

Université de Montréal

**Extraction des patrons de régime de VerbNet pour une
implémentation dans un système de génération
automatique de texte**

par

Daniel Galarreta-Piquette

Département de traduction et linguistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en linguistique

24 janvier 2018

SOMMAIRE

Ce mémoire explore les patrons de régime en anglais provenant de la ressource lexicographique *VerbNet* et leur implémentation dans un système de génération automatique de texte.

SUMMARY

English summary and keywords. . .

TABLE DES MATIÈRES

Sommaire	iii
Summary	v
Liste des tableaux	xi
Liste des figures	xiii
Liste des sigles et des abréviations	xv
Dédicaces	xvii
Remerciements	xix
Introduction	1
Chapitre 1. Extraction des entrées du dictionnaire <i>Verbnet</i>	3
1.1. VerbNet	3
1.1.1. Levin et l'organisation en classes verbales	4
1.1.2. Composantes de VerbNet	6
1.1.2.1. Organisation hiérarchique	7
1.1.2.2. Rôles thématiques	9
1.1.2.3. Restrictions sélectionnelles	9
1.1.2.4. Cadres	9
1.1.2.5. Prédicats sémantiques	10
1.1.2.6. Exemples	10
1.1.3. Raffinement de Korhonen et Briscoe	10

1.1.4.	Mapping de VerbNet à d'autres ressources NLP	10
1.1.5.	Utilisation de VerbNet dans des applications NLP	10
1.2.	Python	11
1.2.1.	Extraction de données des documents XML provenant de <i>VerbNet</i> ...	11
1.2.1.1.	Extraction des descriptions des patrons de régime	11
1.2.2.	Création du dictionnaire de patron de régimes (gpcon)	13
1.2.2.1.	Script pour faire les actants syntaxiques	14
1.2.2.2.	dictionnaire dans Python pour faire correspondre	14
1.2.2.3.	14
1.2.3.	Extraction des membres de chaque classe verbale	14
1.2.3.1.	14
Chapitre 2.	Génération automatique de texte	15
2.1.	Mécanisme dsynt=created->constrained->OK pour générer un énoncé simple (sujet, verbe, objet)	15
2.1.1.	application de la règle root_standard	15
2.1.2.	application de la règle lex_standard pour lexicaliser le verbe principal	15
2.1.3.	application de la règle actant_gp_selection	15
2.1.4.	application des règles actanciennes	16
2.1.5.	application de la règle constraints_gp	16
2.1.6.	application des règles de lexicalisation	17
2.1.6.1.	lex_standard	17
2.1.7.	Avantages d'utiliser ces mécanismes	17
2.2.	Énumérations	18
2.3.	Équations mathématiques	18
2.4.	Définitions, théorèmes et preuves	19

2.5. Construction d'un tableau	19
2.6. Référence à une entrée bibliographique.....	19
2.7. Insertion de figures	19
Bibliographie	21
Annexe A. Le titre	A-i
A.1. Section un de l'annexe A.....	A-i
Annexe B. Le titre2	B-i
Annexe C. Le titre3	C-i
Annexe D. Le titre4	D-i

LISTE DES TABLEAUX

1. I	Un tableau simple dans le premier chapitre.....	10
2. I	Un tableau simple dans le second chapitre.....	19
A. I	Titre alternatif pour la table des matières.....	A-i

LISTE DES FIGURES

2.1	Un cercle.....	20
2.2	Un carré et un triangle.....	20

LISTE DES SIGLES ET DES ABRÉVIATIONS

GAT	Génération automatique de texte
GP	Patron de régime, de l'anglais <i>Government Pattern</i>
DPOS	Partie du discours profonde, de l'anglais <i>Deep Part of Speech</i>
TST	Théorie Sens-Texte
VN	<i>VerbNet</i>

DÉDICACES

Vos dédicaces.

REMERCIEMENTS

Remerciements. . .

INTRODUCTION

Parler de la génération de texte automatique, de *VerbNet*, de la problématique (pas de consensus quant à l'architecture de la classe verbale en TAL) de VerbNet, des dictionnaires. Pourquoi les verbes sont si importants ? (Kipper, 2005, dissertation) Puisque les verbes sont porteurs du sens principal de la phrase, il faudrait donc créer une ressource qui puisse rendre compte du sens des verbes si on souhaite un bon fonctionnement des applications NLP.

Chapitre 1

EXTRACTION DES ENTRÉES DU DICTIONNAIRE *VERBNET*

Dans ce chapitre, nous verrons l'apport que la ressource lexicale *VerbNet* peut offrir à des applications de TAL. Nous avons comme objectif d'extraire l'architecture concernant l'organisation de la classe verbale qu'on retrouve dans le dictionnaire *VerbNet* pour les implémenter dans le dictionnaire de notre système de génération de texte automatique (GAT). Cet objectif provient d'une problématique que nous avons rencontré auparavant. Nous voulions savoir comment organiser la classe verbale au sein de notre système, car les verbes sont si riches et complexes qu'il nous fallait trouver un moyen systématique d'encoder cette partie du discours, et l'encoder pour que notre dictionnaire puisse interagir correctement avec notre grammaire. Justement, en ce qui concerne les dictionnaires, parmi la communauté TAL, il ne semble pas y avoir de consensus quant à la manière de procéder pour modéliser la classe verbale. La raison est simple, les entrées lexicales faisant partie de cette PDD démontre des comportements variables, très riches au niveau de l'éventail de patrons de régime possibles pour un même verbe, et assez complexes ce qui nécessite beaucoup plus d'attention que d'autres parties du discours comme les noms qui démontrent beaucoup moins de variétés d'usage quant au nombre de patrons de régime. Cette problématique nous a donc amené à voir comment une ressource comme VN, qui s'est donnée comme mission de combler cette lacune, a fait pour résoudre le problème.

