est purement automatique et indépendante de toute langue et du formalisme du treebank source. Ceci favorise sa réutilisation. Selon nos connaissances, la GP que nous avons obtenue, induite à partir d'un treebank, représente le premier essai produit pour l'arabe.

Cet article est organisé comme suit : la section 2 est consacrée à une brève présentation de l'état de l'art. Ensuite, l'étude de l'ATB est l'objet de la section 3. La section 4 décrit ensuite la démarche d'induction que nous proposons. La section 5 présente les expérimentations et les résultats obtenus de l'application de notre démarche. La section 6 termine par une conclusion et des perspectives.

2 Etat de l'art

Pour pouvoir aborder la problématique de notre travail, nous avons mené des recherches sur deux volets différents : un aperçu sur les approches d'induction de GP et une observation des différentes améliorations effectuées sur l'ATB.

D'une part, ce qu'il y a en commun dans les approches d'induction de GP est leur entrée qui est la CFG. Leurs formalismes sources ou leurs usages quant à eux diffèrent d'une approche à une autre. Concernant l'entrée de ces approches, elle est sous forme de suites d'étiquettes décrivant les unités syntaxiques observées dans un corpus annoté (étiqueté). Ces suites sont en fait représentées par une CFG. Il est vrai que l'entrée, étant la CFG, est commune à toutes les approches d'induction de GP, mais son induction elle-même à partir d'un treebank peut être faite selon des techniques différentes. En effet, elle peut être une CFG simple comme dans (Marcus et al. 1993 ; Hajic, 1998 ; Abeillé et al., 2003 ;Telljohann et al., 2004), ou bien une CFG intégrant des ajustements spécifiques aux suites d'étiquettes. C'est le cas notamment des CFG probabilistes affectant des probabilités à chacune des suites d'étiquettes obtenues, comme dans (Charniak, 1996 ; Mohri et Roark, 2006 ; Rebein et VanGenabith, 2007 ; Tounsi et VanGenabith, 2010). Il existe également plusieurs exemples montrant la différence entre les approches d'induction des GP au niveau de leurs formalismes sources et de leurs usages : Dans (Blache et al., 2003) par exemple, les auteurs ont préparé leur propre corpus étiqueté à partir d'un corpus français brut en passant successivement par une étape de segmentation et une étape d'étiquetage. Chacune de ces deux étapes ont recours à un dictionnaire constitué plus particulièrement d'un lexique appelé DicoLPL et composé d'environ 450 000 formes[1]. La ressource que représente DicoLPL est constituée sur la base d'un lexique interne au LPL[2] et complétée en s'appuyant sur des ressources existantes et des ressources acquises manuellement ou automatiquement par vérification sur corpus. La GP obtenue a été utilisée ensuite dans le cadre d'analyseurs syntaxiques à granularité variable (VanRullen et al., 2005). La base de données Aix-MARSEC (Auran et al., 2004) a formé aussi un formalisme source pour l'induction des GP. En effet, Aix-MARSEC est formée de deux principaux composants : les enregistrements numérisés du corpus MARSEC[3] et leurs annotations. Ces annotations ont

[1] Une version évoluée du lexique DicoLPL, avec plus de formes, a été présenté dans (VanRullen et al., 2005).

[2] www.lpl.univ-aix.fr/

[3] MARSEC (Machine Readable Spoken English Corpus) contient des enregistrements acoustiques numérisés et c'est une extension du corpus SEC disponible en version treebank et en version étiquetée (www.comp.leeds.ac.uk/ccalas/tagsets/sec. html).

été présentées au début à neuf niveaux différents (tels que le niveau phonèmes, syllabes, mots, etc). A ces niveaux, deux niveaux supplémentaires ont été spécifiés : l'annotation syntaxique ainsi qu'un système de GP relatif. De plus, les treebanks représentent également un autre formalisme source pour l'induction des GP. Ainsi, dans (Blache et Rauzy, 2012) par exemple, les auteurs ont bénéficié de ces qualités en utilisant un sous-ensemble du treebank français FTB[4] pour induire leur GP. Ils ont effectué des modifications sur ce sous-ensemble pour assurer une meilleure homogénéité avec les ressources existantes dans d'autres langues ou pour d'autres domaines. Ces modifications sur les niveaux morphologique et syntaxique et sur les positions des marqueurs de ponctuation. La grammaire obtenue a été exploitée ensuite pour enrichir automatiquement le treebank source par une représentation à base de contraintes (Blache et Rauzy, 2012) tout en appliquant un ensemble de solveurs de contraintes. Les travaux d'induction de GP ne se limitent pas uniquement au treebank français (Blache et Rauzy, 2012), mais aussi au treebank chinois (CTB) (Blache, 2014).

D'autre part, l'ATB a été également enrichi en lui intégrant différentes améliorations et corrections pour pouvoir surmonter les défis liés à certaines spécificités de la langue arabe. En effet, cette langue est caractérisée par sa morphologie complexe. Ce problème a été examiné au niveau de l'ATB par (Kulick et al., 2010) en représentant les mots ayant une forme agglutinative par des unités séparées dans une structure arborescente. Par exemple, le mot arabe « كتبه » (ktbh/ *ses livres*), s'il n'est pas voyellé, l'ATB le représente en deux parties tels que « ktb » (*livres*) est un groupe nominal et « h » (*ses*) est un pronom possessif. Pour réaliser cette tâche, les auteurs ont utilisé l'outil SAMA (Standard Morphological Analyzer) pour générer des solutions d'analyse morphologique pour chaque mot de l'ATB. De même, l'absence de voyelles dans la langue arabe peut générer des ambigüités. Pour surmonter cette difficulté, les auteurs de (Kulick et al., 2010) ont intégré une représentation syntaxique abstraite de la structure arborescente tout en autorisant le passage entre les différents niveaux de représentation syntaxique et tout en fournissant différents niveaux de voyellation pour chaque mot de l'ATB. Concrètement, la procédure d'annotation morphosyntaxique que les auteurs ont suivie est basée sur deux grandes étapes : la première consiste en une annotation syntaxique décomposant le texte de l'ATB en mots (appelés jetons sources). Ces mots sont intégrés dans l'outil SAMA, puis générés sous forme voyellée. La deuxième étape, quant à elle consiste à séparer ces jetons des pronoms liés durant l'annotation syntaxique. Par exemple, l'analyse du mot arabe « كتبه » par l'outil SAMA génère une solution qui le décompose en trois segments. Cette solution inclut une séquence d'informations pour chaque segment portant sur trois champs : la forme voyellée, l'étiquette Part-Of-Speech (POS) et la traduction du segment. La solution SAMA de ce mot est présentée comme suit :

[kutub, NOUN, books]	[i, CASE_DEF_GEN, def.gen]	[hi, POSS_PRON_3MS, its/his]

Les auteurs de (Kulick et al., 2010) ont cherché également des procédures spécifiques pour traiter les mots arabes ayant un caractère particulier, tels que les mots ayant une forme agglutinative ne pouvant pas être explicitement décomposée. Le mot عما (EmA/*de ce que*) par exemple est une préposition suivie d'un pronom relatif. La solution proposée par SAMA est composée de deux segments. Elle inclut un « n » dans « Ean », n'ayant pas été présent dans le mot source « EmA ».

