Annotation en rôles sémantiques du français en domaine spécifique

Quentin Pradet

Université Paris Diderot (Paris 7)
École Doctorale de Sciences Mathématiques de Paris-Centre n° 386

Doctorat d'Informatique

Quentin PRADET

Annotation en rôles sémantiques du français en domaine spécifique

Thèse sous la direction de :

Laurence DANLOS et Gaël DE CHALENDAR

Soutenue publiquement le 6 février 2015
devant le jury composé de :

M. Guy LAPALME, rapporteur
M. Patrick SAINT-DIZIER, rapporteur
M^me^ Laurence DANLOS, directrice
M. Gaël DE CHALENDAR, co-directeur
M^me^ Brigitte GRAU, examinatrice

Résumé

Cette thèse de Traitement Automatique des Langues a pour objectif l'annotation automatique en rôles sémantiques du français en domaine spécifique. Cette tâche consiste à la fois à désambiguïser le sens des verbes d'un texte et à annoter leurs syntagmes avec des rôles sémantiques tels qu'Agent, Patient, ou Destination. Elle aide de nombreuses applications dans les domaines où des corpus annotés existent : on peut alors entraîner des algorithmes supervisés performants. Nous cherchons au contraire à annoter des domaines ne disposant pas de tels corpus annotés. Nous considérons ici trois domaines : le réchauffement climatique, Informatique/Internet, et le football, leurs corpus annotés ne nous servant que pour l'évaluation. Nous montrons que nos traductions vers le français de lexiques sémantiques pour l'anglais donnent la possibilité d'annoter en rôles sémantiques des textes aussi bien en domaine général qu'en domaine spécifique sans avoir à entraîner un modèle statistique.

Nos travaux portent sur deux grands axes : les ressources puis les méthodes servant à l'annotation en rôles sémantiques.

Concernant les ressources, nous commençons par traduire la base de données lexicales WordNet vers le français à l'aide d'un modèle de langue syntaxique issu du web. Cette ressource, WoNeF, est disponible en trois versions : une à haute précision (93,3 %), une à haut F-score (70,9 %), et l'autre à haute couverture, plus large mais plus bruitée. Nous traduisons ensuite le lexique VerbNet dans lequel les verbes sont regroupés suivant leur traits syntaxiques, morphologiques et sémantiques. La traduction, nommée Verb∋Net, a été obtenue à la fois en réutilisant au maximum les lexiques verbaux du français (le Lexique-Grammaire et Les Verbes Français) mais aussi avec un travail manuel important pour contrôler au mieux son contenu.

Concernant les méthodes, nous commençons par évaluer notre méthode basée sur VerbNet sur le corpus annoté FrameNet en suivant les travaux de Swier and Stevenson [2005]. Nous montrons que des améliorations conséquentes peuvent être obtenues à la fois d'un point de vue syntaxique avec la prise en compte de la voix passive et d'un point de vue sémantique en filtrant les syntagmes ne correspondant pas aux restrictions de sélection indiquées dans VerbNet et en réutilisant les résultats des premières annotations automatiques non ambigües.

Enfin, une fois ces briques en place, nous évaluons la faisabilité de l'annotation en rôles sémantiques du français dans nos trois domaines spécifiques. Nous évaluons en effet quels sont les avantages et inconvénients de se baser sur VerbNet et Verb∋Net pour annoter ces domaines en anglais et en français.

Table des matières

Première partie

L'annotation en rôles sémantiques

1. Introduction

1.1. Motivation

1.1.1. Historique

En Intelligence Artificielle et en Traitement Automatique des Langues, les années 50 et 60 étaient pleines d'optimisme. D'après [Russell and Norvig, 2010], Simon Herbert annonçait en 1957 :

> Mon intention n'est pas de vous surprendre ou de vous choquer, mais la manière la plus simple de résumer les choses consiste à dire qu'il existe désormais des machines capables de penser, d'apprendre et de créer. En outre, leur capacité d'accomplir ces choses va rapidement s'accroître jusqu'à que, dans un futur proche, le champ des problèmes qu'elles pourront aborder soit coextensif à celui auquel s'applique l'esprit humain.

Effectivement, les réussites sur des petits problèmes étaient prometteuses. Slagle [1963] a proposé un système de calcul de primitives du niveau d'un bon étudiant de première année à l'université. Winograd [1972] a lui proposé un système de compréhension de l'anglais au sein du monde des blocs, un micromonde très utilisé à l'époque pour sa simplicité. Malheureusement, la réussite sur des petits problèmes ne s'est pas étendue à des problèmes plus complexes, ce qui a conduit notamment à un arrêt des financements portant sur la traduction automatique aux États-Unis [Pierce and Carroll, 1966] et à limiter les travaux en Intelligence Artificielle à deux universités en Grande Bretagne [Lighthill et al., 1973].

Depuis, l'Intelligence Artificielle a continué à progresser jusqu'à devenir une industrie et une science, grâce à :

- des techniques comme les systèmes experts, les réseaux de neurones et diverses approches d'apprentissage automatique,
- de gros volumes de données disponibles depuis le début des années 2000,
- des ordinateurs de plus en plus puissants disposant de stockages de plus en plus rapides,
- et à diverses applications tel que la planification logistique, la reconnaissance de la parole ou encore la robotique.

De la même manière, au fil des années, le Traitement Automatique des Langues a muri, et s'appuie aujourd'hui sur des applications, des méthodes et des sous-tâches plus accessibles que les applications envisagées initialement. Nous citerons ici deux de ces sous-tâches.

— **Étiquetage morpho-syntaxique** Le Brown Corpus a été annoté en parties du discours entre le milieu des années 60 et la fin des années 1970, ce qui a permis d'entraîner divers algorithmes, tels que les chaînes de Markov cachées et plus tard des méthodes d'apprentissage supervisées telles que les SVMs ou les CRFs. Un plateau a été atteint autour de 97 % d'exactitude depuis le milieu des années 2000 [Manning, 2011].
— **Analyse syntaxique** Au début des années 1990, le corpus du Penn Treebank [Marcus et al., 1993] a permis d'avancer la recherche en analyse syntaxique. Deux représentations relativement équivalentes (constituants et dépendances) se sont largement imposées [Rambow, 2010], ce qui a facilité la comparaison des systèmes. Différents chercheurs ont introduit un certain nombre d'algorithmes ayant chacun leurs avantages et leurs défauts. Depuis le début des années 2010, et de la même manière que pour l'annotation des parties du discours, un plateau a été atteint autour de 90 % d'exactitude, et ceci que la méthode soit statistique ou plus symbolique [De La Clergerie, 2014].

Pour un certain nombre de chercheurs [Bos et al., 2012, Banarescu et al., 2013], c'est le moment de se tourner vers de nouvelles tâches plus sémantiques. C'est pour cette raison qu'à la manière des corpus annotés en parties du discours ou en syntaxe qui ont tant fait progresser leurs domaines respectifs, des corpus « sémantiques » ont vu le jour dans le passé tels que FrameNet, PropBank ou le Penn Discourse Treebank, mais d'autres, plus ambitieux, voient aussi le jour aujourd'hui, tels que GMB [Bos et al., 2012] ou l'AMR Bank [Banarescu et al., 2013] (section 2.3.2). L'objectif affiché est de « faire progresser la sémantique comme la syntaxe a progressé dans les années 1990 ».

1.1.2. Au-delà de l'analyse syntaxique

Dès lors, si l'on considère que l'analyse syntaxique n'est plus la priorité, quelle direction prendre ? Commençons par identifier les informations manquantes une fois que l'analyse syntaxique d'une phrase a été effectuée.

Le problème principal que nous voyons est que le sujet et les objets syntaxiques d'un verbe ne suffisent pas à déterminer les sujets et objets sémantiques, c'est-à-dire l'agent, le patient, etc. Par exemple, étant donné la phrase *Le ballon repoussé par Léa a cassé la vitre des voisins*, il s'avère que le sujet syntaxique (*Le ballon repoussé par Léa*) correspond parfaitement à l'agent sémantique. Dans d'autres situations, ce n'est pas le cas : pour *La vitre des voisins a cassé sous le choc du ballon tiré par Léo*, le sujet syntaxique est *La vitre des voisins* [1]. Pourtant ce sujet

1. Si la voix passive avait été employée, comme dans *La vitre des voisins a été cassée par le choc du ballon tiré par Léo*, alors le sujet syntaxique aurait bien été *le choc du ballon tiré par Léo*, mais ce n'est pas le cas ici.

syntaxique n'est pas l'agent sémantique, mais bien le patient, étant donné que c'est la vitre qui subit l'action ici.

De manière plus marquée que pour les sujets, l'analyse syntaxique en tant que telle ne fournit pas suffisamment d'information pour désambiguïser le rôle des objets du verbe. Prenons les phrases *Luc a posé un livre sur la table* et *Luc a posé sur la table son livre préféré traitant de la génétique des chimpanzés*. Ici, l'ordre des objets ne suffit pas, il faut identifier que parmi les deux objets syntaxiques :

— l'un est un syntagme prépositionnel introduit par une préposition locative (*sur*),
— tandis que l'autre est un syntagme nominal direct.

On peut alors déterminer que le livre est le thème sémantique pour le prédicat *poser* et que la table est la destination sémantique pour ce même prédicat. Ici, même si la syntaxe en elle-même ne résout pas le problème, c'est bien grâce à elle qu'on peut déterminer le rôle de chaque syntagme.

Dans d'autres cas, la syntaxe ne suffit plus. Par exemple, pour la phrase *When you've booted the machine you've built yourself* (extraite de la version anglaise de DiCoInfo), le sujet et l'objet de *boot* sont tous les deux des syntagmes nominaux, et la syntaxe ne suffit alors pas à désambiguïser entre le sens informatique de *boot* (démarrer un ordinateur) et le sens géographique de *boot* (exclure un individu de quelque part). Ici, des informations sémantiques peuvent nous aider. Dans le cas informatique, l'objet n'est pas animé, alors que dans le second il l'est. Si on sait que *the machine* n'est pas animé, il devient alors possible :

1. d'exclure le sens géographique du verbe,
2. et d'attribuer les rôles corrects aux syntagmes associés aux verbes (respectivement thème et destination),

Ces deux informations sont toutes les deux importantes : c'est bien le sens du verbe qui permet d'interpréter les rôles sémantiques attribués aux syntagmes dans une application.

Cette étape d'analyse au-delà de l'analyse syntaxique s'appelle l'annotation en rôles sémantiques.

1.2. Objectifs

1.2.1. L'annotation en rôles sémantiques

L'annotation en rôles sémantiques répond à la question « Qui a fait Quoi à Qui, Comment, Où et Quand ? ». Prenons pour exemple la phrase *Mrs. Aouda essaya vainement de retenir Mr. Fogg* (extrait du *Tour du monde en quatre-vingts jours* de Jules Verne). En considérant pour exemple le cadre de FrameNet (section 1.3.4), une annotation en rôles sémantiques du prédicat *essayer* déterminera que cette utilisation du verbe correspond à une situation de Tentative, puis identifiera parmi les syntagmes liés aux verbes quel est l'Agent, l'Activité tentée, et le Résultat. La Figure 1.1 montre le résultat de l'annotation.

[Agent]	**Tentative**	[Résultat]	[Activité]
Mrs. Aouda	**essaya**	vainement	de retenir Mr. Fogg.

FIGURE 1.1. – Le verbe *essayer* déclenche la situation *Tentative*. Les différents syntagmes liés au verbe jouent chacun un rôle sémantique ici, mais ce n'est pas toujours le cas. Par exemple, si la phrase précisait *après le dîner*, ce syntagme aurait été un complément sans rôle sémantique associé.

Différentes informations sont disponibles après l'annotation en rôles sémantiques :

— Le prédicat ayant déclenché la situation est identifié. Dans la Figure 1.1, c'est un verbe. D'autres parties du discours peuvent déclencher une frame, mais nous nous concentrons dans ce travail essentiellement sur les verbes.
— La frame est identifiée, ici *Tentative*.
— Enfin, les rôles exprimés sont annotés. Par exemple, *Mrs. Aouda* est l'Agent.

1.2.2. Applications

Selon Gildea and Jurafsky [2002], l'annotation en rôles sémantiques est historiquement une évolution naturelle de certains travaux sur l'extraction d'information où les systèmes traitent des situations très spécifiques, par exemple la détection de résultats d'évènements sportifs ou la détection d'acquisitions d'entreprises dans des corpus journalistiques. À chaque nouveau système d'extraction d'information dans un domaine différent, il est nécessaire de redéfinir les différents patrons sémantiques et d'entraîner un nouveau système sur de nouvelles données, et l'annotation en rôles sémantiques est vue comme une solution à ce problème. C'est dans cette optique que Gildea and Jurafsky s'appuient sur le corpus FrameNet et présentent le premier système d'annotation en rôles sémantiques.

Aujourd'hui, les systèmes d'annotation en rôles sémantiques n'ont pas remplacé les systèmes

d'extraction d'information. Une des raisons est que les difficultés sont différentes [Boros et al., 2014] :

— un système d'extraction d'information pourra utiliser plusieurs phrases pour remplir un évènement alors que l'annotation en rôles sémantique annote encore les phrases indépendamment,
— l'extraction d'informations se concentre sur un petit nombre d'évènements et d'étiquettes au sein d'un même évènement alors que les systèmes d'annotations en rôles sémantiques doivent traiter un grand nombre de situations différentes, chacune ayant ses propres étiquettes.

L'annotation en rôles sémantiques a, par contre, été utilisée dans un grand nombre d'autres applications, notamment les systèmes de questions-réponses [Shen and Lapata, 2007], l'extraction d'évènements [Exner and Nugues, 2011], la fouille d'opinion [Das et al., 2012] ou la traduction automatique [Bazrafshan and Gildea, 2013, 2014]. Un des intérêts de la généralité de l'annotation en rôles sémantiques est de s'adapter facilement à de nouvelles tâches. Ainsi, l'annotation en rôles sémantiques à aussi été utilisée sur des tâches peut-être moins classiques : l'évaluation de la traduction automatique [Lo and Wu, 2011, Chuchunkov et al., 2014], la détection de plagiat [Osman et al., 2012], la prédiction des cours de bourse [Xie et al., 2013], la génération de scènes 3D [Chang et al., 2014], la recommandation fondée sur le contenu [De Clercq et al., 2014], la détection de comparaisons de produits [Kessler and Kuhn, 2013] ou l'interprétation de recettes de cuisine [Malmaud et al., 2014].