1.1. VERBNET

(Karin Kipper, Hoa Trang Dang, and Martha Palmer, 2000) et Dissertation de Kipper Ainsi, tel que mentionné précédemment, *VerbNet* a été créé dans un contexte où il y avait un réel besoin de réfléchir à la meilleure manière de procéder pour faire un dictionnaire qui saura tenir compte de la complexité que renferment les verbes. C'était en réponse au manque de lignes directrices sur l'organisation des verbes dans les dictionnaires destinés à des

applications TAL et ce malgré la quantité impressionnante de dictionnaire computationnels existants. Parmi ceux qui ont tenté la chose, nous nommerons : le " generative lexicon" de Pustojevsky WordNet, ComLex, LCS et FrameNet, ceux-ci comportaient des lacunes que du point de vue du traitement des entrées lexicales verbales, selon les auteurs de VerbNet. Les lacunes étaient diverses mais importantes, par exemple : le Generative Lexicon se centre sur les noms, WN ne donne pas de détails concernant les cadres syntaxiques possibles pour un verbe ainsi qu'un vide complet sur la structure prédicat-argument, ComLex offre des cadres syntaxiques mais ne distinguent pas les différents sens qu'un verbe peut prendre, le lexicon LCS essaye de palier à ces lacunes aussi, mais ne couvre pas assez large.

1.1.1. Levin et l'organisation en classes verbales

Les classes verbale sont importantes car elles exposent le caractère général qu'on retrouve dans le comportement des verbes. Ainsi, au lieu de construire un dictionnaire où chaque entrée verbale est prise individuellement, on regroupe les entrées lexicales en classe en fonction de leur comportement syntaxique et sémantique. À la base le traitement est fait en fonction du comportement syntaxique, puis selon Levin, les verbes qui partagent le même type d'alternations syntaxiques (patrons de régime) partagent aussi la sémantique sous-jacente à ces constructions, donc cela fait en sorte qu'ils partagent généralement les mêmes caractéristiques sémantiques aussi. Le fait d'organiser le dictionnaire en classe verbale accélère le processus de création du dictionnaire, car une fois qu'on détermine le nombre de verbes qui fonctionnent de la même manière, il ne reste plus qu'à faire une entrée qui retient l'information syntaxique et sémantique partagée par l'ensemble de la classe, et passer à la classe suivante. Après, lorsqu'on a fait le tour des verbes, on a juste à les ajouter à l'une des classes pré-existantes. Les créateurs de VerbNet sont conscients du fait qu'on va parfois tenter de faire rentrer un verbe dans une classe et que c'est un brin tirer par les cheveux, mais il s'agit d'une infime partie du dictionnaire. (p-t trouver un exemple,Dissertation p.2). Ainsi, les verbes d'une même classes partagent des comportements syntaxiques similaires. Ce qui a été démontré par Jackendoff, qui montre qu'à travers les langues, on peut identifier des classes verbales. Plusieurs applications TAL ont tiré profit d'organiser les verbes en classes verbales (p.2, Kipper), mais contrairement à ce que VerbNet et nous essayons de faire, ces dictionnaires étaient domaine-spécifique et non domaine-indépendant.

Tel que Levin le dit (Levin,1993), son travail a été guidé par l'idée que les comportements syntaxiques des verbes sont motivés par les composantes sémantiques qui les représente. À ces fins, Levin a observé les comportements des verbes en anglais pour y tester cette hypothèse linguistique car il pensait que c'était une bonne manière de démontrer la chose. Si des verbes qui partagent le même type de comportement syntaxique, ils doivent aussi probablement partager des caractéristiques sémantiques. Son travail a été de classer les verbes

systématiquement. Déjà, Levin soulignait l'importance d'encadrer les verbes dans les lexicons car les verbes qui sélectionnent des arguments démontrent un ensemble complexe de propriétés.

Les créateurs de VerbNet ont jugé qu'il serait nécessaire d'organiser leur information en classes verbales. Ce qui est un premier pas vers une compréhension de leur hypothèse sur la manière d'organiser un dictionnaire autour de verbes. Ce que nous avons fait par le fait même puisque c'est cette idée que nous avons recyclé de VerbNet. Ça s'est avéré effectivement tel que VerbNet l'a explicité. L'idée d'organiser l'information ainsi leur est venu de Levin qui avait déjà commencé classer les verbes et avait fait un travail considérable sur ce sujet.

Comment fonctionnent les classes de Levin ? Elles démontrent très clairement le comportement syntaxique des verbes, mais ne tiennent pas compte des composantes sémantiques d'une classe de verbe. Les classes regroupent des verbes qui agissent de la même manière d'un point de vue syntaxique. Autrement dit, des verbes qui participent aux mêmes cadres syntaxiques sans que ça n'altère leur sens. De plus l'aspect sémantique de la chose vient du fait que Levin soutenait que sans que les verbes veuillent dire la même chose, les verbes avec une sémantique similaire démontreront aussi des caractéristiques syntaxiques similaires. (d'un point de vue de la structure argumentale). Les cadres syntaxiques associés aux classes sont supposés représenter les composantes sémantiques sous-jacentes et communes à une classe de verbe. Ainsi, en fonction du sens d'un verbe, dans les aspects primitifs des caractéristiques sémantiques, certains verbes vont agir de la même manière car ils partagent cette caractéristique sémantique. Certains verbes partageront ainsi une partie des cadres syntaxiques, mais pas tout, car une composante sémantique leur appartenant fera en sorte qu'un autre cadre syntaxique serait incorrect. Voir l'exemple suivant, tiré de (Kipper, 2005) et à la base de (Levin, 1993).

Voici la mentalité de Levin en action :

- (1) Cadre syntaxique : transitif direct
 - a. John broke the window.
 - b. John cut the bread.
- (2) Middle construction
 - a. Glass breaks easily.
 - b. This loaf cuts easily.
- (3) Intransitive construction
 - a. The window broke.
 - b. *The bread cut.
- (4) Conative construction

- a. *John broke at the window.
- b. John valiantly cut/hacked at the frozen loaf, but his knife was too dull to make a dent in it.