[Ean, PREP, from/about/of]	[mA, REL_PRON, what]

Par ailleurs, l'application d'une analyse statistique apprise sur le treebank arabe (ATB) et examinant des incohérences dans les annotations a généré des scores d'analyse inférieurs aux prévisions. Ces incohérences résident au niveau de certaines constructions syntaxiques ou de la relation entre les étiquettes POS et les annotations syntaxiques. La résolution de ces incohérences va corriger largement les directives d'annotation. Le travail (Maamouri et al., 2008) s'inscrit dans ce cadre. En effet, il présente des corrections et des améliorations des incohérences d'annotation dans le but d'améliorer la qualité d'analyse du corpus de l'ATB. Ce travail utilise les étiquettes POS pour corriger et améliorer les directives d'annotation syntaxique. Ces corrections sont proposées aussi bien au niveau morphologique qu'au niveau syntaxique. Au niveau morphologique, les auteurs de (Maamouri et al., 2008) ont proposé de raffiner les étiquettes POS des noms et des adjectifs pour spécifier les noms quantifieurs (NOUN_QUANT), les nombres (NOUN_NUM), les adjectifs comparatifs (ADJ_COMP), les nombres ordinaux (ADJ_NUM), etc. De même, ils ont distingué des catégories d'étiquettes POS pour les différentes particules, telles que l'étiquette CONJ qui a été décomposée en quatre catégories. Et pour distinguer les pseudo-verbes des verbes, ils ont ajouté l'étiquette PSEUDOVERB pour les sœurs de la particule arabe « إنّ » (inna / *que*). Au niveau syntaxique, les auteurs de (Maamouri et al., 2008) se sont focalisés sur la désignation du nom par une étiquette spécifique s'il est un quantifieur dans le syntagme « idafa » pour spécifier correctement la tête sémantique du syntagme. En effet, pour le syntagme « كل مجموعة » (*chaque collection*), la tête sémantique n'est pas le segment « chaque » mais plutôt le segment « collection » parce que celle-là est un nom quantifieur et non pas un simple nom. Il faut alors le spécifier par l'étiquette NOUN_QUANT. Les auteurs de (Maamouri et al., 2008) ont marqué aussi les gérondifs et les participes, si ceux-ci présentent une lecture verbale, par des étiquettes regroupant toute l'unité syntaxique. Par exemple, la phrase « احتفل الفريق بفوزه بكأس الأبطال» (*L'équipe a célébré son gain de la coupe des champions*), ne contient pas un gérondif suivi d'un complément (« فوزه بكأس الأبطال ») ce qui le caractérise par une lecture verbale plutôt qu'un simple gérondif. Cette lecture verbale est entièrement analysée.

[4] Le FTB est constitué de 12,891 phrases annotées contenant plus que 383,000 mots (Abeillé et al., 2000).

L'ATB dans sa nouvelle forme, après les améliorations et les corrections effectuées permettant d'aligner étroitement son annotation aux catégories de la grammaire arabe traditionnelle, représente une source assez robuste offrant plusieurs spécificités servant à la réalisation d'une induction de GP réussie. Ces spécificités sont citées dans la section suivante.

3 Etude de l'ATB

Avant d'expliquer les spécificités liées à l'ATB, une brève présentation de cette ressource linguistique s'avère nécessaire. En effet, l'ATB a été construit dans le cadre d'un projet en 2001 au LDC[5] (Maamouri et Bies 2004). Il représente un corpus composé de 23,611 phrases extraites d'articles de presse annotées manuellement. Ce corpus a été divisé en ensembles de textes (divisions) pour répondre aux besoins de recherche variés dans le domaine de TALN comme l'apprentissage et l'évaluation (Diab et al., 2013).

Doté d'une annotation très riche, l'ATB se caractérise par un ensemble de particularités et de qualités servant à une meilleure induction de GP. En effet, ses annotations présentent l'avantage d'être fiables. Ceci est prouvé par son efficacité dans un grand nombre de travaux dans différents domaines de TAL (Habash, 2010). Son texte source a prouvé également sa pertinence par son exploitation pour la création d'autres treebanks arabes comme le PADT (Hajic et al., 2001) et le CATiB (Habash et Roth, 2009) qui ont converti l'ATB vers leurs représentations syntaxiques en plus d'autres textes qu'ils ont annotés. De plus, l'ATB est disponible en cinq formats différents (voir la sous-section 3.1). Une autre particularité qui peut être remarquée est celle de la granularité forte qui caractérise son annotation (voir la sous-section 3.2). Finalement, l'ATB a prouvé son aptitude à représenter correctement certains phénomènes particuliers de la langue arabe (voir la sous-section 3.3).

3.1 Les formats de représentation des données dans l'ATB

L'ATB est fourni sous différents formats pour étendre son utilisation pour différents besoins de recherche. Ces formats que nous citons sont au nombre de cinq. Le format *« sgm »* représente les documents sources. Par contre, le format *« pos »* affiche des informations (comme la translitération, la voyéllation et la traduction) décrivant chaque mot source sous forme de champs avant et après la séparation des agglutinations. Le format *« xml »* quant à lui affiche les annotations de l'arbre de mots sources après la séparation des agglutinations. Le format *« penntree »* représente le corpus en deux versions (voyellée ou non) sous forme d'arborescence affichant chaque mot dans sa structure hiérarchique et devant son étiquette POS. Finalement, le format *« integrated »* affiche des informations aussi bien sur la structure arborescente que sur chaque mot source avant et après la séparation des agglutinations.

Après la présentation de ces différents formats, il faut prendre une décision concernant le choix du format de l'ATB à utiliser pour induire la CFG qui sera l'entrée de l'étape d'induction de la GP. Ce choix dépend de trois critères à prendre en compte à savoir : la simplicité de représentation, la présence d'une structure arborescente et l'annotation du niveau syntaxique des documents sources. Nous avons défini ces critères en nous appuyant sur la forme de la CFG à induire. Cette grammaire se limite au niveau des étiquettes POS et non pas au niveau des mots sources. Le format « penntree » a été le seul sélectionné pour sa satisfaction à tous les critères de choix indiqués. Plus particulièrement, nous avons utilisé la version voyellée du format « penntree » pour éviter les ambigüités liées à l'absence de voyelles en arabe.