1.2.3. Contraintes

Nous souhaitons que notre système d'annotation en rôles sémantiques puisse être utilisé dans un environnement industriel dans lequel d'une part cette annotation en rôles sémantiques fournit des informations utiles au développement de tâches applicatives et d'autre part les domaines à couvrir ne sont pas connus à l'avance. Les contraintes suivantes découlent de ces prérequis.

Cadre ouvert Se contenter de certaines situations dans un domaine fermé n'est pas satisfaisant. Les inventaires de sens utilisés doivent couvrir au maximum les différents sens des mots d'une langue.

Langue française Le français dispose encore d'un nombre limité de ressources sémantiques en cadre ouvert, même si cet écart est en train de se combler rapidement. En effet, au-delà des ressources développées dans les années 1970 que nous présenterons à la section 1.3.6, des progrès récents laissent entrevoir un futur brillant. Le projet ANR ASFALDA [Candito et al., 2014] produit un FrameNet annoté du français de grande qualité mais avec une couverture encore limitée, et WOLF [Sagot and Fišer, 2008b] et WoNeF (Chapitre 3) sont des traductions automatiques

de WordNet [Fellbaum, 1998] qui ont progressé à la fois en précision et en couverture au fil des ans. Pour rester dans un cadre ouvert, le système présenté doit pouvoir se contenter de telles transpositions automatiques de ressources anglaises vers le français. [2]. Avec cette approche de mutualisation des ressources au niveau de la langue, chaque nouvelle utilisation d'une de nos ressources est l'occasion de l'améliorer à la fois pour utilisation immédiate mais aussi pour les utilisations futures.

Simplicité Nous voulons que notre système soit très simple à mettre en place et qu'il soit tout aussi facile de corriger quelques erreurs spécifiques, même au prix d'une performance moins bonne que des approches plus complexes dans le cas général. La stratégie que nous adoptons est de simplifier nos systèmes afin que toute intervention manuelle sur les ressources soit possible, puis de les améliorer une fois qu'ils ont montré leurs limites.

Efficacité Cette contrainte est moins forte que les autres, mais reste nécessaire pour que les systèmes présentées puissent être utilisées à large échelle. L'annotation en rôles sémantiques est un problème difficile de classification automatique et certains systèmes ont des temps d'entraînement et d'exécution trop longs pour l'utilisation que nous voulons en faire ici.

Libre diffusion Enfin, il est important que les outils et ressources soient au maximum libre d'accès afin que d'autres puissent les utiliser et les étudier. Pour cette raison, la grande majorité du code source écrit pendant cette thèse est disponible sur GitHub : en particulier, le système d'annotation en rôles sémantiques est disponible sur `https://github.com/aymara/knowledgesrl`. Il est utilisable indépendamment mais a aussi été intégré dans l'analyseur linguistique libre LIMA (`http://aymara.github.io/lima/`) par Clémence Filmont.

Ces contraintes seront utilisées pour évaluer à la fois l'état de l'art et les approches présentées.

1.2.4. Moyens

Le Traitement Automatique des Langues requiert des lexiques et de larges quantités de données annotées pour analyser efficacement des textes dans le domaine général. Obtenir cette quantité de données est un problème en soi connu sous le nom de "knowledge acquisition bottleneck" en désambiguïsation lexicale [Gale et al., 1992, Navigli, 2009]. Le problème se pose aussi pour l'annotation en rôles sémantiques où la quantité de données annotées est limitée [Das et al., 2012, section 1]. Il est possible de résoudre ce problème domaine par domaine en annotant de grandes quantités de données pour chaque domaine, mais d'autres stratégies sont nécessaires

2. L'exception ici est VerbNet, ressource traduite « semi-manuellement » vers le français (Chapitre 4) mais qui n'est encore ni finalisée ni homogénéisée.

pour mieux généraliser et atteindre nos objectifs dans un grand nombre de domaines. Une possibilité est d'utiliser au mieux les données annotées en perfectionnant les algorithmes existants, une autre est d'utiliser intelligemment les données non annotées qui existent en quantité bien plus importante. Une troisième possibilité, celle que nous choisissons d'explorer ici, est d'exploiter des lexiques couvrant l'interface syntaxe-sémantique sur une large partie du vocabulaire. C'est ce qui est fait dans VerbNet où les traits syntaxiques et sémantiques partagés par les mêmes verbes sont explicitement notés, ce qui permet à chaque modification dans VerbNet d'améliorer le traitement de plusieurs verbes au lieu d'un seul.

Deux difficultés majeures qu'affrontent les créateurs de lexique sont la granularité de sens et la distinction des sens. Ces deux difficultés sont traitées par les classes de Levin [Levin, 1993] qui sont à l'origine de VerbNet. Dans ces classes, les verbes sont classifiés principalement à travers leur alternances syntaxiques, ce qui fournit un critère qui est à la fois facilement observable et qui produit des distinctions sémantiques intéressantes (section 1.3.2) validées par des expériences empiriques impliquant un grand nombre d'annotateurs [Hartshorne et al., 2014]. VerbNet [Kipper-Schuler, 2005], basé sur les classes de Levin, encode non plus les alternances mais les cadres de sous-catégorisation valables pour chaque classe, et rajoute des informations de rôle et de sémantique à travers une logique des prédicats simplifiée. De nouvelles classes, constructions et verbes ont étés ajoutés à VerbNet au fil des ans. Au-delà de son encodage efficace, VerbNet est un lexique adapté à la tâche d'annotation en rôles sémantiques : on peut utiliser un cadre de sous-catégorisation pour associer des syntagmes à des rôles thématiques [Swier and Stevenson, 2005, Pradet et al., 2013b]. Grâce à sa couverture élevée (plus de quatre mille verbes distincts) et son groupement de verbes utile, VerbNet est bien adapté à l'annotation en rôles sémantiques.

WordNet [Fellbaum, 1998] est une autre ressource qui complète VerbNet d'au moins deux façons.

— D'une part, WordNet peut être utilisé pour désambiguïser le sens des verbes et donc aider à choisir la bonne classe VerbNet, ce qui est une méthode qui a fait ses preuves en désambiguïsation lexicale [Agirre and Soroa, 2009].
— D'autre part, comme indiqué à la section 1.1.2, certaines propriétés sémantiques des syntagmes (animé, humain, organisation, etc.) sont utiles à l'annotation en rôles sémantiques, et la hiérarchie offerte par WordNet est un moyen pour déterminer ces propriétés (section 5.4.1), malgré le résultat négatif que nous avons observé dans notre implémentation.

Plus généralement, là où VerbNet nous apporte des informations plutôt syntaxiques sur les verbes, WordNet est une source plus globale qui peut apporter des informations utiles sur le reste des mots de la phrase.

Cependant, un VerbNet et un WordNet validés manuellement et à large couverture n'existent pour le moment que pour l'anglais [3]. De telles ressources seraient pourtant encore plus utiles

3. L'exception est plWordnet [Maziarz et al., 2012], un wordnet du polonais de grande taille validé manuellement

pour les langues moins dotées où les corpus annotés en rôles sémantiques n'existent pas. VerbNet a un potentiel inter-linguistique, visible notamment avec le portuguais [Kipper-Schuler, 2005, section 2.2.2]. Adapter VerbNet vers une nouvelle langue suffisamment proche de l'anglais permet de conserver sa structure, ainsi que l'information sémantique et les rôles thématiques, ce qui donne la possibilité de produire un lexique utile en économisant beaucoup de travail manuel.

Une fois que ces ressources ont été traduites vers le français, il faut les utiliser pour réaliser la tâche d'annotation en rôles sémantiques. Cette thèse fournit les *ressources* nécessaires en traduisant WordNet et VerbNet (Partie II) et les *méthodes* (Partie III) répondant à ces objectifs.

1.3. Ressources lexicales utilisées

1.3.1. WordNet

La première ressource lexicale à tirer parti de la possibilité de représenter le lexique sous la forme d'un graphe est WordNet [Fellbaum, 1998]. Son élaboration a commencé en 1985 [Miller et al., 1990]. Établi sur des principes psycholinguistiques, WordNet propose quatre graphes pour les quatre parties du discours formant une classe ouverte : noms, verbes, adjectifs et adverbes. Les nœuds du graphe sont des ensembles de synonymes (*synonym sets* ou *synsets*). Un synset regroupe plusieurs mots, une définition, et potentiellement des exemples.

Chaque synset est lié à d'autres synsets à travers un certain nombre de relations telles que l'hypéronymie, la méronymie de partie (*guidon* est un méronyme de partie de *vélo*), l'antonymie, etc. Si on ne considère que l'hypéronymie, WordNet peut être visualisé comme un arbre (Figure 1.2). En considérant les autres relations, WordNet est un graphe (Figure 1.3).

Les sens proposés par WordNet ont étés utilisés pour annoter différents corpus [Petrolito and Bond, 2014], ce qui a permis d'entraîner des systèmes supervisés. WordNet est rapidement devenu le standard de la désambiguïsation lexicale et a été utilisé dans de nombreuses campagnes d'évaluation [Navigli, 2009]. C'est aussi une ressource très utilisée pour beaucoup d'applications : au moment de l'écriture de ce manuscrit, ses 10 000 citations sur Google Scholar sont le meilleur moyen de l'attester.

1.3.2. Les classes de Levin

Les classes de Levin [Levin, 1993] sont une classification des verbes anglais établie suivant un principe simple : le comportement syntaxique des verbes détermine en partie leur sens. Après

et toujours développé activement

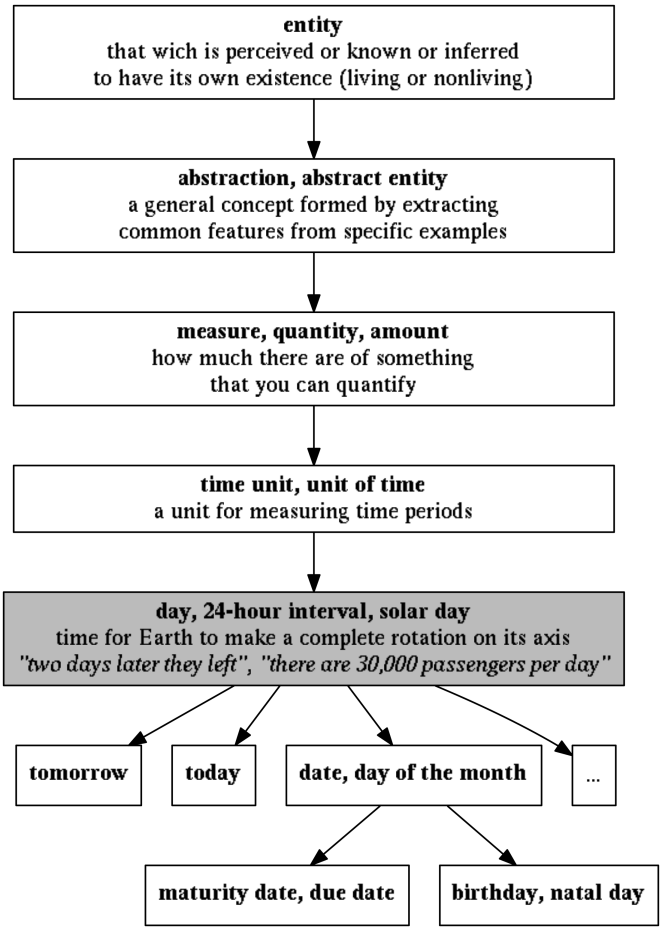


FIGURE 1.2. – Hypéronymie dans WordNet autour du synset *day*. Les synsets au-dessus de *day* sont ses hypéronymes (*day* est-un *time unit*), et les synsets au-dessus font partie de ses hyponymes (*tomorrow* est-un *day*).

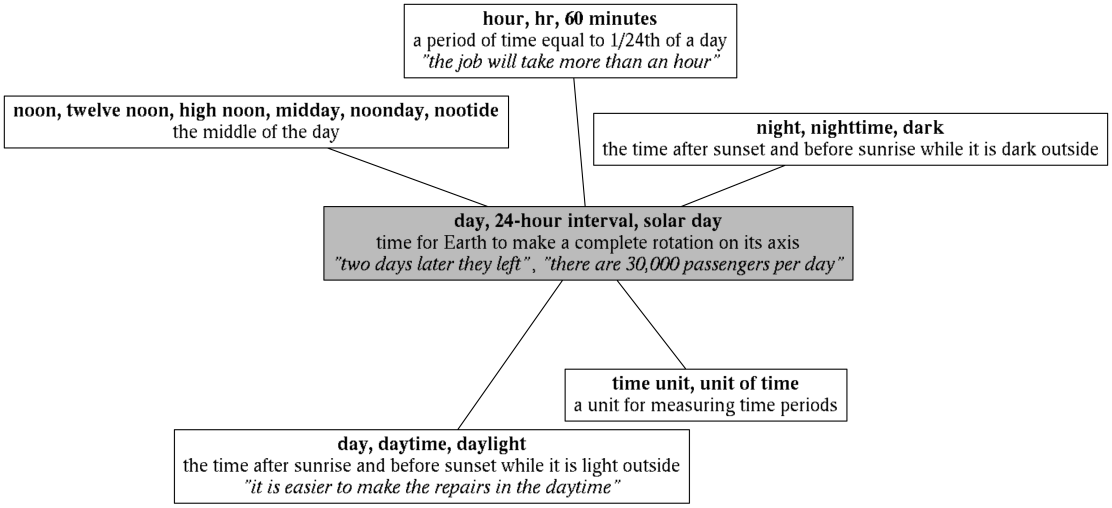


FIGURE 1.3. – Le synset *day* est aussi lié à d'autres synsets si on considère d'autres relations que l'hypéronymie et l'hyponymie.

avoir défini un certain nombre d'alternances de diathèses possibles, les verbes sont classés en groupes partageant les mêmes alternances.