Ainsi, tel que démontré en (1) et (2) les verbes appartenant à la classe *break* et à la classe *cut* se ressemblent car ils peuvent tous les 2 prendre ces cadres syntaxiques (ce genre de construction syntaxique). Toutefois, en (3) et en (4), on voit très bien qu'il ne partagent pas ces constructions syntaxiques. *Break* prend seulement la construction intransitive et exclut l'autre, tandis que *cut* prend la construction conative et exclut l'intransitive. La raison que Levin donne est la suivante. Le verbe *cut* décrit une série d'actions ciblant la complétion d'un but (séparer un objet en morceau). Toutefois, il est possible de faire ces séries d'actions sans que l'objectif final ne soit atteint, mais l'action de couper peut quand même être perçue. En ce qui concerne *break*, la seule chose qui importe dans l'évènement, c'est le changement d'état d'un objet (qui devient séparés en morceaux). Si on n'arrive pas au résultat, une tentative de briser quelque chose ne peut être perçue. Ce qu'on peut tirer de cet exemple, est que les classes verbales regroupent des verbes qui partagent des comportements syntaxiques similaires, les membres des classes ne sont donc pas nécessairement des synonymes, il ne s'agit que de verbes qui s'utilisent de la même façon. Ainsi, certaines classes vont effectivement regrouper des verbes qui signifient à peu près la même chose, mais aussi des verbes qui en surface, ne partagent pas bcp avec une majorité des verbes de cette classe, mais syntaxiquement, ces membres se comportent comme le reste de la classe. Ainsi, selon Levin, il y a quelque chose derrière les composantes sémantiques de ces membres qui les unirait avec les autres qui se ressemblent. (à réécrire, ce passage est horrible)

Les auteurs de VerbNet ont aussi ajouter des classes à ce que Levin avait fait. Ce qu'ils appellent intersective Levin classes, ce sont des sous-ensembles de classes qui s'entrecoupent et forment des classes à part. (p.26)

De plus, les structures de Moens et Steedman (temps et aspect), basés sur les travaux de Vendler. démontrent que les classes de Levin se tiennent. La sémantique des évènements qui relie des verbes entre eux, correspond sans problème aux verbes qui se retrouvent dans les classes de Levin (ce qui est confirmé par le postulat de Levin qui pense que le fait que certains verbes agissent de la même manière est due au fait qu'ils partagent une sémantique sous-jacente similaire bien qu'il ne soient pas des synonymes).

1.1.2. Composantes de VerbNet

VerbNet est composé de classes verbales, tel que nous l'avons mentionné précédemment, et pour les raisons évoquées. Chaque classe contient un ensemble de membres qui lui sont attribués en fonction du caractère commun que les membres possèdent avec la classe, des

rôles thématiques pour la structure argument-prédicat et des restrictions sélectionnelles sur ces arguments ainsi que les cadres syntaxiques qui contiennent une courte description du cadre, un exemple le démontrant et une description syntaxique du cadre. Puis finalement, un ensemble des prédicats sémantiques. Une classe peut être divisé en sous-classe si cela est nécessaire. Tel sera le cas lorsque un sous-ensemble des membres de la classe partage des cadres syntaxiques et des prédicats sémantiques spécifiques. Ce qui est l'aspect hiérarchique de la chose.

1.1.2.1. *Organisation hiérarchique*

(provient de Guidelines) et Dissertation

Les raisons qui ont motivé VerbNet a organisé son information en hiérarchie. D'abord, il se sont fortement inspiré de Acquilex Lexical Knowledge Base. Ensuite, pour des raisons pratiques, ils ont trouvé que c'était la meilleure manière d'organiser cette montagne d'information lexicale car ils ont revu l'information fournie par Levin et ont ajouté des sous-classes aux classes ainsi que des sous-sous-classes aux sous-classes, allant jusqu'à un niveau de 3 steps de profondeur. Donc, des sous-ensembles des sous-ensembles. Ce qu'ils considèrent comme une raffinement des classes originales de Levin (intersective, adding subclasses, removing classes, adding classes). En gros, c'est l'ajout de sous-classes aux classes originales de Levin (et aux sous-sous-classes) qui ont rendu VerbNet hiérarchique.

Le fonctionnement est assez simple. Une sous-classe fille hérite de toute l'information de sa classe mère. Les sous-classes sont à l'origine de vouloir spécifier le comportement qui rassemble un sous-ensemble des membres d'une classe. Ainsi, une sous-classe est un ajout d'information par rapport à des restrictions d'usages de rôles thématiques, de cadres syntaxiques ou de prédicats sémantiques.

Prenons un exemple, la classe verbale *Transfer of a Message* séparer l'exemple en 3 parties (en fonction des niveaux).

Cette classe comporte trois niveaux, en termes de hiérarchie. Au premier niveau, nous avons les rôles thématiques, cadres syntaxiques et prédicats sémantiques partagés par tous les membres de la classe. En ce qui concerne cette classe, il y a au premier niveau, la spécificité qu'il s'agit d'une classe verbale mère avec son ID et VNCLASS. Puis les membres de la classe. Puis les rôles thématiques. Puis la section FRAMES qui contient tous les FRAME. Chaque FRAME est composé de : une description, un exemple, la syntaxe, puis la sémantique. Ainsi, au premier niveau, on retrouve 10 FRAME dans la section FRAMES. Par après, on accède au second niveau, qui est celui des sous-classes. Il n'existe qu'une sous-classe dans cet exemple, mais certaines classes verbales de VN contiennent plus d'une sous-classe au même niveau. Lorsque c'est le cas, il s'agit de classes soeurs. qui sont toutes deux enfant du même parent.

Revenons à notre cas, ainsi, nous n'avons qu'une sous-classe au second niveau. Cette sous-classe garde la même structure que la classe mère. Ainsi, d'abord, on a la précision qu'il s'agit d'une sous-classe et son identification. Puis on a ses membres, suivi des rôles thématiques, dans ce cas-ci, il n'y a pas d'ajout à faire pour cette section puisque la sous-classe utilise les mêmes rôles thématiques que la classe mère, ainsi, par héritage, elle hérite de tous les rôles thématiques mentionnées précédemment. Puis on entre dans sa section de FRAMES qui contient tous les FRAMES appartenant à cette sous-classe. Ainsi, les membres de cette classe hérite des FRAMES de la classe mère en plus des FRAMES qui leurs sont plus spécifiques, car ces membres partagent des caractéristiques syntaxiques et sémantiques de plus que par rapport aux membres de la classe mère. Dans cet exemple, il y a un FRAME. Puis on accède au troisième niveau. Une sous-classe de la sous-classe précédente. Ainsi, le système qu'a adopté VerbNet pour distinguer les concepts de child/sister/parent sont des numérotation des classes (et sous-classes). Ainsi, on sait à coup sûr qu'il s'agit d'une sous-classe de la sous-classe car, au second niveau il s'agissait de la sous-classe transfer mesg-37.1.1-1 et au troisième niveau on a transfer mesg-37.1.1-1-1 et s'il y avait plus qu'une sous-classe à ce niveau, on aurait vu transfer mesg-37.1.1-1-2.