3.2 Les niveaux de granularité des catégories

L'annotation dans l'ATB est caractérisée par une granularité forte. En effet, cette annotation inclut plus que 400 étiquettes POS différentes offrant des informations morphosyntaxiques comme la déclinaison, le mode, le genre, la définition, etc (Maamouri et al., 2009). Parmi ces étiquettes, 22 sont syntagmatiques (des catégories syntaxiques), 20 sont des relations syntaxiques et sémantiques et 24 représentent les étiquettes POS de base. L'ATB prend en compte également des pronoms vides qui peuvent apparaitre dans les phrases arabes tout en leur affectant une étiquette spécifique. Cette annotation a été toujours améliorée pour résoudre les incohérences morphosyntaxiques liées à certaines spécificités de la langue arabe (Maamouri et al., 2008 ; Kulick et al., 2010). La figure 1 illustre les différents traits caractérisant la plupart des catégories lexicales décrites dans l'ATB (Maamouri et al., 2009).

Nature	**Nom**	**Verbe**	**Verbe au présent**	**Pronom**	**Pronom Relatif**

[5] LDC (Linguistic Data Consotium) : Consortium de données linguistiques https://www.ldc.upenn.edu/

Traits		ou Adjectif				
Type		Nom : NUM, PROP, QUANT, VN Adjectif : COMP, NUM, VN	I, C, P	---	POSS, REL, DEM	---
Fonction		---	SUBJ	---	---	---
Déclinaison (mode)		NOM, ACC, GEN	---	I, JUS, SJ	---	NOM, ACC, GEN
Définition		DEF, INDEF	---	---	---	DEF, INDEF
Déterminant		DET,	---	---	---	---
Forme		---	PASS,	---	---	---
Accord	Genre	MASC, FEM	M, F	---	M, F	---
	Nombre	SG DU PL	S, D, P	---	S, D, P	
	Personne	---	1, 2, 3	---	1, 2, 3	
Exemples		NOUN_NUM+NSUFF_FEM_SG+CASE_INDEF_NOM	CV+CVSUFF_SUBJ:2MP	IV1P+IV_PASS+IVSUFF_MOOD:I	POSS_PRON_2MP	REL_PRON+CASE_DEF_GEN
		DET+ADJ_COMP+NSUFF_MASC_DU_ACC	PV_PASS+PVSUFF_SUBJ:3MD	IV3MP+IV+IVSUFF_SUBJ:MP_MOOD:SJ	DEM_PRON_MD	REL_PRON+CASE_INDEF_ACC

FIGURE 1 : Les traits caractérisant des catégories lexicales de l'ATB

Maintenant si nous diminuons cette granularité, plusieurs sous-ensembles de catégories seront factorisés en une seule catégorie. Prenons comme exemple, le sous-ensemble {NOUN_PROP, NOUN_PROP+CASE_DEF_ACC, NOUN_PROP+CASE_DEF_NOM} qui marque les noms propres par trois étiquettes dans le cas d'une forte granularité. Si nous avons une faible granularité, ces étiquettes sont généralisées, et factorisées en une seule étiquette sous le nom NOUN_PROP. Ceci s'explique par le fait que le manque de précision dans les catégories grammaticales dû à la diminution du niveau de granularité permet de les factoriser. Cette factorisation influence le nombre de règles de la CFG, et par conséquent sa taille. En effet, plus ce niveau est bas, plus le nombre de catégories grammaticales est réduit et plus la CFG est compacte mais abstraite et générale. Inversement, plus ce niveau est élevé, plus le nombre de catégories est grand et plus la CFG est détaillée mais précise et significative. La dégradation de la significativité dans la CFG est due à la perte d'informations au niveau des règles lorsqu'on réduit la granularité de ses catégories. Il faut alors contrôler le niveau de granularité des catégories pour pouvoir faire un compromis entre la taille et la qualité de la CFG.

3.3 Représentation de certains phénomènes particuliers de la langue arabe

Comme nous l'avons déjà mentionné dans l'introduction de cet article, la langue arabe présente plusieurs défis lors de son traitement automatique. Parmi ces défis, nous citons les phénomènes linguistiques particuliers comme les relatives, les coordinations et les anaphores. L'ATB a aussi contribué à traiter ces phénomènes en les représentant conformément à la grammaire arabe (Maamouri et al., 2009), ce qui favorise la robustesse des ressources qui peuvent l'exploiter.

Dans le cas des propositions relatives, il est bien remarquable, comme le montre la figure 2, que la relative SBAR « التي لم تحترق» (Al~atiy lam taHotariq+o/ *qui ne sont pas brulées*) est réellement jointe au syntagme nominal « المواد الهيدروكربونية» (Al+mawAd~+i Al+hiydoruwkarobuwniy~ap+i/ *les matières hydro-carboniques*) qui la modifie.

```
(NP (NP -Al+mawAd~+i::-□ المَوادِّ ::the+substances/materials+[def.gen.]
        Alhydrwkrbwnyp:: الهيدروكربونية ::nogloss )
    (SBAR (WHNP-2 Al~atiy:: الَّتي ::which/who/whom_[fem.sg.] )
          (S (VP (PRT lam:: لَم ::did_not )
                 ta+Hotariq+o:: تَحْتَرِقْ ::it/they/she+burn_up/be_burned+[jus.]
                 (NP-SBJ-2 *T*)))))
```

FIGURE 2 : Un exemple de proposition relative représenté par l'ATB (Maamouri et al., 2009)

Pour les coordinations, elles sont composées en arabe généralement de deux conjoints ainsi que la conjonction qui les réunit. L'ATB spécifie plusieurs formes de coordinations selon la manière de représentation des trois composants de la coordination. Concernant les anaphores, l'ATB indique uniquement ceux des catégories vides et des cas exceptionnels comme les structures écartant les syntagmes verbaux (Maamouri et al., 2009).
La richesse et la fiabilité des annotations de l'ATB, ainsi que la pertinence de ses documents sources nous ont incités à l'utiliser pour générer automatiquement une GP héritant les qualités de ce treebank. Ceci permet de favoriser sa robustesse tout en gagnant en temps de construction. Nous proposons alors une démarche à appliquer pour exploiter cette ressource et obtenir notre GP.