Par exemple, les *fill verbs* tels que *staff*, *coat* ou encore *pollute* [Levin, 1993, p. 119] acceptent notamment la *locatum subject alternation* [Levin, 1993, p. 85], donc ces deux constructions sont possibles :

— *Leslie staffed the store with employees*
— *The employees staffed the store*

mais n'acceptent pas par exemple, la *conative alternation* [Levin, 1993, p. 41], donc seule la première construction est possible :

— *Leslie staffed the store with employees*
— * *The store staffed with employees*

Ainsi, pour chaque groupe de verbes, différents critères précis permettent de distinguer ces verbes d'autres qui n'ont pas les mêmes propriétés.

Malgré quelques critiques [Riemer, 2011] sur le principe, cette classification est très pertinente pour l'annotation en rôles sémantiques :

— Une grande majorité des verbes anglais sont couverts, rendant la ressource utile pour des annotations à large échelle.
— La classification est hiérarchique et regroupe de nombreux verbes : avec une cinquantaine

de classes principales, des généralisations et partages d'informations entre les verbes de classes identiques ou proches sont possibles.
— Les comportements syntaxiques déterminent les comportements sémantiques, ce qui correspond au schéma classique de l'annotation en rôles sémantiques qui s'appuie avant tout sur une analyse syntaxique.

Dans les classes de Levin, toute distinction de classe doit s'appuyer sur un critère clairement observable, tel que le comportement syntaxique ou les propriétés morphologiques d'un verbe. Ainsi, même dans les cas où l'on veut identifier une différence sémantique provenant d'une intuition linguistique, il faut établir des critères clairs et précis permettant d'assurer la cohérence et la robustesse de l'ensemble.

Pour cette raison, les groupements de verbes sont parfois grossiers d'un point de vue sémantique. Ainsi, dans les *put verbs* se trouve la classe *Pocket* qui regroupe les verbes "mettre dans sa poche" (*put*) et "mettre en prison" (*jail*). Cependant, le comportement syntaxique est le même : le regroupement est logique et les différences de sens ne sont a priori pas gênantes pour des tâches telles que l'annotation en rôles sémantiques qui ne se basent pas sur le sens précis des verbes. Si ce manque de finesse est gênant, il est possible d'affiner automatiquement les classes de Levin en réalisant des intersections de classes : les verbes obtenus seront alors définis plus strictement [Dang et al., 1998]. Cette possibilité n'est cependant pas utilisée dans nos travaux.

1.3.3. VerbNet

VerbNet [Kipper-Schuler, 2005] est une version électronique des classes de Levin qui ont été améliorées sur plusieurs fronts avec :

— de nouvelles classes provenant de Korhonen and Briscoe [2004] intégrant les verbes acceptant des complétives, mais aussi des syntagmes adjectivaux et adverbiaux ou encore des particules,
— de nouveaux verbes provenant de Dorr et al. [2001],
— la liaisons des verbes à WordNet, OntoNotes, PropBank et FrameNet dans le cadre du projet SemLink [Palmer, 2009],
— de nombreuses corrections au fil des versions, une des améliorations prévues dans le futur étant d'ajouter d'autres verbes via l'étude de large corpus [Bonial et al., 2013].

Ces améliorations ont à la fois contribué à VerbNet en largeur (nouvelles classes) et en profondeur (nouveaux verbes, nouvelles constructions syntaxiques). La base de données continue d'évoluer, la dernière version au moment de l'écriture de ce manuscrit étant la 3.2. Malheureusement, la version 3.2 évolue sans que soient marqués clairement les sous-versions : suivant le jour de téléchargement de VerbNet, nous pouvons avoir à disposition la version 3.2.1, 3.2.2 ou 3.2.3, sans pouvoir clairement distinguer ces trois versions. De plus, nous avons du modifier la ressource pour corriger certaines incohérences que nous espérons voir incluses

pour-9.5
Members: 8, Frames: 5

Class Hierarchy
pour-9.5*
no subclasses

Members KEY

dribble (fn 1; wn 1, 2; g 1, 2)	spew (fn 1, 2; wn 1, 2, 3; g 1, 2)	pour (fn 1; wn 1, 3, 4; g 1, 2)	trickle (wn 1)
drip (fn 1, 2; wn 1, 2; g 1)	spill (fn 1; wn 1, 2, 3; g 1)	slop (wn 1)	slosh (wn 3)

Roles REF

- Agent [+animate]
- Theme [+substance | [+concrete & +plural]]
- Destination [+location & -region]
- Initial_Location [+location & -region]

Frames REF KEY

NP V NP PP.destination
- **example** "Tamara poured water into the bowl."
- **syntax** Agent **V** Theme {{+path & -dest_dir}} Destination
- **semantics** motion(during(E), Theme) not(Prep(start(E), Theme, Destination))
  Prep(E, Theme, Destination) cause(Agent, E)

NP V NP ADVP
- **example** "Tamara poured water here."
- **syntax** Agent **V** Theme Destination <+adv_loc>
- **semantics** motion(during(E), Theme) not(Prep(start(E), Theme, Destination))
  Prep(E, Theme, Destination) cause(Agent, E)

NP V PP.destination
- **example** "Water poured onto the plants."
- **syntax** Theme **V** {{+path & -dest_dir}} Destination
- **semantics** motion(during(E), Theme) not(Prep(start(E), Theme, Destination))
  Prep(E, Theme, Destination)

NP V NP PP.initial_location PP.destination
- **example** "Maria poured water from the bowl into the cup."
- **syntax** Agent **V** Theme {{+src}} Initial_Location {{+dest_conf}} Destination
- **semantics** not(Prep(start(E), Theme, Destination)) Prep(E, Theme, Initial_Location)
  Prep(E, Theme, Destination) cause(Agent, E)

NP V PP.initial_location PP.destination
- **example** "Water poured from the bowl into the cup."
- **syntax** Theme **V** {{+src}} Initial_Location {{+dest_conf}} Destination
- **semantics** not(Prep(start(E), Theme, Destination)) Prep(E, Theme, Initial_Location)
  Prep(E, Theme, Destination)

FIGURE 1.4. – La classe `pour-9.5` dans VerbNet. Huit membres sont listés avec des mappings vers FrameNet (FN), WordNet (WN) et les groupements OntoNotes (G). Les rôles sont associés à des restrictions de sélections : ici l'Agent est toujours animé. Les frames listent les constructions syntaxiques valides, avec un exemple et une interprétation sémantique. Cette classe n'a pas de sous-classes, mais si elle en avait, les frames de cette classes seraient valides pour les verbes de la sous-classe.

dans la prochaine version de VerbNet. Pour cette raison, nous rendons disponibles la version de VerbNet que nous utilisons dans l'ensemble de ce travail à l'adresse suivante : `https://github.com/aymara/verbnet/archive/pradet-thesis.zip`.

VerbNet contient 3769 lemmes, 5257 entrées réparties en 500 sous-classes dont 270 classes de niveau 2 et 109 classes de niveau 1. Pour chaque sous-classe, ce lexique indique :

— la liste des verbes de la classe,
— les rôles thématiques en jeu ainsi que leur restrictions de sélection,
— la liste des *frames* VerbNet,
— les éventuelles sous-classes.

Une frame inclut une phrase d'exemple, une formule syntaxique donnant la liaison entre les syntagmes et les rôles thématiques, une formule sémantique basée sur la logique des prédicats explicitant la relation entre les participants et les évènements. La figure 1.4 montre ces différentes informations telles qu'elles sont listées sur le site web de VerbNet pour la classe `pour-9.5` (disponible sur `http://verbs.colorado.edu/verb-index/index.php`).

VerbNet a montré la cohérence de sa classification et est très utilisé, notamment pour l'annotation en rôles sémantiques [Swier and Stevenson, 2005, Palmer et al., 2013] où il présente l'intérêt de ne pas être restreint à un domaine spécifique tout en couvrant une large partie des occurrences des verbes anglais dans un texte donné.

1.3.4. FrameNet

FrameNet [Baker et al., 1998] repose sur la théorie des *Frame Semantics*, élaborée par Fillmore [1982] en modifiant sa théorie des cas [Fillmore, 1968]. Ici, les rôles sémantiques sont spécifiques à chaque situation. Par exemple, la situation *Infecting* définit les rôles sémantiques *Infected_entity*, *Infection* et *Infection_cause*, alors que la situation *Commerce_buy* définit les rôles sémantiques *Buyer*, *Seller* et *Goods*. Les rôles sont classifiés selon leur importance dans la situation : centraux (nécessaires), périphériques (toujours liés à la situation mais optionnels) et circonstanciels (pas spécifiques donc potentiellement présent dans toutes les situations, par exemple le lieu ou le temps).

FrameNet peut s'appliquer à d'autres domaines et langues (section 2.2.3). En particulier, le projet ANR ASFALDA produit un FrameNet du français.

1.3.5. Différences de vocabulaire entre VerbNet et FrameNet

FrameNet et VerbNet utilisent un vocabulaire différent, ce qui pose des problèmes d'ambiguïté : une frame FrameNet n'a aucun lien avec une frame VerbNet. La Table 1.1 liste les différences.

VerbNet	FrameNet
Classe (get-13.5.1)	Frame (Commerce_buy)
Rôle thématique (Agent)	Rôle sémantique (Buyer)
Frame (NP.Agent V NP.Theme)	-
Membre, verbe (buy)	Unité lexicale (buy, purchase)

TABLE 1.1. – Comparaison des terminologies FrameNet et VerbNet. Les exemples pour chaque concept se trouvent entre parenthèses.

Nous ne traduisons pas le terme *frame*, *cadre* étant suffisamment ambigu. En dehors de tout contexte, nous utilisons généralement la terminologie VerbNet. Les ambiguïtés sont résolues en précisant la ressource considérée quand c'est nécessaire, par exemple *la frame FrameNet*.

1.3.6. Les Verbes Français et le Lexique-Grammaire

À partir des années 1970 deux ressources lexicales pour les verbes français ont été développées : LVF [Dubois and Dubois-Charlier, 1970, Dubois and Dubois, 1971] et LG. Jean Dubois et Maurice Gross entretenaient une certaine forme de collaboration même s'ils ont développé des ressources lexicographiques avec des propriétés différentes ; la sémantique d'un côté, la syntaxe de l'autres. Plus tard, dans les années 1990, une autre ressource a été développée : Dicovalence. Nous ne l'utilisons presque pas dans nos travaux.

LVF (Les Verbes Français, Dubois and Dubois-Charlier [1997]) contient environ 25000 entrées classées en quatre niveaux :

— 14 classes génériques (par exemple *E : Verbes de mouvement*),
— 54 sous-classes sémantico-syntaxiques (par exemple *C1 : s'exprimer par un son, une parole*),
— 246 sous-classes syntaxiques (par exemple *M3b : imprimer tel mouvement à quelquec chose* et
— 533 sous-types (par exemple *R3b.2 : défaire l'opération faite*).

Chaque sous-type contient une liste de verbes à laquelle est associée diverses informations. Nous prendrons l'exemple de *aboyer 2*, le sens figuré d'aboyer (*Luc aboie des ordres*). Les informations associées sont :

— le numéro de sens du verbe (aboyer 2)
— un opérateur (*f.cri chien* pour faire le cri du chien),
— les constructions possibles (notamment T1300 pour transitif direct, humain, chose),
— un sens à l'aide de synonymes (hurler, crier après),
— des exemples (*On a˜ contre les voisins. On a˜ des injures*),
— et les dérivations possibles (en général -ment ou -ant).

Les tables du Lexique-Grammaire [Gross, 1975, Boons et al., 1976] sont l'autre ressource que nous utilisons, en abrégeant souvent par son nom par LG. La ressource comporte 14 000 entrées classifiées en 67 « tables », chaque table groupant des verbes partageant la même propriété définitoire syntaxique et potentiellement une sémantique similaire. Chaque colonne de la table encode des restrictions supplémentaires s'appliquant à certains des verbes de la table.

Par exemple, la table 6 [4], définie par N_0 V Qu P, contient des verbes d'attitude propositionnelle tels que penser ou estimer. Différentes propriétés s'appliquent ou non aux verbes de la table. Par exemple, pour la colonne N_0 V contre Nhum, argumenter a un + (*Léa a argumenté contre Jean* est correct), mais *abroger* a un - (**Léa a abrogé contre Jean* est incorrect). Ces colonnes peuvent être un moyen d'identifier des sous-classes de verbes partageant tous la même propriété représentée par une ou plusieurs colonnes.

Pourquoi ne pas utiliser ces ressources riches directement dans nos travaux ? Pour plusieurs raisons :

— LG ignore presque complètement la sémantique : de nombreuses classes auraient pu être séparées plus utilement en considérant des différences sémantiques exprimées par ailleurs dans la syntaxe.
— LVF, au contraire, se base sur des critères difficiles à comprendre et manifestement subjectifs, au contraire de VerbNet et LG.
— Ni LVF ni LG n'encodent de rôles thématiques ni de formules sémantiques [5].
— LVF inclut de nombreux usages techniques : il y a en réalité *trop* de verbes et de sens : nous voulons nous concentrer sur les verbes les plus fréquents, environ 5 000. Un moyen de pallier ce problème précis est d'utiliser le niveau de lexique L présent dans le LVF. Malheureusement, le niveau 1 est trop pauvre (1500 mots) alors que le niveau 2 est peut-être trop riche (15 000 mots).
— LVF (et dans une moindre mesure, LG), incluent de nombreux emplois métaphoriques, ce qui nuit à la ressource : des usages tels que *Il galopait dans son esprit que Marie allait venir* (extrait de LG) ne sont pas souhaitables. Certains usages métaphoriques peuvent certes être prise en compte dans VerbNet, mais ils n'y sont pas par défaut. En effet, Brown and Palmer [2012] proposent une analyse systématique des emplois métaphoriques de deux verbes représentatifs et montrent qu'utiliser VerbNet pour raisonner sur les emplois métaphoriques d'un texte est en partie possible au prix d'une complexité plus importante

4. `https://verbenet.inria.fr/verbes-html/V_6.lgt.html`
5. Les notions de rôles thématiques et d'évènement n'étaient pas répandues dans les années 1970.

et de prédicats sémantiques moins précis.