Alimenter cette partie avec les infos provenant du Guidelines. Les classes verbales de VerbNet sont numérotées par des chiffres allant de 9-109. Ainsi, le numéro apparaissant devant une classe verbale est associé à des caractéristiques sémantiques et syntaxiques communes. Par Exemple, les classes verbales associées à des verbes de type "mettre quelque chose" commenceront par le chiffre 9. Ce qui nous donne quelque chose comme :

put 9.1 put spatial 9.2 funnel 9.3 put direction 9.4 pour 9.5 coil 9.6 spray 9.7 fill 9.8 butter 9.9 pocket 9.10

Certains numéros n'impliquent qu'une seule classe, car il n'y a pas d'autres classes qui partagent ce genre de traits sémantiques ou syntaxiques communs.

Par la suite, on se sert encore de la numérotation pour expliciter la hiérarchie à l'intérieur même d'une classe. Chaque classe peut inclure des classes filles, qui sont des classes sœurs entre elles, et qui peuvent avoir des classes filles à leur tour. La classe verbale Spray en démontre bien la chose.

D'abord, il y a la classe supérieure, qui est la plus haute de la hiérarchie, toutes les caractéristiques de cette classe sont partagées par tous les verbes de la classe. Dans la top classe nous avons les constructions syntaxiques et les prédicats sémantiques partagés par la classe, ainsi que les rôles thématiques.

Une classe mère domine une sous-classe, toutes ses caractéristiques sont partagées avec les classes subordonnées à celle-ci.

Une sous-classe : Les sous-classes dans VerbNet héritent des caractéristiques provenant de la classe dominante, mais elles spécifient particulièrement des constructions syntaxiques et une sémantique entre les verbes membres de cette sous-classe. Ce qui est spécifié dans une sous-classe peut être de différents ordres : ajouter des constructions syntaxiques propres à ce sous-groupe, ajouter des restrictions sélectionnelles sur les étiquettes des rôles sémantiques.

Des classes sœurs, ne partagent pas de caractéristiques hormis celle héritées par leur classe mère. Le reste de l'information encodée dans la classes sœurs n'est valide qu'à l'intérieur de leur classe respectivement.

1.1.2.2. *Rôles thématiques*

VerbNet utilise un ensemble de 23 rôles thématiques pour identifier les arguments dans les classes verbales. On étiquette les arguments dans les classes verbales en leur associant un rôle thématique. La raison pour laquelle VerbNet a opté pour les rôles thématiques est que, contrairement à un étiquetage générique où on énumère les arguments en procédant comme "Argument 1" Verbe "Argument 2" pour illustrer un cadre syntaxique est parce que les rôles thématiques peuvent offrir de l'information sémantique de plus que juste un argument numéroté. La spécification du rôle fournit de l'information sémantique sur le type d'argument en jeu, tandis que numéroté ne fournit rien du tout. Chaque argument se fait donné un rôle thématique unique, et ces rôles thématiques sont partagés par tous les membres d'une classe. Donc, ils doivent être assez large pour que ce soit cohérent avec tous les membres, mais pas trop imprécis non plus.

(Dissertation) :

VerbNet utilise les rôles thématiques pour étiquetter les arguments figurant des les cadres syntaxiques (et sémantiques). Elle puise dans une banque de 23 rôles thématiques pour associer le bon rôle à l'argument en question. Ils ont choisi ces rôles car ils étaient assez généraux pour se prêter à toutes les événements que dégagent les verbes dans le dictionnaire. Ils voulaient capturer l'essence des verbes et démontrer encore une fois le caractère général des verbes en démontrant qu'une poignée d'argument peut bel et bien rendre compte des arguments sélectionnés par les verbes en anglais. À l'intérieur d'une même classe verbale on y retrouve un nombre x de rôles thématiques qui seront mappés aux arguments sélectionnés dans les cadres syntaxiques et sémantiques fournis par VerbNet. Ils ont choisi les rôles thématiques au lieu d'autres moyens d'étiquettage des arguments car ils trouvaient que ça ajoutait de l'information sémantique sur une classe verbale.

(guidelines) :

À la fin de ce paragraphe, peut-être faire un clin d'oeil au fait qu'on ne va pas utiliser cette technique et qu'on a procédé autrement.

1.1.2.3. *Restrictions sélectionnelles*

Les restriction sélectionnelles vont sur les rôles thématiques. Est-ce que c'est pertinent d'en parler ? Probablement, pas, on ne s'en sert pas du tout.

1.1.2.4. *Cadres*

Pour une classe donnée, on y retrouve soit un ou des cadres syntaxiques qui servent à représenter le type de réalisation de surface possible pour cette classe verbale. D'ailleurs, ces cadres syntaxiques sont partagés par l'ensemble des membres d'une classe ou d'une sous-classe. Chaque cadre syntaxique décrit une construction verbale de type transitives directes/indirectes, des intransitives, des phrases prépositionnelles, etc. Un cadre syntaxique est composé de rôles thématiques dans leur position argumentale ainsi que le verbe qui les régie (ainsi que d'autres unités lexicales nécessaires au bon fonctionnement d'une construction). Agent V Patient

1.1.2.5. *Prédicats sémantiques*

Est-ce que ça vaut la peine que j'en parle ? Leur manière de voir la sémantique est wierd as fuck

1.1.2.6. *Exemples*

Chaque cadre syntaxique est accompagné d'un exemple pour exemplifier ce que le cadre représente. On peut ainsi mieux comprendre comment la classe verbale fonctionne. Il me faudrait faire des screenshots. Et les rajouter à cette section pour montrer comment les frames fonctionnent.

1.1.3. Raffinement de Korhonen et Briscoe

1.1.4. Mapping de VerbNet à d'autres ressources NLP

FrameNet, WordNet, Xtag

1.1.5. Utilisation de VerbNet dans des applications NLP

Encore du texte... et un tableau

Le tableau 1. I n'est pas très garni.

exemple premier element

second exemple

TABLEAU 1. I. Un tableau simple dans le premier chapitre.

Option	g	c	d	p{0.4\textwidth}
Effet	À gauche	Au centre	À droite	Le texte de cette colonne est justifié et la largeur de la colonne est fixée à 40 % de la zone de texte (hors tableau).