4 Le formalisme de GP

Le formalisme de GP représente une approche s'appuyant sur les contraintes (Blache, 2001) tout en permettant un accès direct aux valeurs des variables. En effet, il représente les informations syntaxiques directement en fonction de catégories, et non pas en fonction de structures comme le font les autres approches basées sur la satisfaction de contraintes (Pollard, 1994 ; Maruyama, 1990). Le formalisme de GP s'inscrit dans le cadre des grammaires syntagmatiques tout en adoptant une structure syntaxique hiérarchisée. Ainsi, une GP est formée par un ensemble de propriétés exprimant différentes relations entre les catégories qui forment la structure syntaxique. Dans ce qui suit, nous présentons ses éléments essentiels ainsi que son fonctionnement.

4.1 Les catégories

Une catégorie en GP est formée par une structure de traits définissant les informations pouvant intervenir dans la spécification de contraintes. Chaque trait est un couple <étiquette, valeur>. Les traits partageant des caractéristiques communes sont regroupés dans un type spécifique. A ce type, plusieurs sous-types peuvent être associés héritant ses traits, ainsi que des traits spécifiques. Les types et leurs sous-types peuvent être organisés sous la forme de hiérarchies, distinguant ainsi différents niveaux de spécification de l'information, de telle sorte que chaque catégorie ayant un ensemble de traits est associée à un certain niveau de spécification (granularité) de son type. Une hiérarchie est représentée sous la forme d'un arbre où la racine représente un type et les nœuds descendants sont des sous-types plus spécifiques de leurs nœud parent. La figure 3 ci-dessous illustre la hiérarchie du type « *cat_lexicale* » caractérisant les catégories lexicales de l'ATB. Pour ce type, un seul trait appelé CATLEX est spécifié ayant une valeur catlex complexe dont le premier trait est appelé NATURE. Ces traits sont représentés dans une matrice sous le type concerné, leurs étiquettes sont en majuscules et leur type est en italique.

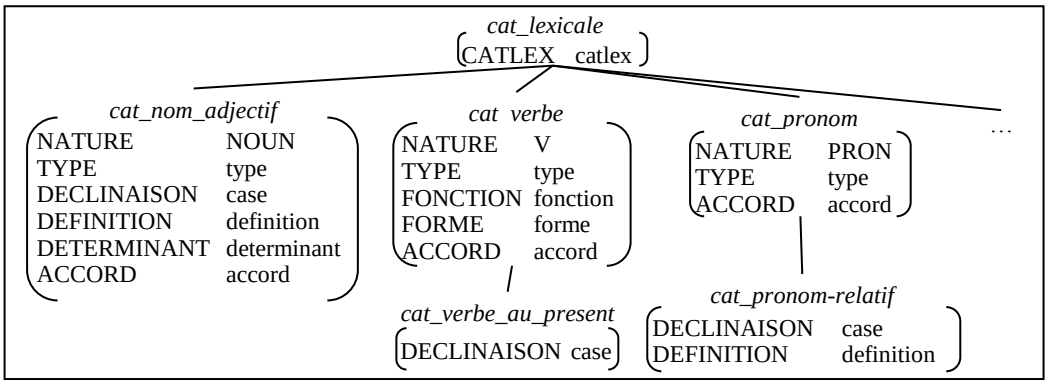


FIGURE 3 : Hiérarchie du type « *cat_lexicale* » caractérisant les catégories lexicales dans l'ATB

4.2 Les Propriétés

Une propriété est une contrainte portant sur un ensemble de catégories et décrivant une certaine catégorie. Ce qui caractérise les propriétés est le fait qu'elles sont toutes définies au même niveau, c'est à dire qu'elles ne sont ni dépendantes les unes des autres ni ordonnées entre elles. En plus, elles représentent des relations traitant toutes les informations de manière explicite à la différence des représentations de constituants qui se limitent à la définition explicite d'une relation unique : la hiérarchie. Les relations hiérarchiques représentent l'information syntaxique de manière holistique. Cette représentation ignore les cas où cette information est incomplète ou mal formée. Ce type de relations ne traite pas les phénomènes linguistiques complexes comme les relatives, les anaphores et les coordinations. Les propriétés par contre peuvent décrire ces phénomènes grâce à leur représentation décentralisée et locale des informations linguistiques. Ces propriétés peuvent être du niveau lexical (comme les propriétés morphologiques ou phonologiques) ou bien du niveau syntaxique. Les propriétés syntaxiques portent sur six différents types de contraintes montrées dans la figure 4 suivante :

Propriétés	Symboles	Fonctions
Linéarité (Lin)	≺	Relations de précédence linéaire entre les constituants d'un constituant d'un niveau syntaxique
Unicité (Unic)	Unic	Ensemble des constituants ne devant apparaître qu'une seule fois

Obligation (Oblig)	Oblig	Ensemble des têtes possibles du constituant de niveau syntaxique
Exigence (Exig)	⇒	Cooccurrence obligatoire entre les constituants
Exclusion (Excl)	⊗	Restriction de cooccurrence entre les constituants
Dépendance (Dep)	∾	Relations de dépendance entre les constituants

FIGURE 4 : Fonctions des propriétés dans les GP

Les propriétés d'unicité et d'obligation sont des relations unaires. Les autres sont par contre des relations binaires.

4.3 Vérification de la satisfaction de contraintes

Le formalisme de GP représente les contraintes de manière indépendante. Ces contraintes sont regroupées dans des sous-systèmes caractérisant chacun une catégorie syntaxique. L'analyse avec ce formalisme revient en fait à vérifier pour chaque catégorie syntaxique la satisfaisabilité de son sous-système de contraintes. Pour analyser un énoncé donné, un processus de trois étapes est à appliquer : L'énumération de l'ensemble des catégories possibles de cet énoncé y compris celles syntaxiques susceptibles d'être des catégories mères pour les catégories dégagées, la construction des suites possibles des catégories énumérées, et finalement le calcul de la caractérisation des suites par la vérification de la consistance des sous-systèmes de contraintes correspondant aux catégories syntaxiques de ces suites. L'apport du formalisme de GP est très bien décrit grâce à cette notion de caractérisation. En effet, aucune règle syntagmatique ni schéma de règle n'est nécessaire pour décrire syntaxiquement un énoncé. Il suffit de fournir un ensemble de systèmes de contraintes décrivant cet énoncé de façon simple et directe et peu importe sa forme, et de vérifier leur satisfaction.