Ce sont les raisons principales pour laquelle nous voulons construire une nouvelle ressource française, Verb∋Net (Chapitre 4). Cette ressource tire profit d'une part des ressources existantes pour le français avec un encodage sémantique et syntaxique riche, et d'autre part de l'information sémantique présente dans VerbNet pour l'anglais, une langue proche du français. Verb∋Net peut être vu dans une certaine mesure comme une réorganisation des classes LG, étant donné que les deux ressources sont essentiellement basées sur des critères syntaxiques.

Conclusion

Après avoir introduit le cadre de nos travaux et présenté les différentes ressources que nous utilisons, le chapitre suivant présente l'état de l'art qui est le socle sur lequel nos travaux s'appuient.

2. État de l'art

> Un jour, alors que je venais de connaître Burrich, il m'avait ordonné de défaire le harnais d'un équipage de chevaux. [...] Quand Burrich revint voir ce qui me prenait tant de temps, il demeura muet de stupéfaction mais ne put me reprocher de n'avoir pas obéi à son ordre. Quant à moi, j'étais effaré du nombre de pièces qui entraient dans la composition d'un objet apparemment d'une seule pièce quand je m'y étais attaqué. J'avais la même impression sur la route : tous ces sons pour faire un mot, tous ces mots pour former une pensée ! Le langage tombait en morceaux entre mes mains. Jamais je n'y avais réfléchi.
>
> *(Robin Hobb, La Voie royale)*

La représentation du sens des mots (section 2.1) occupe une place importante dans nos travaux, en particulier pour la traduction de la ressource WordNet (Chapitre 3). La traduction de ressources lexicales (section 2.2) est un moyen de faire profiter une langue cible d'une ressource existante dans une autre langue, comme nous le faisons dans la partie II. Enfin, l'annotation en rôles sémantiques (section 2.3) sera utile pour la partie III.

2.1. Représentation des mots

Nous abordons certains aspects du sens des mots à travers l'étude de différentes ressources lexicales, en commençant par les dictionnaires. Nous présentons ensuite les modèles de langue qui sont une manière directe de représenter les mots et leurs sens à partir d'un corpus brut.

2.1.1. Représentation du sens des mots

« La lexicographie est la science qui consiste à recenser les mots, les classer, les définir et les illustrer, par des exemples ou des expressions, pour rendre compte de l'ensemble de leurs significations et de leurs acceptions au sein d'une langue, afin de constituer un dictionnaire » [Wikipédia, 2014]. La lexicographie est donc un socle sur lequel le Traitement Automatique des Langues peut s'appuyer pour représenter le sens des mots.

Pour pouvoir identifier les différents sens d'un mot, les lexicographes n'opèrent plus par intuition linguistique [Kilgarriff, 1997]. Ils commencent par établir un corpus équilibré et de taille

assez importante pour représenter la langue étudiée. Ce corpus peut par exemple être constitué de textes de journaux, de fiction, ou encore de blogs, le tout étant supposé être représentatif de ce qu'une personne lambda lit durant sa vie. Pour un mot donné, le lexicographe examine ses différents usages dans ce corpus à l'aide d'un concordancier dans le but de séparer ces différents usages en sens. Certains sens, jugés trop peu fréquents, sont laissés de côté. Le lexicographe étudie ensuite la séparation obtenue pour établir des critères objectifs distinguant les différents sens du mot étudié. Une phase d'ajustement de la séparation suit pour vérifier que les critères ont été correctement appliqués, ce qui peut amener à raffiner ces critères. Une fois les critères définitifs établis, la définition peut être rédigée, les occurrences étudiées pouvant servir d'exemples. L'avantage principal est que le processus lexicographique est basé sur des données réelles et non pas sur des intuitions linguistiques.

Ainsi, les sens ne sont pas définis en tant que tels, mais sont avant tout des occurrences dans un contexte donné. C'est une façon de comprendre la citation de [Firth, 1957] : *You shall know a word by the company it keeps* [1]. En effet, selon [Kilgarriff, 1997], un ensemble de sens n'est défini que par rapport à un corpus, et il est illusoire de vouloir définir un dictionnaire parfait pour tous les sens possibles d'un mot. Néanmoins, il n'est pas concevable de réaliser manuellement un dictionnaire par corpus : il faut surtout être conscient des difficultés théoriques posées par le sens des mots.

On considèrera dans ce travail que les sens définis dans un petit dictionnaire classique relèvent du « domaine général », et que les sens qui apparaissent dans d'autres domaines sont des sens « de spécialité ». Par exemple, le dictionnaire DicoInfo [OLST, 2014] spécialisé dans les domaines de l'Informatique et d'Internet mentionne un sens spécifique pour le nom *compilation* : *action effectuée par un compilateur qui consiste à transformer du code créé au moyen d'un langage de programmation évolué en un langage compréhensible par l'ordinateur*. Ce sens est notamment absent du TLFi [Pierrel, 2003] parce qu'il ne faisait pas partie des sens du mot dans le corpus utilisé pour établir les définitions. Le lexique *Les Verbes Français* (section 1.3.6) contient bien le sens informatique du verbe *compiler* mais le domaine (Informatique) et le niveau de lexique (5, soit « grands dictionnaires de langue ou encyclopédiques, lexiques spécialisés ») indiquent bien que c'est un sens de spécialité.

2.1.2. Ressources lexicales actuelles

D'autres moyens existent pour représenter le sens des mots. La qualité du travail lexicographique exposé dans les dictionnaires n'a pas été remise en cause, mais :

— les dictionnaires traditionnels, même dans leur version en ligne, ne tirent pas profit des nouveaux moyen d'organisation rendus possibles avec un ordinateur : il n'est plus nécessaire de trier les mots, on peut les représenter par un graphe [Miller et al., 1990, Polguère,

1. Vous devriez connaître un mot par ce qui l'accompagne.

2013],
— les dictionnaires traditionnels sont basés sur l'histoire des mots au lieu de considérer les progrès en linguistique et psycholinguistique proposant des organisations plus utiles et plus proches du lexique mental [Miller et al., 1990].

De plus, l'utilisation de dictionnaires récents implique un coût d'achat et le respect de la licence restrictive, ce qui explique que ces dictionnaires ont rapidement étés abandonnés au profit d'autres ressources généralement disponibles sous une licence libre comme WordNet, Wikipédia ou encore le Wiktionnaire. En effet, ces ressources autorisent une utilisation à la fois à des fins de recherche mais aussi pour un usage commercial, ce qui leur a assuré une large diffusion.

La première ressource lexicale à tirer partie de la possibilité de représenter le lexique sous la forme d'un graphe est WordNet, décrit à la section 1.3.1.

Hovy et al. [2006], Ide and Wilks [2006], Navigli et al. [2007], Snow et al. [2007] ont jugé que la trop grande finesse de distinction des sens de WordNet justifiait une alternative avec des sens distingués plus grossièrement. Ce problème est attribué selon Edmonds and Kilgarriff [2002] au manque de rigueur lexicographique de WordNet, et à la mise en avant de la similarité entre les mots au détriment de la distinction des différents sens de chaque mot. Il s'est en effet avéré que l'accord inter-annotateurs pour un étiquetage avec WordNet est faible : de l'ordre de 70% [Snyder and Palmer, 2004]. Utiliser un autre inventaire est un moyen efficace de s'adapter à différentes applications [Palmer et al., 2004], ce qui a entraîné des travaux utilisant des inventaires plus grossiers [Navigli et al., 2007].

Au-delà des fusions de synsets automatiques [Snow et al., 2007], de nouveaux inventaires de sens moins fins ont étés développés.

— OntoNotes [Hovy et al., 2006] a choisi de regrouper manuellement les sens WordNet jusqu'à obtenir un accord inter-annotateur de 90%.
— DANTE [2] [McCarthy, 2010] est un inventaire entièrement nouveau conçu dans l'objectif de corriger les erreurs faites avec WordNet [Kilgarriff, 2010].
— Le Réseau Lexical du Français [Polguère, 2014] lie des sens de mots avec de nombreuses fonctions lexicales associés à un degré de confiance, le tout permettant de produire des articles de dictionnaires.

Ces inventaires semblent plus adaptés que WordNet pour la désambiguïsation lexicale [Navigli, 2012]. Cependant, les deux premiers ne sont pas librement utilisables (en particulier à des fins commerciales) et le troisième, bien qu'il sera diffusé sous une licence libre, n'est pas encore terminé.

Une approche complètement différente est celle de la structure de qualia [Johnston and Busa, 1996] qui s'inscrit dans le contexte plus général du lexique génératif introduit par Pustejovsky

2. Les entrées pour les mots entre M et R sont disponibles sur `http://webdante.com/`.

[1991] et qui considère qu'une approche énumérative n'est pas viable. Le sens d'un mot est alors défini selon plusieurs aspects prédéfinis (constitution, rôles, facteurs impliqués dans la création, etc.) qui peuvent se retrouver dans plusieurs mots. Par exemple, un couteau contient une lame et sert à couper. Cette approche est semblable à celle qui définit le sens d'un mot comme une liste de sèmes [Rastier, 1987]. À notre connaissance, CoreLex [Buitelaar, 1998] est le seul inventaire et système de désambiguïsation lexicale suivant cette approche.

Différents travaux mentionnent la possibilité d'utiliser plus d'un sens pour un mot donné. Smith [2011] propose d'utiliser des distributions de probabilité sur les différents sens possibles pour définir un sens précis dans un corpus. Dans SemCor, les annotateurs pouvaient choisir plusieurs sens si besoin, mais seulement 0.3% des occurrences de SemCor sont étiquetées avec plus d'un sens. Erk et al. [2013] montrent qu'un accord inter-annotateur élevé peut être obtenu en demandant aux annotateurs d'indiquer pour chaque sens sa correspondance avec l'usage sur une échelle de 1 à 5. Une campagne d'évaluation a eu lieu en 2013 à ce sujet [Jurgens and Klapaftis, 2013]. Les annotations obtenues avec Amazon Mechanical Turk ont été abandonnées au profit de l'annotation par les deux organisateurs de la tâche. Dans les deux cas, l'accord inter-annotateur était modéré, ce qui remet en question la pertinence de l'annotation graduée de chacun des sens de l'inventaire.

Dans le cadre de l'adaptation aux domaines, d'autres travaux s'attachent à la prise en compte du changement des sens suivant les domaines. Agirre et al. [2010] ont proposé une campagne d'évaluation sur le domaine environnemental en anglais, chinois, néerlandais et italien. Pour chacune des langues, au moins un système a battu la baseline du sens le plus fréquent, et les systèmes fondés sur la connaissance ont pu s'adapter au domaine plus facilement que les systèmes supervisés. Les scores maximaux, entre 52 % et 57 %, restent loin de l'accord inter-annotateur, compris entre 72 % et 96 %. En fouille d'opinion, Marchand et al. [2014] évaluent eux l'apport de la prise en compte du changement de polarité des mots lors d'un changement de domaine. Par exemple, il s'avère que les termes mélioratifs au passé tels que « I loved » sont plutôt positifs pour le cinéma ou le théâtre et plutôt négatif en cuisine ou électroménager. Dans un cas, on parle simplement de la séance au passé ; dans l'autre, on aimait l'objet au début mais ce n'est plus le cas. Un autre exemple est l'adjectif « imprévisible » : c'est une qualité pour un film, et un défaut pour de l'électroménager.

Contrairement aux systèmes généraux de désambiguïsation lexicale qui cherchent à choisir un sens parmi un ensemble donné de sens, Marchand et al. [2014] ne s'intéressent qu'aux changements de sens utiles pour l'application, ce qui rend le système plus robuste. Différentes tâches bénéficiant de tels « inventaires » spécifiques à la tâche ont ainsi étés proposées lors de campagnes d'évaluation, en particulier SemEval. Par exemple :

— la simplification lexicale [Specia et al., 2012, Fabre et al., 2014],
— l'induction de sens et la désambiguïsation lexicale pour le groupement de résultats en recherche d'information [Navigli and Vannella, 2013],
— ou la traduction de fragments en langue maternelle (L1) dans un texte écrit dans une langue apprise (L2) [van Gompel et al., 2014]

Enfin, une tendance récente facilitée par l'existence de ressources telles que WordNet, Wikipédia ou le Wiktionnaire dans de nombreuses langues est de représenter les sens des mots dans différentes langues simultanément [Navigli, 2013]. Lors de SemEval-2010 [Lefever and Hoste, 2010], les meilleurs systèmes de la campagne utilisaient des corpus parallèles. Lors d'une tâche proche à SemEval-2013 [Navigli et al., 2013], aucun système n'a profité du fait que BabelNet soit un inventaire multilingue ni que les données de tests étaient parallèles. Néanmoins, les organisateurs ont remarqué qu'en ne validant que les sens qui étaient les mêmes lorsque les mots étaient alignés, la précision augmentait considérablement, ce qui est une piste à explorer pour de futurs travaux.

2.1.3. Modèles de langue pour la similarité sémantique

Une autre façon de représenter le sens des mots est possible grâce aux modèles de langues. Un modèle de langue prédit la probabilité d'un mot étant donné son contexte dans la phrase. Cette probabilité est directement utile pour des tâches telles que la traduction automatique ou la reconnaissance de la parole dans lesquelles un modèle de langue favorisera des phrases globalement plausibles au lieu d'étudier chaque mot individuellement.

Ces modèles de langue permettent aussi d'obtenir des mesures de similarité sémantiques utiles, ce qui est justifié par l'hypothèse distributionnelle [Harris, 1954, p. 156] :

> ... si l'on considère que le sens de deux mots ou morphèmes A et B diffère davantage que le sens de A et C, alors on observe souvent que les distributions de A et B diffèrent davantage que les distributions de A et C. Autrement dit, la différence de sens est corrélée à la différence de distribution.

Cette observation est utilisée depuis longtemps pour étudier les similarités sémantiques entre les mots à l'aide de matrices de co-occurrences [Miller, 1967]. Les modèles de langue sont une généralisation permettant d'étudier ces distributions de probabilité. On peut étudier deux types de distributions différentes correspondant à deux types de relations entre les mots [Sahlgren, 2008] :

— les relations syntagmatiques identifient les mots qui sont présents ensemble dans un contexte donné ;
— les relations paradigmatiques identifient les mots qui sont présents dans un même contexte, mais sans y être présents ensemble.