1.2. PYTHON

1.2.1. Extraction de données des documents XML provenant de *VerbNet*

Tel que mentionné dans la section précédente concernant l'architecture de VerbNet. Nous avons vu comment les documents XML sont arrangés. Ainsi, à l'aide du module *xml.etree.cElementTree* nous avons pu faire des opérations sur l'ensemble des données de VN.

1.2.1.1. *Extraction des descriptions des patrons de régime*

(arranger mon info en sous-sous-sous-section)

Dans les feuilles XML de VerbNet, l'information que nous cherchions pour améliorer notre système MATE était : tous les patrons de régimes possibles pour une classe de verbe. Au départ, nous ne savions pas encore ce que nous voulions extraire de ces feuilles XML, elles abondent en information. Toutefois, il semblait évident que nous pourrions vraiment en tirer parti. Nous avons donc commencé par extraire l'information se trouvant sous SYNTAX en pensant que le syntactic frame était ce que nous utiliserions pour construire notre lexicon. Toutefois, nous nous sommes vite rendu compte dans le processus que ce n'était pas exactement ce que nous voulions. Ça nous est apparu évident lorsque nous avons extrait les descriptions des syntactic frames, les exemples, puis les données sous SYNTAX. En effet, *NP V S INF ['NP', 'VERB', 'NP'] I loved to write.* un exemple comme celui-ci nous montre que les informations sous SYNTAX ne correspondent pas à ce que nous cherchions, car la description du patron de régime et le patron de régime en soi diffèrent. Ils mettent que "to write" correspond à un NP. Ce qui n'est pas le cas quand on y pense bien. De plus, lorsqu'on extrait les balises sous SYNTAX, on a décidé de ne pas extraire les attributs contenus sous les balises car c'était de l'information soit sur les rôles thématiques ou sur les prépositions.

Nous avons donc utilisé VerbNet de mieux que nous le pouvions. Ainsi, les descriptions des FRAMES SYNTACTIC étaient "accurates" et allait nous donner les brèves descriptions

des patrons de régime que nous ajouterons à notre lexicon. Ainsi, pour l'exemple mentionné NP V S INF, ce que nous en retenions, c'est que pour cette classe de verbe, il existe un patron de régime où on a un sujet, puis un verbe, et ensuite une proposition infinitive comme complément d'objet direct. Cette information était suffisante pour créer le dictionnaire de patron de régime. Car, nous pensions au départ que nous pourrions tout prendre de verbnet pour créer notre lexicon. Nous voulions chercher les descriptions des patrons de régime ainsi que l'information sous SYNTAX sous FRAME pour ainsi créer les patrons de régime en soi en les traduisant dans notre langage (TST). Toutefois, leur manière d'encore les patrons de régime ne correspond pas à la notre sur beaucoup trop d'aspect. Nous le verrons plus tard, il y a de la redondance à certains moments et notre théorie rend mieux compte des patrons de régime. Mais pour l'instant, ce qui est important de noter c'est que nous notons les patrons de régime en attribuant des actants I et des noms de relation pour signifier que est le rôle de cet actant lors du passage de la sémantique à la syntaxe. VerbNet ne fait pas cela de la même manière que nous. Ils ont aussi un volet sémantique, mais qui ne se branche pas à notre modèle théorique. Nous voulions donc nous inspirer quand même de leur méthode pour nos patrons de régime, mais le fait qu'ils utilisent les rôles thématiques nous posait problème. Il était difficile d'associer un rôle thématique à un actant syntaxique (bien que nous ayons tenté [montrer le graphique de F.L]). Donc, nous en sommes venus à la conclusion qu'il nous faudrait créer les patrons de régime en Python pour ensuite les exporter dans un format adéquat pour MATE. Pour ce faire, nous devions utilisé les descriptions des patrons de régime qui se retrouve dans *VerbNet* et à partir de la description, créer une fonction qui nous permettrait de générer rapidement des patrons de régime adéquat pour MATE en ayant simplement à remplir les trous.

Pour extraire les descriptions des patrons de régime nous avons utilisé deux fonctions. D'abord la fonction *treeframe* qui s'occupe de récupérer la description de chaque frame syntaxique à travers tout VN. Nous passons à travers tous les frames existant dans les XML de VN. Autant dans les classes que les sous-classes. Toutefois, cette fonction ne fait pas que récupérer la description du syntactic frame de VN et nous la recrache tel quel. Nous faisons quelques opérations sur les descriptions que nous extrayons. Notamment, nous remplaçons tout espace, tiret, point, barre oblique, parenthèses qui pourraient se trouver dans les descriptions par des `_` à l'aide d'expressions régulières. Puis, nous retirons certaines descriptions de gp à cause de leur caractère problématique à encoder (pour des raisons théoriques, logiques – à expliquer ailleurs). Puis une fois qu'un clean up a été fait des descriptions et que nous avons uniquement celles que nous voulions, nous procédons à une seconde étape de raffinement des descriptions. Nous utiliserons une seconde fois une expression régulière pour trouver tous les occurrences de PP car nous ajouterons les prépositions impliquées dans les PP comme tel. Pour ce faire, nous faisons un search de tous les prépositions existant dans les

patrons de régime de verbnets(et non dans la description, c'est là que les patrons de régime de VN nous ont été utiles) et nous mettons les prépositions que nous soutirons des gp directement dans le nom de la description des gp à la suite du mot PP à chaque fois qu'on retrouve le mot PP dans une description. On obtient ainsi les descriptions nettoyées de celle qu'on ne veut pas, avec uniquement des underscore pour séparer les syntagmes puis les prépositions (lorsqu'il y a lieu) dans les noms des descriptions de gp.

Puis, une fois que nous faisons ces opérations sur les classes de VerbNet, nous avons scripté une méthode pour que la fonction s'applique aussi aux sous-classes.

Par la suite, nous avons créé une fonction *super* afin que la création du lexicon s'agence bien avec la manière que MATE fonctionne. Cette fonction nous permet d'utiliser le mécanisme d'héritage qui existe dans MATE. Ainsi, on peut renvoyer une sous-classe à la classe(ou sous-classe) qui la domine. De plus, cette fonction limite le nombre de descriptions de gp. VerbNet a aussi ce mécanisme d'intégrer autrement. Ainsi, si une classe X a 10 descriptions et qu'une sous-classe Y en a 5, mais que les 10 descriptions de la classe s'appliquent aussi à la sous-classe, on n'aura pas 15 descriptions mais juste 5 dans le sous-classe, car le mécanisme d'héritage s'occupe de faire ça. On a aussi programmé la création de notre lexicon pour que si il n'y a pas de sous-classes, que la classe hérite de la classe verb afin d'avoir les attributs de base de cette classe (la dpos, la spos, voir MATE). De cette manière, tous les classes héritent de la classe verb si on remonte aux classes mères, ainsi, on n'a pas besoin de préciser à chaque fois la dpos/spos.