5 Démarche proposée

Pour élaborer notre démarche, nous nous sommes basées sur l'idée d'induction de GP à partir du FTB adoptée dans (Blache et Rauzy, 2012). Cette démarche s'articule autour de deux grandes étapes : l'induction de la CFG et l'induction de la GP. La démarche proposée est présentée dans la figure 5 suivante :

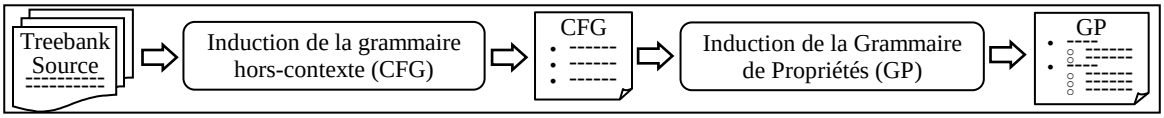


FIGURE 5 : Démarche d'induction de la GP

5.1 Etape d'induction de la CFG

L'induction de notre GP ne peut pas être réalisée directement en appliquant une simple tâche d'acquisition et de manipulation des données du treebank. Il faut en effet introduire une étape intermédiaire permettant de représenter les productions décrivant les unités syntaxiques. Cette étape consiste à parcourir l'ATB et à en extraire les constructions (règles) possibles pour produire une CFG pour l'arabe à différents niveaux de granularité. Ceci est réalisé dans le cadre de trois sous-étapes primordiales : la détection des constituants, l'élaboration des règles et le contrôle de ses niveaux de granularité (voir figure 6).

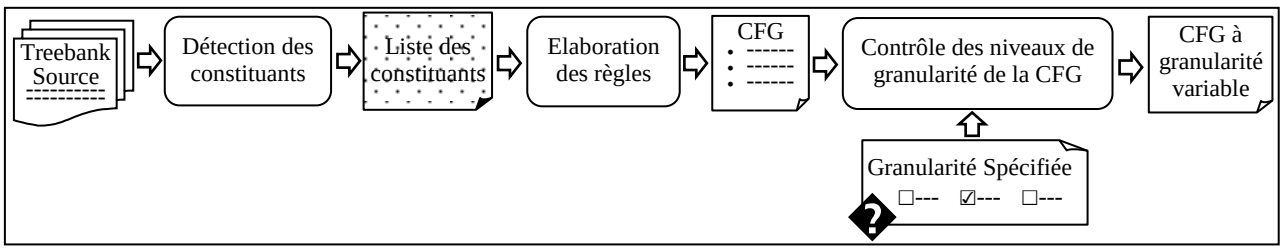


FIGURE 6 : Sous-étapes d'induction de la CFG

5.1.1 Détection des constituants

Cette sous-étape permet tout d'abord de parcourir le treebank ligne par ligne en décomposant chaque ligne en un ensemble de mots. Les mots commençant par une parenthèse ouvrante représentent les constituants. Après l'acquisition de l'ensemble des constituants, elle élimine les doublons puis trie ces constituants par ordre alphabétique. Finalement, afin de présenter quelques statistiques, cette sous-étape calcule le nombre d'occurrences de chacun de ces constituants.

5.1.2 Elaboration des règles

Il s'agit d'extraire les différentes constructions possibles (règles) pouvant être représentées dans l'ATB, ce qui permet de produire implicitement la CFG. Ceci se fait en parcourant ligne par ligne les fichiers des données source du treebank. Pour chaque ligne, trois tâches sont réalisées successivement à savoir : la récupération de l'ensemble des mots formant cette ligne, la détection des extrémités des règles en utilisant les mots récupérés et la génération des règles. La deuxième tâche permet d'affecter des positions incrémentales aux mots récupérés tout en donnant aux bornes (les parenthèses ouvrante et fermante) de chaque règle détectée la même position. A la troisième tâche, il s'agit de former les règles à générer. Chaque règle est composée d'une partie gauche consacrée à l'un des constituants du niveau syntaxique, et d'une partie droite représentant les constituants descendants de ce constituant syntaxique. Cette tâche s'étend également à l'exploration des constituants descendants dans la partie droite à la recherche d'autres règles de niveaux hiérarchiques plus fins. Après chaque parcours d'un fichier courant, les règles extraites sont accumulées à la liste des règles extraites des fichiers déjà parcourus. Cette liste est filtrée et triée. La liste finale des règles forme notre CFG. Il est possible également de calculer le nombre d'occurrences de chaque règle dans la liste pour présenter quelques statistiques.

En se trouvant devant une complexité prévue de la forte granularité caractérisant l'annotation de l'ATB, les catégories ne doivent pas être détectées d'emblée, mais le contrôle de leurs niveaux de granularité s'impose. L'emploi de ce contrôle est également motivé par le fait que nous voulons générer une grammaire à large couverture. Le fait d'offrir une certaine souplesse de présentation à cette grande masse de données favorise la qualité de la grammaire obtenue. Le contrôle à appliquer s'effectue au niveau de l'étape d'induction de la CFG vu qu'il concerne les catégories détectées. La section suivante explique les mécanismes de ce contrôle.

5.1.3 Contrôle des niveaux de granularité de la CFG

Pour pouvoir contrôler la granularité des catégories détectées, il faut spécifier leurs traits. Ces traits peuvent être organisés selon le formalisme de hiérarchies des types (voir la section 4.1). Mais, il faut les extraire tout d'abord de la structure de leur catégorie concernée. Pour cela, nous avons effectué une étude de cette structuration, et constaté que l'ATB a utilisé huit informations dans son annotation, portant sur : la numérotation, la déclinaison, le genre et le nombre, le mode de conjugaison du verbe, son type, le déterminant, la catégorisation lexicale et la celle syntaxique. Chacune de ces informations peut être spécifiée en se basant sur des caractères de séparation comme « : », « - », « + », « _ » ainsi que des mots clés comme « NSUFF », « CVSUFF », « IVSUFF », « PVSUFF », « CASE » et « MOOD ».