Par exemple, étant donné les deux phrases *Je bois du café* et *Je bois du thé*, on peut déduire que les lemmes *boire* et *thé* sont liés par une relation syntagmatique : ils sont présents ensemble dans la phrase. Au contraire, *thé* et *café* ne sont pas ici présents dans la même phrase, mais apparaissent dans un même contexte (*Je bois du*) : ils sont liés par une relation paradigmatique.

Selon l'hypothèse distributionnelle faible [Lenci, 2008], observer les distributions de contexte des mots peut séparer les mots en différents sens selon l'usage de chaque sens [Yarowsky, 1993, Pantel and Lin, 2002, Pedersen, 2010]. Cependant, dans la littérature que nous exposons et dans nos travaux, les modèles de langue ne décrivent que des mots en confondant leurs différents sens. Nous ne mentionnerons plus par la suite cette difficulté, en considérant qu'il suffit que le modèle de langue décrive parmi tous les sens celui que nous souhaitons observer. Cette simplification est largement partagée par la littérature, même si la disponibilité de corpus de plus d'un milliard de mots (*gigaword*) rend plus facile des travaux prenant en compte la polysémie [Kawahara et al., 2014].

Comment observer ces distributions de probabilité ? Un modèle de langue classique indique la probabilité d'un mot dans une phrase étant donné les mots précédents. Par exemple, étant donné le début de phrase *Au-delà des approches ...*, on veut connaître la probabilité du mot suivant, en espérant que celle de *statistiques* ou *supervisées* soit plus importante que celle de *chat*. En prenant par exemple le contexte des deux mots qui précèdent le mot étudié, on calcule sa probabilité simplement avec le maximum de vraisemblance :

$$p(w_i|w_{i-2},w_{i-1}) = \frac{\#(w_{i-2},w_{i-1},w_i)}{\#(w_{i-1},w_i)}$$

La séquence w_{i-2},w_{i-1},w_i est un 3-gramme, et # indique le nombre d'occurrences de cette séquence dans le corpus considéré. La taille du contexte peut varier, ce qui est la raison pour laquelle on parle de manière générale de n-grammes. Le nombre de paramètres à estimer pour une distribution de probabilité conditionnelle est $|V|^N$, $|V|$ étant la taille du vocabulaire et N la taille du contexte étudié. En considérant un petit vocabulaire (10 000 mots) et un contexte de trois mots, il faut déjà estimer 10^9 probabilités, ce qui requiert un corpus très large : Google a utilisé un corpus de livres de 10^{12} mots pour produire des n-grammes allant jusqu'à $n = 5$ [Brants and Franz, 2006]. Diverses techniques de lissage existent pour mieux répartir les probabilités obtenues par maximum de vraisemblance [Jurafsky and Martin, 2008, Chapitre 4]. En effet, la plupart des probabilités sont initialement nulles, que ce soit parce que le n-gramme est grammaticalement invalide ou simplement parce qu'il n'a pas été observé dans le corpus étudié. Il faut alors estimer la probabilité de tels n-grammes à partir d'un n-gramme plus court ou leur assigner une probabilité très faible pour éviter les probabilités nulles.

Modèles de langues syntaxiques Diverses extensions de ces modèles de langues à base de n-grammes existent, l'une d'entre elles étant le modèle de langue syntaxique [Lin, 1998, Goldberg and Orwant, 2013]. Dans ce modèle, on considère les mots présents ensemble dans une relation syntaxique donnée. Pour la relation complément du nom par exemple, on s'attend à ce que *vélo* soit le complément du nom des mots *pédale*, *guidon*, *pneu*... Nous utilisons pour la traduction de WordNet (Chapitre 3) un tel modèle de langue syntaxique (`http://www.kalisteo.fr/demo/semanticmap/`). Il a été entraîné sur un corpus extrait du web

francophone [Grefenstette, 2007]. Le corpus a ensuite été analysé par LIMA [Besançon et al., 2010], une chaîne d'analyse linguistique désormais libre utilisée ici comme un analyseur syntaxique à base de règles produisant des dépendances syntaxiques fines. Pour une relation donnée r et un lemme x, le modèle de langue indique quels sont les 100 premiers lemmes co-occurrant le plus fréquemment avec x dans la relation r. Avec le mot *avion* et la relation de complément du nom, c'est le mot *billet* qui modifie le plus *avion* : *billet d'avion* est fréquent dans le corpus.

billet, pilote, vol, accident, détournement, tour, collision, moteur, monde, crash, aile, bruit, type, carburant, attentat, guerre, chute, commande, construction, descente, nombre, prix, achat, place, réservation, passager, programme, bombe, peur, avion, flotte, transport pilotage, écrasement, vitesse, utilisation, arrivée, ...

FIGURE 2.1. – *avion* est complément de ces noms d'après notre modèle de langue syntaxique. La taille des mots reflète la force de l'association. L'ordre des mots est aléatoire.

Modèles de langues continus D'autres types de modèles de langue représentent les mots de manière distribuée en utilisant un vecteur de nombres réels. LSA [Deerwester et al., 1990], par exemple, considère une matrice de termes et de documents, les termes étant en ligne, les document en colonne. Les éléments de cette matrice sont typiquement calculés avec TF-IDF. Cette matrice est ensuite réduite, ce qui permet de généraliser en rapprochant la similarité cosinus des mots qui apparaissent dans les mêmes documents. LSA fait partie d'une classe de méthodes où on considère une matrice de co-occurrence entre les mots et leurs contexte, une mesure d'association quelconque, une factorisation potentielle, et une mesure de la distance entre les éléments, l'objectif final étant de rapprocher les mots sémantiquement proches.

Modèles de langues neuronaux Depuis la diffusion de word2vec [Mikolov et al., 2013] et de ses bons résultats obtenus dans un certain nombre de tâches, l'utilisation de représentations de mots à l'aide de réseaux de neurones est un champ actif de recherche [3]. La manière la plus répandue pour faire cela est d'utiliser un réseau de neurones peu profond dont une des couches sera le vecteur représentant chaque mot. Hinton [1986] a d'abord proposé l'idée d'un réseau de neurones pour représenter des concepts à l'aide de vecteurs faisant partie du réseau, Bengio et al. [2001, 2003] ont présenté le modèle de langue neuronal tel qu'on le connaît aujourd'hui, et Mikolov et al. [2013] (parmi d'autres) a optimisé un modèle de langue en supprimant notamment une couche cachée pour l'utiliser sur de plus gros corpus. Dans tous ces modèles, lors de la rétropropagation du gradient, la représentation des mots évolue, ce qui a pour effet de rapprocher

3. Les bons résultats obtenus notamment en vision par ordinateur par des réseaux de neurones profonds ont sûrement contribué à cet enthousiasme, il est donc important de rappeller que les réseaux de neurones concernés ici ne sont pas profonds.

la représentation des mots sémantiquement proches afin d'offrir une meilleure généralisation, comme le fait LSA en factorisant la matrice de co-occurrences. Par exemple, Mikolov et al. [2013] montrent que la distance entre le mot représentant un pays (par exemple *Turkey*) et le mot représentant la capitale de ce pays (ici *Ankara*) correspond au vecteur du mot *capital*. Ces représentations de mots peuvent être ensuite utilisées ou apprises pour diverses tâches. Nous citerons ici l'extraction d'évènements [Boros et al., 2014], l'annotation en rôles sémantiques [Léchelle and Langlais, 2014] et la traduction automatique [Devlin et al., 2014], mais toute tâche peut bénéficier d'améliorations (plus ou moins importantes) de telle représentations de mots où les mots sémantiquement proches sont proches dans la représentation choisie. Ces réseaux de neurones sont encore difficiles à entraîner et à paramétrer [Do et al., 2014], mais semblent représenter une alternative plus efficace que les modèles de langue simples à base de n-grammes [Baroni et al., 2014].

La compréhension des mécanismes derrière ces réseaux de neurones est aussi un champ actif de recherche. Levy and Goldberg [2014] ont montré que la méthode recommandée par Mikolov et al. [2013] (*negative sampling*) revient à factoriser la matrice d'information mutuelle (légèrement modifiée) entre les mots et les contextes en accordant plus d'importance aux termes fréquents. Cette méthode est très proche des méthodes couramment utilisées en sémantique distributionnelle et permet d'espérer une optimisation plus directe de l'objectif afin d'améliorer les résultats tout en assurant une meilleure compréhension des méthodes utilisées.

Bien que nous ne tirions pas profit des méthodes les plus récentes dans les travaux que nous avons choisi de présenter ici, ces progrès importants sont à considérer pour tout travail futur sur le sujet.

2.2. Traductions de ressources linguistiques

2.2.1. WordNet

WordNet (section 1.3.1) est une ressource extrêmement utile pour l'anglais : reproduire ce travail pour d'autres langues serait coûteux et difficile à maintenir. Malgré quelques problèmes théoriques, traduire WordNet en gardant sa structure et ses synsets mène à des ressources linguistiques utiles [Fellbaum and Vossen, 2007, de Melo and Weikum, 2008]. Cependant, il n'existe encore que peu d'équivalents de même qualité dans d'autres langues [Bond and Paik, 2012], et il est donc utile de s'atteler à la traduction de cette ressource.

Les traductions automatiques de WordNet emploient une approche dite d'extension (*expand approach*) : la structure de WordNet est préservée et seuls les littéraux sont traduits. Trois techniques principales représentent cette approche dans la littérature. La plus simple utilise des dictionnaires bilingues pour faciliter le travail des lexicographes qui filtrent ensuite manuellement

les entrées proposées [Vossen, 1998, Pianta et al., 2002, Tufiş et al., 2004]. Une deuxième méthode de traduction utilise des corpus parallèles, ce qui évite l'utilisation de dictionnaires qui peuvent entraîner un biais lexicographique. Dyvik [2002] représente cette méthode en s'appuyant sur des *back-translations* entre le norvégien et l'anglais, alors que [Sagot and Fišer, 2008a] combinent un lexique multilingue et les différents WordNets de BalkaNet comme autant de sources aidant à la désambiguïsation. Troisièmement, plus récemment, des ressources telles que Wikipédia ou le Wiktionnaire ont été explorées. Grâce aux nombreux liens entre les différentes langues de ces ressources, il est possible de créer de nouveaux wordnets [de Melo and Weikum, 2009, Navigli and Ponzetto, 2010, Bond and Foster, 2013, Aliabadi et al., 2014, Oliver, 2014] ou d'améliorer des wordnets existants [Hanoka and Sagot, 2012]. Enfin, Fišer et al. [2014] montrent qu'il est possible de faire appel à la myriadisation (*crowdsourcing*).

Concernant le français, l'EuroWordNet [Vossen, 1998] est la première traduction française de WordNet. C'est une ressource d'une couverture limitée qui demande des améliorations significatives avant de pouvoir être utilisée [Jacquin et al., 2006], et qui n'est ni libre ni librement accessible. WOLF est une seconde traduction initialement construite à l'aide de corpus parallèles [Sagot and Fišer, 2008a] et étendue depuis avec différentes techniques [Apidianaki and Sagot, 2012]. WOLF est distribué sous une licence libre compatible avec la LGPL et c'est aujourd'hui le WordNet français standard. JAWS [Mouton and de Chalendar, 2010] est une traduction des noms de WordNet développée à l'aide de dictionnaires bilingues et d'un modèle de langue syntaxique. Enfin, Gader et al. [2014] montrent qu'il est possible de convertir le Réseau Lexical du Français vers une représentation WordNet.

WoNeF, notre traduction de WordNet (Chapitre 3) est le successeur de JAWS [Mouton and de Chalendar, 2010, Mouton, 2010]. Cette section présente JAWS. Le Chapitre 3 traitera lui seulement des différences entre JAWS et WoNeF.

JAWS repose sur un algorithme faiblement supervisé qui ne demande aucune donnée annotée manuellement. Pour traduire un wordnet source, JAWS s'appuie sur un dictionnaire bilingue et un modèle de langue syntaxique pour la langue cible. Dans notre cas, la langue source est l'anglais, la langue cible est le français et le dictionnaire bilingue est une concaténation du dictionnaire bilingue SCI-FRAN-EurADic [4] et des liens [5] entre les Wiktionnaires français et anglais [6]. Le modèle de langue syntaxique a été présenté à la section 2.1.3. Grâce aux dictionnaires, JAWS n'a pas besoin de sélectionner les littéraux de chaque synset parmi l'ensemble du vocabulaire mais seulement parmi un petit nombre de candidats (9 en moyenne). Le processus de traduction se fait en trois étapes :

1. Créer un wordnet vide : la structure de WordNet est préservée, mais les synsets eux-mêmes n'ont pas de littéraux associés.

4. `http://catalog.elra.info/product_info.php?products_id=666`
5. `https://github.com/pquentin/wiktionary-translations`
6. `http://www.wiktionary.org/`

2. Choisir les traductions les plus faciles parmi les candidats des dictionnaires pour commencer à remplir JAWS (sélecteurs initiaux).
3. Étendre JAWS de manière incrémentale en utilisant le modèle de langue, les relations entre synsets et le JAWS déjà existant (sélecteurs syntaxiques).

Sélecteurs initiaux Quatre algorithmes que nous nommons sélecteurs initiaux choisissent des traductions correctes parmi celles qui sont proposées par les dictionnaires.

— Premièrement, les mots qui apparaissent dans un seul synset ne sont pas ambigus et il suffit d'ajouter toutes leurs traductions au WordNet français : c'est le sélecteur par monosémie. C'est le cas de *grumpy* : toutes ses traductions sont validées dans le seul synset où il apparaît.
— Deuxièmement, le sélecteur par unicité identifie les mots n'ayant qu'une seule traduction et la valide dans tous les synsets où cette traduction est présente. Les cinq synsets contenant *pill* en anglais sont ainsi complétés avec *pilule*.
— Un troisième sélecteur vise à traduire les mots qui ne sont pas dans le dictionnaire en utilisant directement la traduction anglaise : c'est le sélecteur des transfuges.
— Un quatrième sélecteur utilise la distance d'édition de Levenshtein : si la distance entre un mot anglais et sa traduction est petite, on peut considérer que c'est le même sens (c'est le cas par exemple pour *portion* ou encore *university*), malgré l'existence de certains faux amis. Ces quatre sélecteurs produisent une première version du WordNet français qui contient assez de traductions pour pouvoir ensuite utiliser le modèle de langue et continuer de compléter les synsets.