Puis finalement, nous utilisons la fonction write qui écrit le tout dans un fichier .dict. Nous loopons à travers tous les fichiers compris dans le dossier VN. Puis nous allons chercher les keys et values du dictionnaire créé à l'aide de la fonction treeframe et du dictionnaire super.(À revoir)

Nous avons aussi pris la peine d'extraire les exemples pour chaque description car il s'agit d'une information utile pour voir dans quel contexte s'utilise ce patron de régime. Et c'est une information pertinente à avoir dans un dictionnaire, donc nous l'ajouterons à notre lexicon pour notre système de GAT. Pour ce faire, nous opérons de la même manière que pour les descriptions de gp dans le sens où nous faisons l'extraction d'exemples en passant par les feuilles XML et le module d'extraction xml.etree. Pour chaque frame, nous passons la fonction à l'entièreté de VN et nous nous arrêtons à la balise EXAMPLES puis à tout texte se trouvant dans cette balise. Nous avons voulu extirper les exemples alongside des descriptions car c'est de l'information pertinente à avoir dans un dictionnaire.

[Avant de passer à la prochaine section, il faut parler de la manière que mon dictionnaire fonctionne avec les keys() et values() avant que je le passe à la fonction write pour en extraire des parties]

1.2.2. Création du dictionnaire de patron de régimes (gpcon)

Soit le mentionner ici, ou ailleurs, mais il a fallu faire un dictionnaire de patron de régime. D'abord, parce qu'on s'est rendu compte que du à toute l'information qu'on allait chercher et la différence dans le type d'information, on a jugé bon de créer un second dictionnaire qui ne contiendrait que les gps, autrement dit un gpcon. Celui l'information sur les patrons de régime (les actants syntaxiques). Il existe x nombre de gps répertoriés. Nous les avons trouvé en faisant un ensemble à partir de tous les descriptions que nous avons obtenus avec le script précédent. Une fois que nous avons l'ensemble des gps différents. Il nous fallait les créer, car tel que mentionné, nous ne pouvions pas extraire les gps de VerbNet dû à une différence trop grande (cadre théorique et application). Notre système de GAT fonctionne avec la théorie Sens-Texte et nous pensons que c'est la théorie qui s'y prête le plus pour faire ce type d'opérations et qui tient le mieux compte de la manière dont le langage fonctionne. Ainsi, nous avons créer le gpcon à partir de Python car un bon nombre d'opérations peuvent être automatisés (éviter les fautes, et c'est plus transparent). Pour la création du gpcon, notre dictionnaire en Python ressemblait à ça. Nos keys() étaient la description du gp et les valeurs étaient les actants syntaxiques impliqués dans ce gp (avec de l'information sur les actants syntaxiques nécessitant une préposition à réaliser). Selon l'ordre dans lequel figure nos objets dans la liste qui est ce qu'on retrouve dans values(), notre fonction va assigner le bon actant syntaxique(I, II, III, etc.) ainsi, cette partie est automatisée grâce à cette fonction. Après, pour l'objet "subj" on va lui assigner une string 'rel=subjective dpos=N' ce qui est encodé dans une autre cellule. Ainsi à chaque fois qu'un gp a un subj, on n'a pas à écrire ce que subj contient. Alors pour l'objet subj, on aura I et 'rel=subjective dpos=N'. C'est l'union de la fonction gp et de la fonction roman qui nous permettent d'assigner les bons actants syntaxiques aux objets dans la liste qui représente les values dans mon dictionnaire de gpcon.

1.2.2.1. *Script pour faire les actants syntaxiques*

1.2.2.2. *dictionnaire dans Python pour faire correspondre*

1.2.2.3.

1.2.3. Extraction des membres de chaque classe verbale

1.2.3.1.

Chapitre 2

GÉNÉRATION AUTOMATIQUE DE TEXTE

Ce sera une section du mémoire sur le mécanisme `dsynt=X`

2.1. MÉCANISME `DSYNT=CREATED->CONSTRAINED->OK` POUR GÉNÉRER UN ÉNONCÉ SIMPLE (SUJET, VERBE, OBJET)

2.1.1. application de la règle `root_standard`

Cette règle crée un noeud qui sera la racine de l'arbre syntaxique. Ce noeud se fait imposer des contraintes. Notamment, on demande à ce que ce soit un lexème appartenant à la partie du discours : verbe et que sa finitude soit de type : fini. On impose à ce noeud le trait `dsynt=constrained` pour que ça s'harmonise avec la règle `lex_standard`, mais c'est à revoir.

2.1.2. application de la règle `lex_standard` pour lexicaliser le verbe principal

On assigne un lexème à un noeud créé en syntaxe profonde. Dans le contexte actuel, on l'utilise pour sélectionner l'unité lexicale qui matche les contraintes énoncées sur le noeud vide créé par la règle `root_standard`. Ce lexème provient du lexicon. Lorsque la règle s'applique et qu'elle consomme le noeud en y mettant la bonne lexicalisation (qui respecte les contraintes sur le noeud), on ajoute un trait `dsynt=OK` pour signifier que le sémantème a été réalisé en syntaxe profonde et qu'on ne fasse plus d'opérations sur ce noeud.

2.1.3. application de la règle `actant_gp_selection`

Cette règle s'applique lorsque nous avons un prédicat. On crée une variable `[?GP]` qui nous fournit un chemin vers l'information encodée sous l'attribut `gp` d'un lexème `[?X]`. Puis, on extirpe les traits `id` et `dia` pour chaque attribut `gp` de notre `[?X]`. Une fois qu'on a récupéré ces informations, on les appose au lexème en question car on se servira de ces informations

pour l'application de règles subséquentes. Le trait *id* représente la description du patron de régime (chaque description se retrouve dans notre gpcon qui est un dictionnaire de *gp*) et le trait *dia* nous renseigne sur la diathèse de ce patron de régime, c'est-à-dire combien d'actants sont en jeu, et dans quel ordre sont-ils placés? Il est essentiel qu'un lexème verbal aille chercher ces traits car il en a besoin pour appliquer les règles actanciennes qui en découlent. Il faut que le système sache quel patron de régime utilisé pour un prédicat donné, et dans quel ordre les actants seront réalisés en syntaxe.