5.2 Etape d'induction de la GP

La CFG étant l'entrée de cette étape et la sortie de celle précédente, est utilisée pour induire automatiquement une GP. Ce qui facilite cette étape est le fait qu'aussi bien la CFG que la GP sont structurées sous forme de constituants de niveau syntaxique auxquels nous affectons des informations de différents types. La différence réside au niveau du type de ces informations. En effet, la CFG donne pour chaque constituant syntaxique, l'ensemble de règles qui le décrit. Ces règles forment en fait des contraintes hiérarchiques. La GP, par contre, représente le constituant syntaxique à l'aide de contraintes non hiérarchiques qui sont les « propriétés ». Pour expliquer comment nous avons induit notre GP, nous nous mettons d'accord sur quelques notations :

- XP représente tout constituant syntaxique décrit dans la grammaire.
- RHS(XP) est l'ensemble des règles spécifiant chaque XP.
- const(XP) est l'ensemble sans doublons des constituants pouvant former XP. Il est obtenu en parcourant RHS(XP) pour récupérer tout constituant faisant partie d'une règle de RHS(XP), il va servir à la représentation des propriétés. Nous avons commencé par la description des cinq premiers types de propriétés. Tandis que la description des propriétés de dépendance, elle n'est pas intégrée dans le présent article. Nous présentons, dans ce qui suit, les descriptions formelles de ces types de propriétés tout en nous appuyant sur celles établies dans (Blache, 2012) :

– La linéarité (lin) : elle vérifie dans tout l'ensemble RHS(XP) la validité de chaque relation de précédence entre chaque constituant de const(XP) et un autre. Pour cela, pour chaque couple de constituants dans const(XP), nous allons considérer que cette relation est vraie tant qu'il n'existe pas un contre exemple.

$\forall (c_i, c_j) \in const(XP) \mid c_i \neq c_j$
  $\forall rhs_a \in RHS(XP)$
    Si $((\exists (c_m, c_n) \in rhs_a \mid c_m = c_i \wedge c_n = c_j)$
    Et $(\nexists (c_m, c_n) \in RHS \mid c_m = c_i \wedge c_n = c_j \wedge c_n \prec c_m))$
    alors ajouter $lin(c_i, c_j)$

– L'unicité (unic) : elle vérifie dans tout l'ensemble RHS(XP) pour chaque constituant de const(XP) s'il n'est pas répété dans la même construction de RHS(XP). Selon cette interprétation, nous supposons que le constituant à traiter est unique tant qu'il n'existe pas un cas contraire.

$\forall c_i \in const(XP)$
  card : 0
  $\forall rhs_a \in RHS(XP)$
    $\forall c_j \in rhs_a$
      Si $(c_j = c_i)$ **alors** card ← card+1
  Si (card =1) **alors** ajouter $unic(c_i)$

– L'obligation (oblig) : elle représente l'ensemble des constituants obligatoires pour former XP. Un constituant obligatoire (tête) est un constituant devant apparaitre au moins une fois dans chacune des constructions de RHS(XP).

$\forall c_i \in const(XP)$
Si $(\forall rhs_b \in RHS(XP) \mid \exists c_j \in rhs_b \wedge c_j = c_i)$ **alors** ajouter $oblig(c_i)$

– L'exigence (Exig) : elle vérifie dans tout l'ensemble RHS(XP) la validité de chaque relation de cooccurrence entre chaque couple de constituants de const(XP). Une catégorie est co-occurrente avec une autre si l'apparition de la première implique l'apparition de la deuxième dans la même construction. Cette relation n'est pas symétrique du fait que si la deuxième catégorie apparait dans une construction sans la première, cette relation est considérée valide.

$\forall (c_i, c_j) \in const(XP) \mid c_i \neq c_j$
  bool ←vrai
  $\forall rhs_a \in RHS(XP)$
    bool ← $(\exists c_n \in rhs_a \mid c_n = c_i) \wedge (\nexists c_m \in rhs_a \mid c_m = c_j)$
Si bool **alors** ajouter $exig(c_i, c_j)$

– L'exclusion (excl) : elle vérifie dans tout l'ensemble RHS(XP) la validité de chaque relation de restriction de cooccurrence entre chaque couple de constituants de const(XP). Une catégorie n'est pas co-occurrente avec une autre s'il s'est arrivé que l'une apparait avec l'autre dans une même construction. Cette relation est totalement contraire à la relation d'exigence. Et même, elle est symétrique du fait que pour que cette relation soit valide, l'apparition de l'une de ses deux catégories empêche l'apparition de l'autre.

$\forall (c_i, c_j) \in const(XP) \mid c_i \neq c_j$
  bool ← faux
  $\forall rhs_a \in RHS(XP)$
    bool ← $(\exists (c_m, c_n) \in rhs_a \mid c_m = c_i \wedge c_n = c_j)$
Si non bool **alors** ajouter $excl(c_i, c_j)$

6 Expérimentations et résultats

Nous avons utilisé comme ressources pour induire notre GP, la deuxième division avec sa version 3.1 (ATB2 v3.1), composée de 501 articles de presse contenant 144.199 segments avant la fragmentation des clitiques. Comme nous l'avons déjà mentionné, le mécanisme d'induction de la GP se déroule sur deux tâches successives qui permettent d'obtenir successivement deux types de grammaires sous le format XML : La CFG et la GP. La taille de ces grammaires dépend du niveau de granularité des catégories qu'elles décrivent, puisque toute catégorie peut être caractérisée par différents traits morphologiques. Plus le niveau de granularité des catégories est élevé, plus ces grammaires sont complexes mais leurs propriétés de plus en plus fidèles à la langue, et inversement. La CFG obtenue est composée d'ensembles des règles décrivant chaque catégorie non terminale XP. Chaque règle prend la forme d'une liste ordonnée de catégories grammaticales représentant une catégorie syntaxique XP. La table 1 montre des informations sur la CFG

obtenue au niveau de granularité le plus élevé. Cette table affiche en particulier la fréquence de la règle « PREP NP » lorsqu'elle décrit la catégorie PP (Prepositional Phrase) y compris ses sous-catégories (e.g. PP-MNR et PP-TMP), qui intègrent plus de détail (Maamouri et al., 2009). A ce niveau, il y a 263 règles de différents types. Selon ce que nous avons observé dans la table1, nous pouvons noter que la granularité la plus élevée ne fait pas une grande différence pour certaines sous-catégories de PP. Par exemple, dans la plupart des cas, la construction « PREP NP » reste la plus fréquente quel que soit la sous-catégorie de PP qu'elle décrit. Les autres règles ne sont pas fréquentes, elles partagent ensembles le reste des occurrences. La sous-catégorie PP-LOC représente un exemple. En plus de la règle « PREP NP », elle est décrite par d'autres règles que chacune d'elles ne dépasse pas les 10 occurrences et qu'elles apportent ensembles uniquement 19 occurrences. Par ailleurs, le signe "#" affecté à quelques paramètres signifie leur cardinalité.