Expansion de JAWS JAWS étant partiellement rempli, une nouvelle étape d'expansion tire parti des relations entre les synsets de WordNet pour valider de nouvelles traductions. Par exemple, si :

— un synset S1 est méronyme d'un synset S2 dans WordNet,
— dans notre modèle de langue, un littéral de S1 est méronyme d'un littéral candidat C dans S2,

alors ce littéral est considéré comme correct. La tâche de traduction est ainsi réduite à une tâche de comparaison entre d'une part les relations lexicales entre les synsets de WordNet et d'autre part les relations lexicales entre les lexèmes du français.

Prenons l'exemple de *quill* qui peut se traduire par *piquant* ou *plume* (Figure 2.2). Dans WordNet, *quill* est méronyme de *porcupine* qui a déjà été traduit par *porc-épic* par un sélecteur initial. Dans le modèle de langue, *piquant* fait partie des compléments du nom de *porc-épic* mais ce n'est pas le cas de *plume*. Ici, la relation de complément du nom implique la méronymie et c'est

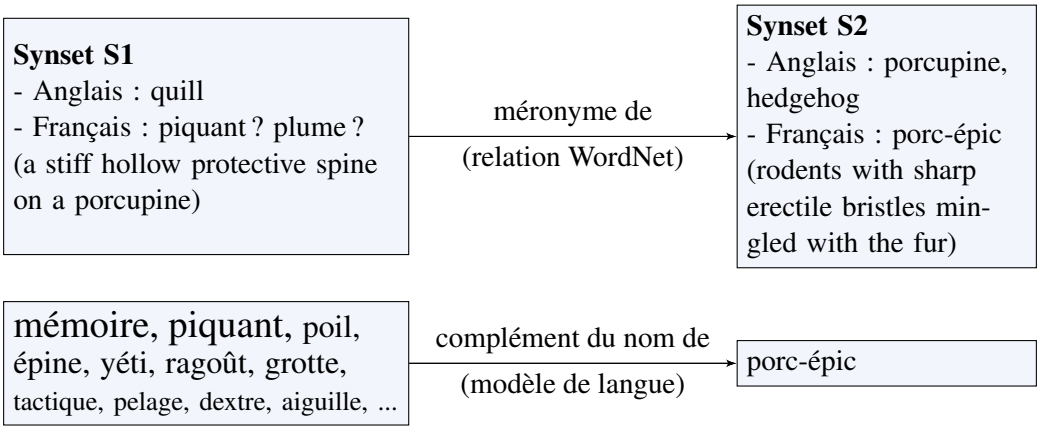


FIGURE 2.2. – Traduction via la relation de méronymie de partie.

donc *piquant* qu'il faut choisir comme la traduction correcte de *quill*. Le modèle de langue a permis la désambiguïsation parmi les deux traductions possibles.

Un problème potentiel avec cette approche est que la relation de complément du nom n'est pas limitée à la méronymie. Par exemple, le mot *mémoire* qui est lui aussi associé à *porc-épic* dans le modèle de langue vient d'un livre intitulé *Mémoires d'un porc-épic*. Heureusement, *mémoire* n'est pas dans les candidats de *quill* et ne peut pas être choisi comme une traduction. Paradoxalement, le modèle de langue ne peut pas choisir entre deux mots très différents, mais est capable de faire le bon choix parmi les différentes traductions d'un mot polysémique. Alors que traduire WordNet automatiquement avec un dictionnaire ou un modèle de langue syntaxique est impossible, combiner les deux sources d'information permet de résoudre le problème.

Chaque sélecteur syntaxique suit le même principe que le sélecteur par méronymie de partie et traduit de nouveaux synsets en identifiant les relations entre lexèmes via le modèle de langue syntaxique. La correspondance entre la relation de complément du nom et la relation de méronymie est directe, mais ce n'est pas le cas pour les autres relations : il n'y a par exemple pas de relation syntaxique qui exprime directement la synonymie entre deux lexèmes. Pour ces relations, il est nécessaire d'employer soit des motifs lexicaux [Hearst, 1992] soit des relations paradigmatiques [Lenci and Benotto, 2012]. Ce sont ces dernières (section 2.1.3) que JAWS utilise. Pour la synonymie, si deux mots partagent les mêmes co-occurrents dans une relation syntaxique donnée, alors ils peuvent être synonymes dans ce contexte. Pour les noms, les relations syntaxiques qui donnent les meilleurs résultats sont les relations de complément du nom, d'objet du verbe et d'apposition. Concrètement, si deux noms qui modifient les mêmes noms sont les objets des mêmes verbes ou sont apposés aux mêmes noms, alors il est probable qu'ils soient synonymes et si l'un des deux est déjà dans un synset, alors on peut y ajouter le second. Par exemple, *avant-propos* et *préface* partagent les mêmes compléments du noms : *livre, édition, ouvrage*. Le sélecteur par synonymie peut ajouter *avant-propos* une fois que le littéral *préface* est dans JAWS. [Mouton and de Chalendar, 2010, Mouton, 2010] décrivent d'autres sélecteurs

exploitant notamment les relations d'hyperonymie et d'hyponymie.

2.2.2. VerbNet

Étant donné l'intérêt de la classification de verbes à la manière de VerbNet [Hartshorne et al., 2014] et l'intérêt de la ressource en Traitement Automatique des Langues, VerbNet a été traduit ou recréé dans plusieurs langues.

Acquisitions de VerbNets Premièrement, même si ce n'est pas dans le cadre d'une traduction de VerbNet, un certain nombre de travaux cherchent à catégoriser automatiquement les verbes en classes sémantiques, notamment pour l'espagnol [Ferrer, 2004], l'allemand [im Walde, 2006], le japonais [Suzuki and Fukumoto, 2009] et l'anglais [Stevenson and Joanis, 2003, Lapata and Brew, 2004, Vlachos et al., 2009, Lippincott et al., 2012, Kawahara et al., 2014], où l'évaluation se fait notamment sur les classes de Levin. Pour l'anglais, [Kawahara et al., 2014] montrent qu'un corpus de plusieurs milliards de mot permet de considérer la polysémie. En particulier, un corpus de 20 milliards de mot améliore de manière conséquente l'état de l'art en considérant l'affectation de verbes à plusieurs classes (71.39 % de F1 contre 61.97 % pour LDA-frames [Materna, 2012]).

Traductions de VerbNet Concernant les traductions directes, Merlo et al. [2002] ont utilisé des similarités entres langues pour convertir 20 classes de Levin vers l'italien. Les seules traductions directes de VerbNet dont nous avons la connaissance sont le VerbNet estonien [Jentson, 2014] et le VerbNet portuguais brésilien [Scarton and Aluısio, 2012] (qui utilise des mappings entre VerbNet et WordNet, et entre WordNet.Br et WordNet). Scarton et al. [2014] proposent la création d'un VerbNet en s'appuyant comme seul effort manuel sur la traduction manuelle des frames VerbNet vers la langue cible, puis de ressources lexicales déjà disponibles (WordNet, VerbNet et WordNet traduit dans la langue cible), le tout dans l'optique de créer des VerbNets pour les langues proches de l'anglais disposant déjà d'un WordNet. La méthode a été testée pour le brésilien portuguais ; la comparaison des clusters de verbes de [Scarton and Aluısio, 2012] indique un F-score de 60 %. Comme l'indiquent les auteurs, même si l'effort pour créer un nouveau VerbNet est réduit, le lexique obtenu a encore besoin d'être corrigé manuellement.

VerbNet en français Pour le français, Saint-Dizier [1996] a produit une ressource proche des classes de Levin dans l'objectif de s'en servir en Traitement Automatique de Langues, tout comme VerbNet dix ans plus tard. Chacun des 1700 verbes est décrit par un certain nombre de « contextes » plus proches des alternances de Levin que des frames VerbNet. Ces contextes proviennent de différentes sources : les classes de Levin, les tables du Lexique-Grammaire, de

corpus et d'intuition linguistique. Cependant, l'effort sur cette ressource s'est arrêté et le résultat n'est disponible que sur demande à l'auteur.

Des travaux se sont ensuite concentrés sur le regroupement de verbes en se basant sur :
— des cadres de sous-catégorisation acquis automatiquement,
— et des similarités sémantiques.
En effet, Sun et al. [2010] ont utilisé un large lexique de cadres de sous-catégorisation [Messiant et al., 2010] pour regrouper les verbes en clusters à l'aide de traits sémantiques (colocations et préférences lexicales des verbes) et syntaxiques (cadres de sous-catégorisation). L'évaluation sur une vérité-terrain créée manuellement a mené à une F-mesure de 55.1%. Falk et al. [2012] appliquent un algorithme de clustering différent, utilisent des features différentes, mais s'évaluent sur la même vérité-terrain mais simplifiée, ce qui donne une F-mesure de 70%. Ces ressources mettent en valeur de nouvelles façons de séparer les verbes du français en traitant la polysémie, mais les erreurs qu'elles contiennent seront une nouvelle source d'erreur dans les applications : il est important de les corriger si possible, ce qui est la raison pour laquelle nous adoptons une approche différente qui sera présentée au Chapitre 4.

2.2.3. FrameNet

La théorie de FrameNet, présentée à la section 1.3.4, peut être utilisée dans d'autres domaines et d'autres langues [Boas, 2009]. Il existe ainsi notamment des FrameNet en espagnol [Subirats and Petruck, 2003], japonais [Ohara et al., 2004], suédois [Heppin and Gronostaj, 2012] ou encore français avec le projet ANR ASFALDA [Candito et al., 2014]. Certaines méthodes automatiques utilisent le Wiktionnaire pour traduire le lexique FrameNet, mais pas son corpus [Mouton et al., 2010, Hartmann and Gurevych, 2013], ce qui n'est pas directement utile. Le Kicktionary [Schmidt, 2009] est lui spécifique au football et annote en trois langues (français, anglais, allemand) des dépêches de presses de l'UEFA. Venturi et al. [2009] proposent un FrameNet italien centré sur le domaine du droit. Enfin, il existe plusieurs FrameNets spécialisés en portugais brésilien : un dans le domaine du droit [Bertoldi and Chishman, 2012] et l'autre pour la coupe du monde 2014 [Torrent et al., 2014].

FrameNet est donc une ressource très riche qui peut s'adapter à de nombreux domaines et à de nombreuses langues. Malheureusement, nous considérons que le coût pour développer une telle ressource pour un nouveau domaine est prohibitif et n'utilisons cette ressource que pour l'évaluation. Le projet FrameNet est encore loin de couvrir un éventail complet du vocabulaire anglais [Màrquez et al., 2008, section 5.4], nécessitant de trouver d'autres moyens pour étendre la ressource, notamment avec la myriadisation (*crowdsourcing*) [Fossati et al., 2013, Baker, 2014].

2.3. Annotation en rôles sémantiques

Dans cette section, nous étudions les différentes façons de définir les rôles sémantiques, examinons diverses ressources utiles pour la tâche d'annotation, et présentons les techniques principales pour réaliser l'annotation elle-même.

2.3.1. Les rôles sémantiques

La notion de rôle sémantique semble particulièrement adaptée à notre volonté d'aller au-delà de l'analyse syntaxique (section 1.1.2). Ces rôles ont en effet pour objectif de s'abstraire des alternances de diathèse présentes dans le langage naturel (Figure 2.3).

Carol Agent	crushed V	the ice Patient
The ice Patient	crushes V	easily

FIGURE 2.3. – Ces deux phrases annotées avec la classe VerbNet `carve-21.2` montrent que la position des arguments ne détermine pas à elle seule les rôles sémantiques. Ici, le sujet syntaxique est tour à tour Agent (pour *Carol*) puis Patient (pour *The ice*).

Fillmore [1968] a établi que le cas grammatical exhibe des relations profondes et sémantiques. De nombreuses langues marquent ces relations au niveau morphologique ; l'élatif est un exemple de cas grammatical qui exprime le lieu de l'intérieur duquel provient un mouvement et qui est marqué morphologiquement en finnois, hongrois et estonien. On peut alors, en se basant sur ces observations linguistiques, définir un rôle sémantique pour ces cas grammaticaux, même dans les langues où ils ne sont pas marqués morphologiquement. La théorie des *frame semantics* [Fillmore, 1982] est une modification de la théorie des cas, et a abouti à FrameNet où chaque situation ou *frame* dispose de ses propres rôles sémantiques, ce qui a créé plusieurs centaines de rôles. Ces rôles partageant parfois le même nom, mais ce sont les liens entre les rôles qu'il faut exploiter pour faire des généralisations [Litkowski, 2014]

Il n'y a pas de réel consensus sur un inventaire de cas donnés. Parmi les rôles sémantiques généralement acceptés [Palmer et al., 2010, p. 4], on peut citer :

— l'**Agent** qui est à l'origine de l'action
— le **Patient** qui subit un changement d'état
— l'**Instrument** utilisé pour réaliser l'action
— le **Bénéficiaire** qui tire profit de l'action

Nous adoptons dans ce travail les rôles de notre lexique : VerbNet. Ce sont 27 [7] rôles supposés être suffisamment généraux pour s'adapter à toutes les situations : *Actor*, *Agent*, *Asset*, *Attribute*, *Beneficiary*, *Cause*, *Co-Agent*, *Co-Patient*, *Co-Theme*, *Destination*, *Experiencer*, *Extent*, *Goal*, *Initial_Location*, *Instrument*, *Location*, *Material*, *Product*, *Patient*, *Pivot*, *Predicate*, *Recipient*, *Result*, *Source*, *Stimulus*, *Theme*, *Time*, *Topic*, *Trajectory* et *Value*.

2.3.2. Lexiques et corpus

Il existe différentes ressources utiles dans le cadre de l'annotation en rôles sémantiques. Toutes les ressources que nous utilisons directement dans nos travaux ont déjà été présentées à la section 1.3.