2.1.4. application des règles actanciennes

Une fois qu'un gp est sélectionné, on appliquera la règle actancielle qui lui correspond. Nos règles actanciennes ressemblent à : *actant_gp_ijk*. Les règles prennent en input les arcs sémantiques liant les actants à leur prédicat. Ces règles génèrent des noeuds vides auxquels on appose un trait *dsynt=created* pour signifier qu'on vient de créer des noeuds vides (on dit qu'ils sont vides car ils n'ont pas encore été consommés par une unité lexicale) en syntaxe profonde. On obtient ainsi des arcs syntaxiques au bout desquels se trouve un noeud vide. Ces règles se font imposer des conditions bien strictes. Il faut d'abord que le prédicat qui les gouverne soit lexicalisé. Ce qui se traduisait par l'ajout d'un trait *dsynt=OK* au lexème lexicalisé (avec la règle *lex_tandard*). La règle *actant_gp_selection* nous permettait de soustraire les traits *id* et *dia* et c'est ici que le trait *dia* entre en jeu. On s'en sert pour illustrer la diathèse du verbe. [à revoir]

2.1.5. application de la règle *constraints_gp*

Cette règle applique des contraintes à des noeuds nouvellement créés (autrement dit, des noeuds qui ont le trait *dsynt=created*). C'est à partir de cette règle que le mécanisme de *dsynt=created->constrained->OK* prend vie. Il s'agit d'une étape intermédiaire entre la création du noeud avec les règles actanciennes et la lexicalisation qui consommera le noeud en octroyant un lexème suivant un trait *dsynt=OK* (signifiant que la réalisation en syntaxe profonde est terminée). Ainsi, la règle procède de cette manière. On prend un X (un prédicat) qui lie un Y (un actant) puis ce dernier se fait imposer des contraintes. On cherche à lui imposer ces contraintes car on souhaite qu'il respecte le patron de régime de X. Ainsi, s'il ne convient pas comme actant syntaxique, il ne pourra pas passer à la lexicalisation. Car on souhaite une correspondance entre les traits naturels contenus dans le dictionnaire pour le sémantème en question. Si ses traits ne conviennent pas aux contraintes imposées au noeud, provenant du patron de régime de X, alors l'arbre sera incomplet. Lorsqu'on met les contraintes sur le noeud, on lui met aussi le trait *dsynt=constrained* pour montrer qu'il a été contraint et donc qu'il remplit les critères pour passer à la lexicalisation.

2.1.6. application des règles de lexicalisation

Il y a différentes règles de lexicalisation afin de donner un peu de latitude au système. Ainsi, si quelques informations sont manquantes dans le lexicon ou le semanticon, nous voulons que le système arrive quand même à effectuer une génération de texte. C'est pourquoi il existe la règle standard qui fonctionne lorsque nous avons accès à tous les éléments nécessaires. Les règles de types *guess* opèrent lorsqu'il nous manque certaines infos.

2.1.6.1. *lex_standard*

Nous avons vu cette règle un peu plus haut, car il fallait l'appliquer dès le début pour lexicaliser le noyau principal afin de chercher le bon lexème qui pouvait remplir cette fonction. Une fois que nous l'avons lexicalisé, nous avons pu aller chercher les informations sur la nature de son gp etc. Nous sommes maintenant rendus au moment où il existe des arcs syntaxiques au bout desquels nous avons des noeuds contraints par des traits comme : la dpos, la finitude, la définitude, etc. Nous allons donc lexicaliser ces noeuds afin de poursuivre la construction de l'arbre, du haut vers le bas. En gros, comment ça fonctionne. On cherche la lexicalisation d'un sémantème donné dans le semanticon, en allant chercher le trait *lex* du sémantème. Puis, on colle cette lexicalisation à un noeud déjà existant. Ce noeud existe déjà car il a soit été créé par *root_standard* ou par une des règles actanciellles. On va octroyer au noeud 3 traits : un trait *dlex*, un trait *dpos*, et finalement un trait *dsynt*. Le trait *dlex* est la lexicalisation profonde, celle qu'on retrouve dans le lexicon, le trait *dpos* est la partie du discours profonde qui doit correspondre à la dpos demandée par *constraints_gp*. Et finalement, un trait *dsynt=OK* qui s'ajoute à la lexicalisation du sémantème pour signifier au système que le noeud a été consommé et qu'il a été réalisé en syntaxe profonde. Donc, qu'on ne fasse plus d'opération sur ce noeud. Finalement, on peut uniquement faire des opérations sur des noeuds qui ont le trait *dsynt=constrained* afin de lexicaliser seulement les noeuds qui se sont fait attribués des contraintes. Ainsi, s'ils respectent les contraintes, ils pourront être lexicalisés, sinon, l'arbre sera incomplet et la génération échouera.

2.1.7. Avantages d'utiliser ces mécanismes

"*Dsynt=OK*" indique que le noeud a été consommé. Autrement dit, il existe maintenant une unité lexicale réalisé en syntaxe profonde, là où il y avait un noeud vide. "*Dsynt=constrained*" indique qu'un noeud vide s'est fait imposé des contraintes. Ces contraintes peuvent être de plusieurs ordres. Notamment, on impose une dpos, une finitude, un mood, etc. Il s'agit du passage intermédiaire entre la création d'un noeud et sa lexicalisation. On veut s'assurer qu'il respecte certaines contraintes pour ne pas lexicaliser n'importe quoi lors de la génération de notre arbre syntaxique. Finalement, le trait *dsynt=created*

2.2. ÉNUMÉRATIONS

Voici une énumération avec numérotation :

1. item 1 ;
2. item 2 ;
3. item 3.

Maintenant, une énumération sans numérotation avec des marqueurs différents :

- Marqueur par défaut ;
- \bullet ;
- ★ \star .