Catégories syntaxiques	PP	PP-CLR	PP-PRP	PP-TMP	PP-LOC	PP-PRD	PP-MNR	PP-DIR	Others
Σ# Règles	50	44	15	15	13	13	12	9	--
#Occ de « PREP NP »	12834	3025	445	754	1511	762	246	154	--
Σ#Occ des règles	13814	3781	684	805	1537	805	286	165	222

TABLE 1 : Fréquence de la règle « PREP NP » décrivant les sous-catégories de PP au niveau plus haut de granularité

Mais si nous observons la CFG, illustrée en partie dans la table 2, nous pouvons noter qu'indépendamment du niveau de granularité élevé des catégories syntaxiques, les occurrences de la construction « PREP NP » représentent en tout environ 90% des cas, ce qui rend l'augmentation de la granularité des catégories inutiles dans certains cas. La réduction de cette granularité à 0 nous donne une CFG pour les PP plus compacte. Elle est illustrée en partie par la table 2 et regroupe uniquement 59 types différents de règles intégrant dans la plupart des cas la construction « PREP NP ». Nous pouvons remarquer aussi que, dans les deux CFG, les constructions de PP les plus fréquentes dans le treebank sont formées de deux constituants. Par contre, les constructions complexes formées de plus que deux constituants sont rares. Généralement, nous avons trouvé que, pour toutes les catégories non terminales, le niveau de granularité des catégories affecte également la taille de toute la grammaire. En effet, le nombre de règles dans la CFG s'est divisé par 6 au niveau le plus bas par rapport à celui le plus élevé (2998/14452).

Règles	#Occ	Règles	#Occ	Règles	#Occ
PREP NP	19886	PREP ADVP	32	PP PREP NP	10
PREP SBAR	1346	NP PREP NP	28	PREP NP PUNC	10
PREP S	237	PREP PUNC NP	22	PREP UCP	8
PP CONJ PP	126	PP PP	20	PUNC PREP SBAR PUNC	7
-NONE-	87	PUNC PREP NP	20	PREP NP PP	6
PRT PREP NP	63	ADVP PREP NP	19	14 règles	≤5
PP PUNC CONJ PP	48	PREP PUNC NP PUNC	18	25 autres règles	1
PUNC PREP NP PUNC	42	Σ# Occurrences			22099

TABLE 2 : Extrait de la CFG au plus bas niveau de granularité décrivant la catégorie PP

La GP obtenue à un niveau donné décrit pour chaque catégorie syntaxique, l'ensemble de ses constituants ainsi que les propriétés qui relient ces constituants. La figure 7 et la figure 8 illustrent respectivement des extraits des GP obtenues aussi bien au niveau de granularité le plus élevé que celui le plus bas pour la catégorie PP. Premièrement, nous pouvons remarquer que, grâce au formalisme de GP, des informations implicites de différents types dans le treebank sont rendues explicites. Ce sont les propriétés (ou relations) qui relient les différents constituants. Ces nouvelles informations peuvent servir à la tâche d'analyse syntaxique. A partir de la figure 7, prenons comme exemple la propriété de linéarité « PREP $\prec$ S-NOM » décrivant la sous-catégorie PP-DIR désignant un syntagme prépositionnel de direction. Cette relation indique que si la catégorie PREP (préposition) apparaît avec la catégorie S-NOM (proposition nominative) dans la même réalisation, elle va toujours précéder S-NOM directement ou indirectement. Une information de ce type n'est pas explicite dans le treebank. Toutefois, avec un niveau de granularité totalement élevé des catégories, plusieurs informations implicites peuvent être répétées pour plusieurs sous-catégories, ce qui multiplie la taille de la GP et rend son parcours plus difficile. Plus particulièrement dans la figure 7, c'est le cas des propriétés reliant les catégories PREP et NP. Voyons ici que ces propriétés sont répétées au moins 6 fois dans la grammaire pour les sous-catégories indiquées.

PP-DIR	Const	{PP, PREP, NP, ADVP, S-NOM, SBAR, PRT}
	Unic	{PREP, NP, ADVP, S-NOM, SBAR, PRT}
	Lin	PP $\prec$ {PREP, NP} ; PRT $\prec$ {PREP, NP} ; PREP $\prec$ {NP, ADVP, S-NOM, SBAR}
	Exig	{NP, ADVP, S-NOM, SBAR} $\Rightarrow$ PREP; PRT $\Rightarrow$ {PREP, NP}
	Excl	PP $\otimes$ {ADVP, S-NOM, SBAR, PRT}; NP $\otimes$ { ADVP, S-NOM, SBAR} ADVP $\otimes$ {S-NOM, SBAR, PRT};S-NOM $\otimes$ {SBAR, PRT};SBAR $\otimes$ {PRT}

PP-DTV PP-TPC PP-LOC PP-MNR PP-PRD	Const	{PREP, NP}
	Unic	{PREP, NP}
	Oblig	{PREP, NP}
	Lin	PREP $\prec$ NP
	Exig	NP $\Rightarrow$ PREP PREP $\Rightarrow$ NP

FIGURE 7 : Extrait de la CFG au plus haut niveau de granularité

L'induction de la GP au niveau de granularité le plus bas montre une grande différence. La figure 8 montre un extrait de cette grammaire pour la catégorie PP. La GP devient beaucoup plus compacte, les catégories sont plus simples et les propriétés ne sont pas répétées. C'est parce que ces catégories ont été généralisées et factorisées Toutefois, ce manque de précision peut perdre l'information. Plusieurs exemples dans la GP peuvent prouver cette idée : Ainsi, avant la généralisation des catégories, la propriété de linéarité «PRON_3MS ≺ DET+ADJ+CASE_DEF_NOM» décrit la sous-catégorie NP-ADV-1. Après la généralisation, il faut que la précision de cette propriété soit réduite, et la propriété soit transformée à une autre étant « PRON ≺ ADJ » pour décrire la catégorie de base NP. Mais ceci n'a pas été effectué. Ceci s'explique par le fait que la validité de cette propriété n'a pas été garantie pour toutes les sous-catégories de NP. L'absence de plusieurs propriétés à cause de la généralisation peut produire un degré d'erreur.