PropBank PropBank [Palmer et al., 2005] a décidé d'utiliser les annotations syntaxiques du Penn TreeBank [Marcus et al., 1993] pour annoter en rôles sémantiques les phrases incluant un des 5000 verbes les plus fréquents du corpus. Le principe est le même qu'avec les classes de Levin et VerbNet : la syntaxe joue un rôle important pour la désambiguïsation des sens et l'attribution des rôles sémantiques. Contrairement à VerbNet, PropBank se base sur un corpus pour identifier les différentes constructions syntaxiques, les sens des verbes et le sens à apporter aux rôles sémantiques. L'objectif principal de PropBank est de permettre d'utiliser l'apprentissage automatique pour l'annotation en rôles sémantiques. C'est pour cette raison que les étiquettes disponibles sont très générales. Ainsi, il est fréquent que *ARG0* désigne l'agent, *ARG1* le patient. D'autres arguments sont disponibles pour étiqueter des rôles plus spécifiques (*ARG2*, *ARG3*, etc.) ainsi que des rôles secondaires (*Location*, *Extent*, *Manner*, etc.)

NomBank NomBank [Meyers et al., 2004] a été conçu à l'image de PropBank. La spécificité de cette ressource est de se concentrer sur les noms communs, plus particulièrement sur les 5 000 noms communs les plus fréquents dans le Penn TreeBank. Sur le million de mots présents dans le corpus, 250 000 sont des noms communs. 100 000 d'entre eux sont des noms issus d'un verbe ou qui se comportent à la façon d'un verbe. Par exemple, le nom commun français « achat » est lié au verbe « acheter », et les arguments sémantiques seront probablement les mêmes : dans « Il a acheté un arbre » et « l'achat d'un arbre », *ARG1* sera dans les deux cas l'arbre. D'autres catégories incluent les noms partitifs, relationnels et environnementaux.

Pour une phrase telle que *They gave the chefs a standing ovation*, les annotations PropBank et NomBank proposent la même annotation, l'une étant basée sur le groupe nominal (*a standing ovation*), l'autre sur la structure prédicat-argument autour du verbe *gave*. Cette similarité volontaire a permis de lier ces ressources [Pustejovsky et al., 2005, Verhagen et al., 2007], mais nous ne connaissons pas d'applications tirant profit de ces deux ressources.

7. Il y a 30 rôles en comptant Co-Agent, Co-Patient et Co-Theme.

Gerber and Chai [2010] ont étendu NomBank aux arguments implicites, améliorant ainsi la couverture de NomBank de 65%, c'est-à-dire en augmentant le nombre moyen de rôles remplis dans chaque exemple annoté. Il n'est pas rare que les arguments soient implicites mais présents dans d'autres phrases. En effet, les annotations étant volontairement limitées à la phrase, il n'est pas possible de référer à un argument présent dans une phrase précédente.

Enfin, PropBank lui-même a récemment décidé d'inclure d'autres parties du discours impliquant potentiellement des structures prédicat-argument en s'inspirant notamment de NomBank : les noms, les adjectifs, et les verbes support [Bonial et al., 2014].

Groningen Meaning Bank (GMB) Ce projet [Basile et al., 2012] vise à fournir un large corpus annoté de l'anglais avec un nombre important de couches, le but étant de fournir un grand corpus pour la recherche en « sémantique ». Sont notamment inclus parmi les couches : les lemmes, les parties du discours, la syntaxe (avec CCG). Il y a aussi les sens WordNet, les rôles thématiques VerbNet, des annotations sémantiques avec la DRT et des relations discours avec SDRT. C'est un corpus prometteur qui unifie un nombre important d'annotations complémentaires et jusque-là distinctes.

L'annotation des rôles VerbNet est effectuée directement au niveau de la syntaxe, CCG donnant un moyen élégant d'associer les rôles au token prédicat [Bos et al., 2012]. Quelques difficultés conséquentes subsistent avant de pouvoir s'en servir comme d'un corpus VerbNet :

— Les classes VerbNet ne sont pas explicitement annotées, il faut les retrouver à travers le sens WordNet.
— Les frames VerbNet ne sont pas explicitement annotées, il faut les reconstruire d'après les supertags CCG.
— Il n'y a aucune garantie de cohérence avec VerbNet : si une frame est manquante ou erronée dans VerbNet, les annotateurs utilisent le mode *open* et peuvent choisir n'importe quel rôle VerbNet.

Abstract Meaning Representation Bank (AMR Bank) Ce projet [Banarescu et al., 2013] est un autre corpus à visée sémantique pour l'anglais. Sa particularité est qu'il propose une forme d'interlingue permettant d'annoter sémantiquement des corpus parallèles. Les rôles sémantiques sont ceux de PropBank, ce qui permet d'envisager de s'en servir en tant que corpus VerbNet.

2.3.3. Approches d'annotation

Les systèmes d'annotation en rôles sémantiques utilisent deux types de ressources :

1. Les **inventaires** examinés aux sections 1.3 et 2.3.2 permettent de fournir un socle commun à différents systèmes. Ce sera par exemple la définition des frames, des rôles, des cadre de sous-catégorisation et des prédicats possibles.
2. Les **corpus annotés** par des humains utilisent un inventaire donné pour réaliser la tâche qu'on essaie de faire apprendre aux systèmes. FrameNet contient de nombreux exemples annotés en plus des rôles sémantiques définis pour chacune des situations.

Ces ressources sont utilisées différemment suivant les méthodes, souvent divisées en quatre approches générales : supervisées, fondées sur la connaissance, semi-supervisées et non supervisées.

Supervisées Les méthodes supervisées [Gildea and Jurafsky, 2002, Surdeanu et al., 2008, Das et al., 2014, Hermann et al., 2014, Lluís et al., 2014] utilisent un corpus annoté, et adoptent donc l'inventaire associé. Des techniques classiques d'apprentissage automatique sont utilisées pour déterminer le sens correct de chaque occurrence d'un mot étant donné les informations obtenues à partir du contexte de cette occurrence. L'annotation en rôles sémantiques supervisée est souvent divisée en plusieurs sous-tâches, parfois partiellement regroupées :

— l'identification des prédicats,
— l'identification des frames,
— l'identification des arguments qui établit les syntagmes jouant un rôle dans la phrase,
— et la classification des rôles qui détermine le rôle effectif de chaque syntagme parmi ceux retenus à la phase précédente.

Nous ne rentrons pas ici dans le détail des techniques utilisées par les méthodes supervisées. Elles s'adaptent en général efficacement à leur corpus annoté et possèdent donc les meilleurs résultats sur ces corpus, mais d'autres techniques sont souhaitables pour généraliser à d'autres domaines non couverts par ces corpus annotés.

Fondées sur la connaissance Quelques approchent n'utilisent pas de corpus annoté mais se contentent de la ressource VerbNet [Swier and Stevenson, 2004, 2005, Pradet et al., 2013b]. Les systèmes s'affranchissent alors de la petite taille inhérente à tout corpus annoté et s'appuient sur les cadres de sous-catégorisation pour l'annotation. Un inventaire de sens est utilisé : il faut toujours aussi faire de la classification ; la difficulté principale étant ici d'obtenir des informations utiles sans exemples annotés. Étant donné que ces méthodes continuent à utiliser un inventaire, il reste possible de comparer les résultats entre différents systèmes et de réaliser une évaluation sur une vérité-terrain. C'est l'approche pour laquelle nous optons ici : nos apports sont décrits au Chapitre 5.

Semi-supervisées van der Plas and Apidianaki [2014] annotent le corpus Europarl anglais avec un système automatique entrainé sur PropBank, puis utilisent les alignements du corpus pour obtenir un corpus français automatiquement annoté en rôles PropBank. C'est une piste intéressante pour obtenir des corpus annotés en rôles sémantiques pour le français. Diverses difficultés restent : les données PropBank anglaises ont été directement utilisées pour le français, et les scores encore trop faibles nécessitent une correction manuelle du corpus. De manière similaire, Exner et al. [2014] proposent de développer automatiquement un nouveau PropBank du suédois en se basant sur le PropBank anglais et en utilisant Wikipédia comme un corpus parallèle. La précision des données obtenues est encore inconnue.

D'autres méthodes transfèrent directement les modèles appris d'une langue vers d'autres. [Zeman and Resnik, 2008] appliquent un modèle appris sur un corpus danois et l'appliquent à un corpus suédois, et montrent qu'il faut un corpus de 1500 phrases annotées en suédois pour qu'un modèle entraîné directement sur le corpus suédois soit meilleur que leur système. [Kozhevnikov and Titov, 2013] montrent sur trois paires de langues (Anglais-Chinois, Anglais-Tchèque et Anglais-Français) que les résultats sont meilleurs en modifiant le modèle appris qu'en projetant l'annotation comme au paragraphe précédent.

Non supervisées Ces approches n'utilisent aucune connaissance *a priori*, que ce soit un inventaire ou un corpus annoté. Une approche non supervisée doit nécessairement construire son propre inventaire. Cette construction peut se faire via du *clustering* de sens à partir des occurrences de contextes trouvées dans le corpus [Lang and Lapata, 2011, Garg and Henderson, 2012, Titov and Klementiev, 2012, Materna, 2013].

Les avantages potentiels sont nombreux. Ces algorithmes ne nécessitent aucune ressource, et offrent de fait deux propriétés intéressantes :

— L'inventaire choisi colle au plus près du corpus utilisé, ce qui lui permet à la fois d'éviter des distinctions trop fines et de s'adapter à de nouveaux domaines via de nouveaux corpus, le domaine ayant un impact important sur les sens utilisés.
— Plus la quantité de texte disponible augmente, plus le système peut devenir efficace.

Malheureusement, les systèmes utilisant une approche non supervisée sont difficiles à évaluer et à utiliser directement dans d'autres systèmes en raison de la difficulté à interpréter les rôles obtenus. Par exemple, dans le cadre de la traduction automatique, distinguer les sens ne suffit pas ; il faut aussi savoir quelle traduction appliquer. Ainsi, bien que représentant une voie prometteuse, nous n'avons pas considéré ici ce type de méthode.

2.3.4. Terminologie

En français, le terme « annotation en rôles sémantiques » n'est pas encore stabilisé. En effet, de nombreux termes coexistent encore aujourd'hui :

— *Annotation syntaxico-sémantique des actants* [Hadouche, 2011],
— *étiquetage en rôles sémantiques* [Boros et al., 2014],
— *étiquetage de rôles sémantiques* [Léchelle and Langlais, 2014],
— ou encore *prédiction de la structure sémantique* [Michalon, 2014].

Tous ces travaux traitent de ressources proches de FrameNet. Ce n'est pas le cas en anglais, où la littérature utilise aujourd'hui deux termes différents pour distinguer l'annotation en rôles sémantiques de type PropBank (*Semantic Role Labeling*) et l'annotation en rôles sémantiques de type FrameNet (*Frame-semantic parsing*). En effet, ce sont deux tâches relativement différentes.

— Pour PropBank, il s'agit d'identifier les arguments de chaque sens de verbe (ARG0, ARG1, etc.) sans avoir besoin de désambiguïser le sens du verbe [Carreras and Màrquez, 2005].
— Pour FrameNet, il faut identifier les prédicats (verbes mais aussi noms, adjectifs et adverbes), identifier la frame correspondante (tâche proche de la désambiguïsation lexicale), et ensuite, à la manière de PropBank, identifier les arguments et leurs rôles sémantiques. Pour montrer la différence avec le *semantic role labeling*, le terme *frame-semantic parsing* a été choisi [Das et al., 2010].

Dans ce sens, nos travaux sont effectivement du *frame-semantic parsing*, même si nous ne traitons que des verbes. En effet, la désambiguïsation entre classes VerbNet joue un rôle important dans notre système (Chapitre 5).

2.3.5. Adaptation au domaine

Chen and Mooney [2008] entraînent un système qui apprend à commenter un match de football en utilisant des commentaires existants et des simulations de jeux de football, mais sans connaissance explicite sur la langue anglaise. Leur approche a entraîné des travaux sur le *situated language understanding* (compréhension ancrée du langage) : Bordes et al. [2010], Richardson and Kuhn [2012] ont proposé par la suite d'autres corpus pour cette tâche. Chang et al. [2014] génèrent eux des scènes 3D à partir de textes tels que "Il y a une pièce avec une chaise et un ordinateur" en essayant d'inférer les contraintes implicites telles que la présence d'un bureau. Notre système (Chapitre 6) est similaire à ces systèmes dans le sens où nous minimisons l'effort humain pour annoter de nouveaux domaines, mais nous nous concentrons sur l'annotation en rôles sémantiques *à la FrameNet*.

Le système d'annotation en rôles sémantiques de Gormley et al. [2014] n'a pas besoin de corpus annoté en syntaxe, mais nécessite un corpus annoté en rôles sémantiques. Hadouche [2011] effectue une annotation en rôles sémantiques sur le corpus DicoInfo [OLST, 2014] à l'aide de deux approches :

— en appliquant des règles définies manuellement s'appliquant à la sortie d'un analyseur syntaxique,
— en apprenant un système supervisé en utilisant divers traits issus de la littérature.

Même si nous utilisons le même corpus, nos travaux vont dans la direction opposée : nous souhaitons annoter un grand nombre de phrases provenant de divers domaines sans utiliser de corpus annoté.

Ce travail conclut en indiquant que pour obtenir de meilleurs résultats sur plus de rôles et de prédicats, il faut plus d'exemples d'entraînement. Notre travail prend une autre direction : nous étudions l'utilisation de moins de données créées manuellement pour couvrir plus de phrases dans divers domaines.

En conclusion, les approches pour annoter un texte en rôles sémantiques sont nombreuses mais les difficultés restant à franchir pour annoter un texte français en cadre ouvert restent nombreuses. La partie suivante décrit les efforts réalisés pour adapter au français des ressources qui ont prouvé leur utilité en anglais.

Deuxième partie

Ressources pour l'annotation en rôles sémantiques

Nous avons choisi d'accorder dans ces travaux une part importante à l'utilisation de ressources lexicales. La première, WordNet, représente le sens des mots et les liens entre ces sens, alors que la seconde, VerbNet classifie les verbes selon leurs comportements syntaxiques. Cette partie présentes la traduction de ces deux ressources vers le français (produisant respectivement WoNeF et Verb∋Net). Quant à elle, la partie III traite de l'utilisation de ces ressources pour l'annotation en rôles sémantiques.