2.3. ÉQUATIONS MATHÉMATIQUES

Une équation :

$$\otimes^n \mathbb{C}^2 \cong \bigoplus_{m=-n/2}^{n/2} W_m.$$

Une autre équation, cette fois-ci numérotée :

$$\frac{\partial \mathcal{L}}{\partial \phi^a} - \partial_\mu \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi^a)} = 0, \quad \mu = 0, 1, 2, 3. \quad (2.3.1)$$

Les équations (2.3.1) précédentes sont appelées *équations d'Euler-Lagrange* ou encore *équations du mouvement*. Dans les calculs suivants,

$$\begin{aligned} \delta S &= \int_{\Omega} d^d x \mathcal{L}(\phi'^a(x), \partial_\mu \phi'^a(x)) - \int_{\Omega} d^d x \mathcal{L}(\phi^a(x), \partial_\mu \phi^a(x)) \\ &= \int_{\Omega} d^d x \left[\delta \phi^a \frac{\partial \mathcal{L}}{\partial \phi^a} + \partial_\mu \delta \phi^a \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi^a)} \right] \\ &= \int_{\Omega} d^d x \left[\left(\delta \phi^a \frac{\partial \mathcal{L}}{\partial \phi^a} + \partial_\mu \left(\delta \phi^a \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi^a)} \right) \right) - \delta \phi^a \partial_\mu \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi^a)} \right] \\ &= 0, \end{aligned}$$

aucune ligne n'est numérotée. Alors que dans ce qui suit, la dernière ligne l'est :

$$\begin{aligned} \delta S &= \int_{\Omega'} d^d x' \mathcal{L}(\phi'^a(x'), \partial'_\mu \phi'^a(x')) - \int_{\Omega} d^d x \mathcal{L}(\phi^a(x), \partial_\mu \phi^a(x)) \\ &= \int_{\Omega} d^d x \left[\bar{\delta} \phi^a \frac{\partial \mathcal{L}}{\partial \phi^a} + \partial_\mu \bar{\delta} \phi^a \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi^a)} \right] + \int_{\partial \Omega} d\sigma_\mu \mathcal{L}(\phi^a, \partial_\mu \phi^a) \delta x^\mu \\ &= \int_{\Omega} d^d x \partial_\mu \mathcal{J}^\mu(x). \end{aligned} \quad (2.3.2)$$

2.4. DÉFINITIONS, THÉORÈMES ET PREUVES

Voici une définition.

Définition 2.4.1 (La définition). *La définition.*

Voici un théorème.

Théorème 2.4.1 (Titre). *Ceci est vrai !*

DÉMONSTRATION. Voici la preuve. □

Démonstration. Voici la preuve en gras. □

2.5. CONSTRUCTION D'UN TABLEAU

TABLEAU 2. I. Un tableau simple dans le second chapitre.

Option	g	c	d	p{0.4\textwidth}
Effet	À gauche	Au centre	À droite	Le texte de cette colonne est justifié et la largeur de la colonne est fixée à 40 % de la zone de texte (hors tableau).

Le tableau 2. I n'est pas très garni.

2.6. RÉFÉRENCE À UNE ENTRÉE BIBLIOGRAPHIQUE

Les documents par Lamport [3], Goossens, Mittelbach et Samarin [1] ainsi que Spivak [5] sont des références en matière de L^AT_EX. Le manuel par Goossens *et al.* [1] est probablement le plus populaire du lot.

L'article de Martin [4] est, manifestement, très riche en rebondissements.

Les entrées du fichier `.bib` qui ne sont pas référencées dans le texte ne sont pas ajoutées à la bibliographie : un avantage de plus en faveur de BibT_EX.

Dans ce paragraphe, on teste une cette référence Hastie [2].

2.7. INSERTION DE FIGURES

La figure 2.1 est un *cercle*. À la figure 2.2, le triangle (a) et le carré (b) ont été placés côtes-à-côtes grâce à la commande `\subfigure`.

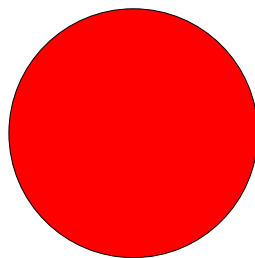
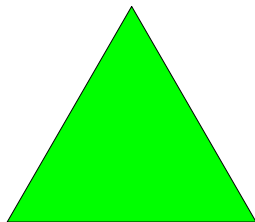
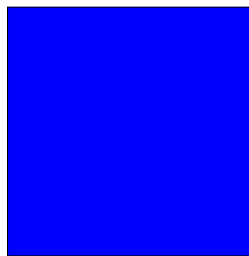


FIGURE 2.1. Un cercle.



(a) Un triangle.



(b) Un carré.

FIGURE 2.2. Un carré et un triangle.

Bibliographie

- [1] Goossens, M., F. Mittelbach et A. Samarin. 1994, «The L^AT_EX companion», New-York.
- [2] Hastie, D. 2005, «Towards automatic reversible jump markov chain monte carlo», thèse de doctorat, University of Bristol.
- [3] Lamport, L. 1986, «L^AT_EX – a document preparation system», Reading.
- [4] Martin, P. 1992, «On Schur-Weyl duality, A_n Hecke algebras and quantum $\mathfrak{sl}(N)$ on $\otimes^{n+1}\mathbb{C}^N$ », *Int. J. Mod. Phys. A*, vol. 7, p. 645–673.
- [5] Spivak, M. D. 1990, «The joy of T_EX», Providence, 2^e éd..

Annexe A

LE TITRE

A.1. SECTION UN DE L'ANNEXE A

La première annexe du document.

Pour plus de renseignements, consultez le site [web de la FESP](#). Pour plus de renseigne-

TABLEAU A. I. Liste des parties

Les couvertures conformes	obligatoires
Les pages de garde	obligatoires
La page de titre	obligatoire
Le résumé en français et les mots clés français	obligatoires
Le résumé en anglais et les mots clés anglais	obligatoires
Le résumé de vulgarisation	facultatif
La table des matières, la liste des tableaux, la liste des figures	obligatoires
La liste des sigles, la liste des abréviations	obligatoires
La dédicace	facultative
Les remerciements	facultatifs
L'avant-propos	facultatif
Le corps de l'ouvrage	obligatoire
L'index analytique	facultatif
Les sources documentaires	obligatoires
Les appendices (annexes)	facultatifs
Le curriculum vitæ	facultatif
Les documents spéciaux	facultatifs

ments, consultez le site [web de la FESP](#).

Annexe B

LE TITRE2

Annexe C

LE TITRE3

Annexe D

LE TITRE4