Const	{-NONE-, NP, S, SBAR, PP, PREP, ADVP, PRT, CONJP, UCP, NAC, FRAG, PRON, TYPO}
Unic	{-NONE-, PREP, NP, ADVP, SBAR, PRT, PRON, UCP, NAC, FRAG}
Lin	-NONE- ≺ NP ; NP ≺ {UCP, NAC, FRAG ; PRT ≺ {PREP, NP, S, PRON, PP, SBAR, ADVP} ; PP ≺ {PREP, NP, S, NAC}; TYPO ≺ {PP, S}; PREP ≺ {NP, ADVP, S, UCP, PRON}; UCP ≺ PP
Exig	{NP, ADVP, S, SBAR} ⇒ PREP ; PRT ⇒ {PREP, NP};{CONJP, NAC, FRAG} ⇒NP
Excl	{TYPO, PRT, CONJP, PRON, ADVP, S, SBAR, -NONE-} ⊗ {UCP, NAC, FRAG} ; PRON ⊗ {NP, CONJP, TYPO} -NONE- ⊗ {S, SBAR, PP, ADVP, TYPO, CONJP, PRON, PRT}; NAC ⊗ FRAG; CONJP ⊗ {PRT, TYPO} S ⊗ {SBAR, PP, NP, ADVP, CONJP, PRT}; SBAR ⊗ {PP, NP, PRT, CONJP, PRON, ADVP, TYPO} PP ⊗ {ADVP, PRT, CONJP, PRON, FRAG} ; ADVP ⊗ {NP, PRON, PRT, CONJP, TYPO};PRT ⊗ TYPO

FIGURE 8 : Extrait de la GP au plus haut niveau de granularité

Dans la figure 8, nous pouvons remarquer qu'aucune propriété d'obligation n'est présentée. C'est parce que l'interprétation de ce type de propriétés exige la présence d'un constituant en une seule forme dans toutes les règles. En principe, ce constituant est la catégorie NOUN dans notre cas. Ceci n'est pas assuré, puisqu'à chaque fois, nous avons soit une catégorie plus détaillée (NOUN_NUM ou NOUN_PROP) ou bien une autre totalement différente. Tandis que les propriétés d'exclusion, elles sont très nombreuses en adoptant l'interprétation que nous venons de présenter dans la section 5.2. Considérer que toute absence de deux constituants d'une catégorie syntaxique dans toutes ses constructions représente une relation de restriction de cooccurrence n'est pas toujours valide. En effet, la rareté de ces règles dans la langue ou la non richesse du treebank en entrée peuvent être les vraies raisons de cette absence.

Selon les résultats obtenus, nous constatons que la variation du niveau de granularité des catégories a une grande influence sur la réduction de la complexité du problème permettant ainsi de diminuer la taille de la GP induite. Mais, même la généralisation pose un problème du faite qu'elle engendre une perte dans la précision de l'information. Ceci favorise beaucoup plus l'adoption d'un mécanisme de contrôle du niveau de granularité pour faire le compromis entre la généralisation et la spécification des catégories grammaticales de l'ATB.

7 Conclusions et perspectives

Nous avons proposé dans cet article une démarche de construction d'une GP à granularité variable à partir de l'ATB, ce qui la rend une ressource à large couverture héritant des qualités de l'ATB comme sa fiabilité, sa soumission à des consensus et sa richesse en annotations de différents types. La technique que nous avons adoptée pour construire cette ressource présente l'avantage d'être générique. En effet, elle est indépendante non seulement de toute langue, mais aussi, du formalisme source, puisque la génération des propriétés se fait directement à partir de la CFG. En plus, avec les types de propriétés définis jusqu'à présent, cette technique est automatique, ce qui favorise sa réutilisabilité pour des treebanks de différentes langues et de différents formalismes sources.

Dans le but, d'offrir une représentation très précise de l'information syntaxique, l'ensemble des relations présentées dans la GP peut toujours être enrichi ou modifié. Cette grammaire peut être également utilisée pour enrichir le treebank arabe ATB, qui est à base de syntagmes, avec la représentation à base de propriétés, ce qui peut permettre d'améliorer la qualité du treebank. Pour optimiser cet enrichissement, plusieurs mécanismes de contrôle peuvent être intégrés au niveau de la détermination des unités syntaxiques, et de l'efficacité de leurs propriétés linguistiques.

Références

ABEILLÉ A., CLÉMENT L., KINYON A. (2000). Building a treebank for French. Proceedings of *the Second International Language Resources and Evaluation Conference*. Athens, Greece.

AURAN C., BOUZON C., HIRST D.J. (2004). The Aix-MARSEC project: an evolutive database of spoken British English. Proceedings of *the Second International Conference on Speech Prosody*, 561-564. Nara.

BLACHE P. (2001). *Les Grammaires de Propriétés : Des contraintes pour le traitement automatique des langues naturelles*. Hermès Sciences Publications.

BLACHE P. (2014). A Chinese Constraint Grammar Extracted from the Chinese Treebank. Proceedings of *APCLC*.

BLACHE P., GUÉNOT M.-L. & VANRULLEN T. (2003). Corpus-based grammar development. Proceedings of *Corpus Linguistics-03*.

BLACHE P., RAUZY S. (2012). Hybridization and Treebank Enrichment with Constraint-Based Representations. Proceedings of *LREC-2012*.

DUCHIER D., PROST J.-P., DAO T.-B-.H. (2009). A Model-Theoretic Framework for Grammaticality Judgements. FG, volume 5591 of *Lecture Notes in Computer Science*, page 17-30. Springer.

KAY P., FILLMORE C. (1999).Grammatical Constructions and Linguistic Generalizations: the what's x doing y construction. *Language.*

LAMMIE GLEEN M., STRASSEL S. (2005). Linguistic Resources for Meeting Speech Recognition. *MLMI-05*.

MAAMOURI M., BIES A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. Proceedings of *COLING-04*. Geneva, Switzerland.

MAAMOURI M., BIES A., KROUNA S., GADDECHE F., BOUZIRI B. (2009). Penn Arabic Treebank guidelines v4.8. *Technical report, Linguistic Data Consortium*, University of Pennsylvania.

MAAMOURI M., ZAGHOUANI W., VIOLETTA CAVALLI-SFORZA V., GRAFF D., CIUL M. (2012). Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement. *NAACL-HLT 2012*. Montreal.

MOHRI M., ROARK B. (2006). Probabilistic Context-Free Grammar Induction Based on Structural Zeros. Proceedings of *the Seventh Meeting of the Human Language Technology conference- North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*. New York.

REHBEIN I., VAN GENABITH. J. (2007) Treebank annotation schemes and parser evaluation for German. Proceedings of *the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 630–639*. Prague, Czech Re public.

SONG Z., ISMAEL S., GRIMES S., DOERMANN D., STRASSEL S. (2012). Linguistic Resources for Handwriting Recognition and Translation Evaluation. *LREC 2012*. Istanbul.

TOUNSI L., VAN GENABITH J. (2010). Arabic Parsing Using Grammar Transforms. *LREC 2010*.

VANRULLEN T., BLACHE P., PORTES C., RAUZY S., MAEYHIEUX J.-F., GUENOT M.-L., BALFOURIER J.-M., BELLENGIER E. (2005). Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales. Actes de *TALN, pp. 41-48*. Paris, France.