3. WoNeF : une traduction de WordNet

> « S'il vous plaît ! fis-je. Taisez-vous, s'il vous plaît ! » Je désirais seulement qu'ils fassent silence, qu'ils cessent les bruits qu'ils faisaient avec leur bouche, mais le son de mes propres paroles capta mon attention. « S'il-vous-plaît, répétai-je, en m'étonnant de tous les mouvements que devait effectuer ma propre bouche pour produire ces sons imprécis. « Taisez-vous ! » Je m'aperçus que ces deux mots avaient trop de sens pour en avoir un véritable.
>
> *(Robin Hobb, La Voie magique)*

Identifier les sens possibles des mots du vocabulaire est un problème difficile demandant un travail manuel très conséquent (section 2.1.1). Ce travail est cependant nécessaire pour pour pouvoir identifier le sens correct d'un mot utilisé en contexte. WordNet [Fellbaum, 1998] regroupe les mots en ensembles de synonymes (*synonym sets* ou *synsets*) et lie ces synsets entre eux par différentes relations sémantiques comme par exemple l'antonymie, l'hyperonymie, l'hyponymie ou encore la méronymie (section 2.1.2). Nous présentons ici une traduction de WordNet vers le français. Ce travail a été réalisé en collaboration avec Jeanne Baguenier Desormeaux. Une partie conséquente de ce chapitre provient d'articles déjà publiés [Pradet et al., 2013a, 2014b].

Nos travaux sont la suite de JAWS, dont le fonctionnement est détaillé à la section 2.2.1. Cette nouvelle version de JAWS se nomme WoNeF : le nouveau nom évite la confusion avec une API Java pour WordNet (*Java API for WordNet Searching*). Dans la suite de ce travail, le terme *JAWS* ne fera pas référence à l'API mais à la version précédente de WoNeF [Mouton and de Chalendar, 2010, Mouton, 2010].

Nous étendons et améliorons les techniques utilisées dans JAWS et l'évaluons à l'aide d'une vérité-terrain obtenue par adjudication de deux annotateurs. WoNeF se décline en trois versions pour répondre à différents besoins. Le WoNeF principal a un F-score de 70.9 %, une autre version a une précision de 93.3 %, et une dernière contient un plus grand nombre de paires (littéral, synset).

3.1. WoNeF : un JAWS amélioré et étendu

3.1.1. Limites de JAWS

JAWS souffre d'un certain nombre de limites. Avant tout, il ne contient que des noms alors que WordNet contient des noms, verbes, adjectifs et adverbes.

Ensuite, la façon dont il a été évalué rend difficile tout jugement sur sa qualité. En effet, JAWS a été évalué en le comparant à l'EuroWordNet du français et à WOLF 0.1.4 (qui date de 2008). Ces deux WordNets du français ne sont pas des annotations de références : ils souffrent soit d'une précision limitée soit d'une couverture limitée. Cette évaluation limitée a été complétée par une évaluation manuelle des littéraux n'existant pas dans WOLF, mais elle n'a été faite que sur 120 paires (littéral, synset). La précision de JAWS est évaluée à 67,1 % [Mouton, 2010], ce qui est plus bas que celle de WOLF 0.1.4 et considérablement plus bas que la précision de WOLF 1.0b [1]. Ce score est à prendre avec précaution étant donné la taille de l'échantillon de test : l'intervalle de confiance est d'environ 25 %. Une autre limite de JAWS est qu'il ne contient qu'une seule et unique ressource qui ne correspond pas à tous les besoins.

À notre connaissance, les traductions automatiques de WordNet actuelles n'existent qu'en une seule version. Soit les auteurs décident eux-mêmes quelle métrique optimiser, soit il faut choisir un seuil de confiance pour ne garder que les traductions les plus sûres. Nous fournissons aussi une telle version, mais ajoutons aussi deux ressources qui peuvent servir des besoins différents. Même si notre WoNeF à haute précision est petit, il peut être utilisé comme une annotation de référence et servir pour entraîner un système d'apprentissage. Une ressource à haute couverture peut servir de base à une correction manuelle ou servir pour une intersection avec d'autres ressources, ce qui est la raison pour laquelle nous en fournissons une aussi.

Cette section présente les trois améliorations essentielles qui ont étés apportées à JAWS. Un changement non détaillé ici est celui qui a mené à une meilleure rapidité d'exécution : JAWS se construit en plusieurs heures contre moins d'une minute pour WoNeF, ce qui a facilité les expérimentations.

3.1.2. Sélecteurs initiaux

Les sélecteurs initiaux de JAWS ne sont pas optimaux. Alors que les sélecteurs par monosémie et par unicité sont conservés, nous avons changé les autres sélecteurs. Premièrement, le sélecteur des transfuges est supprimé : sa précision était très basse, même pour les noms.

Deuxièmement, un nouveau sélecteur considère les traductions candidates provenant de plu-

1. Nous remercions Benoît Sagot pour nous avoir fourni cette version préliminaire de WOLF 1.0.

sieurs mots anglais différents dans un synset donné : c'est le sélecteur par sources multiples. Par exemple, dans le synset *line, railway line, rail line (the road consisting of railroad track and roadbed)*, les littéraux français *ligne de chemin de fer* et *voie* sont des traductions à la fois de *line* et *railway line*, et sont donc choisis comme traductions.

Troisièmement, le sélecteur de la distance de Levenshtein a été amélioré. 28 % du vocabulaire anglais est d'origine française [Finkenstaedt and Wolff, 1973], et l'anglicisation a produit des transformations prévisibles. Il est possible d'appliquer ces mêmes transformations aux littéraux candidats français, et seulement alors d'appliquer la distance de Levenshtein [2]. Nous commençons par supprimer les accents, puis appliquons différentes opérations (Table 3.1. Par exemple, l'inversion des lettres "r" et "e" prend en compte (*order*/*ordre*) et (*tiger*/*tigre*). Toutes les transformations ne s'appliquent qu'à la fin des mots : *-que* est transformé en *-k* ou *-c* (*marque* devient *mark*), *-té* vers *-ty* (*extremité* devient *extremity*), etc. Les faux-amis ne peuvent pas être détectés puis exclus par cette méthode qui est purement orthographique.

-que	-k	banque	→	bank
-aire	-ary	tertiaire	→	tertiary
-eur	-or	chercheur	→	cherchor
-ie	-y	cajolerie	→	cajolery
-té	-ty	extremité	→	extremity
-re	-er	tigre	→	tiger
-ais	-ese	libanais	→	libanese
-ois	-ese	chinois	→	chinese
-ant	-ing	changeant	→	changeing
-er	-	documenter	→	document
-ose	-osis	osmose	→	osmosis
-ment	-ly	confortablement	→	confortably

TABLE 3.1. – Règles appliquées à la fin des mots français avant d'utiliser le sélecteur de la distance de Levenshtein.

Il serait possible d'améliorer cette table des correspondances en réutilisant les correspondances listées dans le Wiktionnaire français : pour chacun des suffixes du français, une liste de suffixes correspondants dans d'autres langues (dont l'anglais) est parfois indiquée, par exemple pour -ment : `http://fr.wiktionary.org/wiki/-ment`.

3.1.3. Apprentissage de seuils

Dans JAWS, chaque littéral anglais ne peut avoir qu'une traduction française correspondante. La traduction choisie est celle qui a le meilleur score, indépendamment des scores des traductions

2. La distance de Damerau-Levenshtein qui prend en compte les inversions n'importe-où dans un mot [Damerau, 1964] a donné de moins bons résultats.

moins bien notées. Cela a pour effet de rejeter des candidats valides et d'accepter des candidats erronés. Par exemple, JAWS n'inclut pas *particulier* au synset *(a human being) "there was too much for one person to do"* parce que *personne* est déjà inclus avec un score supérieur.

Dans WoNeF, nous avons donc appris un seuil pour chaque partie du discours et sélecteur. Nous avons d'abord généré les scores pour toutes les paires (littéral, synset) candidates, puis trié ces paires par score. Les 12 399 paires présentes dans l'évaluation manuelle associée à WOLF 1.0b (notre ensemble d'apprentissage) ont été jugées correctes tandis que les paires n'y étant pas ont été jugées erronées. Nous avons ensuite calculé les seuils maximisant la précision et le F-score. Le seuil qui maximise le F-score est utilisé dans les ressources à haut F-score et à haute couverture, tandis que le seuil maximisant la précision est utilisé dans la ressource à haute précision.

Une fois que ces seuils sont définis, les sélecteurs choisissent tous les candidats au-dessus du nouveau seuil, ce qui a deux effets positifs :

— des candidats valides ne sont plus rejetés simplement parce qu'un meilleur candidat est aussi sélectionné, ce qui améliore à la fois le rappel et la couverture ;
— les candidats invalides qui étaient jusque-là acceptés sont maintenant rejetés grâce au seuil plus strict : la précision s'en retrouve augmentée.

3.1.4. Vote

Après l'application des différents sélecteurs, notre WordNet est large mais contient des synsets bruités. Comme toutes les traductions automatiques de WordNet, WoNeF doit alors être nettoyé [Sagot and Fišer, 2012b]. Dans WoNeF, le bruit provient de différents facteurs.

— Les sélecteurs essaient d'inférer des informations sémantiques à partir d'une analyse syntaxique sans prendre en compte toute la complexité de l'interface syntaxe-sémantique.
— L'analyseur syntaxique produit lui-même des résultats bruités.
— Le modèle de langue syntaxique est produit à partir d'un corpus extrait du web lui-même bruité (texte mal écrit, contenu non textuel, phrases non françaises) et n'est pas une « distribution idéale » [Copestake and Herbelot, 2013].
— Les traductions déjà choisies sont considérées comme valides dans les étapes suivantes alors que ce n'est pas toujours le cas.

Pour la ressource haute-précision, il fallait donc un moyen de ne garder que les littéraux pour lesquels les sélecteurs étaient les plus confiants. Étant donné que, contrairement à JAWS, plusieurs sélecteurs peuvent choisir une même traduction (sous-section 3.1.3), notre solution est simple et efficace : les traductions validées par un bon sélecteur ou par plusieurs sélecteurs moyens sont conservées tandis que les autres sont supprimées. Ce principe de vote est aussi appelé méthode d'ensemble en apprentissage automatique. Les sélecteurs performants varient d'une partie du

discours à une autre : le choix est fait sur un ensemble de développement contenant 10 % de notre référence.

Cette opération de nettoyage ne conserve que 18 % des traductions (de 87 757 paires (littéral, synset) à 15 625) mais la précision grimpe de 68,4 % à 93,3 %. Cette ressource à haute précision peut être utilisée comme donnée d'entraînement. Un défaut classique des méthodes de vote est de ne choisir que des exemples faciles et peu intéressants, mais la ressource obtenue ici semble être équilibrée entre les synsets ne contenant que des mots monosémiques et d'autres synsets contenant des mots polysémiques et plus difficiles à désambiguïser (section 3.3.2).

3.1.5. Extension aux verbes, adjectifs et adverbes

Les travaux sur JAWS ont commencé par les noms parce qu'ils représentent 70 % des synsets dans WordNet. Nous avons continué ce travail sur les autres parties du discours qui sont aussi importantes pour examiner le sens d'un texte donné : verbes, adjectifs et adverbes. Les sélecteurs génériques ont ici été modifiés, mais il s'agira dans le futur d'implémenter des sélecteurs prenant en compte les spécificités des différentes parties du discours dans WordNet.

Verbes Les sélecteurs choisis pour les verbes sont le sélecteur par unicité et par monosémie. En effet, la distance de Levenshtein a donné des résultats médiocres pour les verbes : seuls 25 % des verbes choisis par ce sélecteur étaient des traductions correctes. Concernant les sélecteurs syntaxiques, seul le sélecteur par synonymie a donné de bons résultats, alors que le sélecteur par hyponymie avait les performances d'un classifieur aléatoire.

Adjectifs Les adjectifs sont traduits de la même manière que les noms : tout d'abord un nombre limité de sélecteurs initiaux remplit un WordNet vide, puis les sélecteurs syntaxiques complètent cette traduction avec le modèle de langue syntaxique. Tous les sélecteurs initiaux sont ici choisis, et le sélecteur syntaxique choisi est le sélecteur par synonymie. Ils ont donné de bons résultats qui sont présentés dans la section 3.3.3.

Adverbes Les adverbes sont traduits avec les quatre sélecteurs : par monosémie, par unicité, par sources multiples et enfin par la distance de Levenshtein. Dans le cas des adverbes, nous n'avons pas trouvé de cas où la granularité des synsets de WordNet n'était pas applicable en français, comme c'était le cas pour les autres parties du discours. L'accord inter-annotateur était de 0.57, ce qui peut s'expliquer par le fait qu'un des deux annotateurs n'avait jamais effectué cette tâche, alors que l'autre avait déjà annoté les autres parties du discours. Les adverbes obtenus avec notre approche sont de bonne qualité tout en étant complémentaires avec WOLF : 87 % des

adverbes proposés ne sont pas dans WOLF. Une fusion entre WoNeF et WOLF aurait trois fois plus d'adverbes que WOLF seul.

La section 2 décrit en annexe l'ensemble des sélecteurs utilisés dans WoNeF.

3.2. WoNeF : un JAWS évalué

3.2.1. Développement d'une annotation de référence

L'évaluation de JAWS souffre d'un certain nombre de limites (section 3.1.1). Pour évaluer rigoureusement notre propre traduction de WordNet, nous avons produit une annotation de référence. Pour chaque partie du discours, 300 synsets ont été annotés par deux annotateurs locuteurs natifs du français. Pour chaque traduction candidate fournie par nos dictionnaires, il fallait décider si elle appartenait au synset. Puisque les dictionnaires ne proposent pas de candidats pour tous les synsets et que certains synsets n'ont pas de candidat valable, le nombre réel de synsets non vides est inférieur à 300 (section 3.2.2).

Durant l'annotation manuelle, nous avons rencontré une difficulté importante découlant de la tentative de traduire WordNet dans une autre langue. Dans le cas de l'anglais vers le français, la plupart des difficultés proviennent des verbes et adjectifs figurant dans une collocation. Dans WordNet, ils peuvent être regroupés d'une manière qui fait sens en anglais, mais qui ne se retrouve pas directement dans une autre langue. Par exemple, l'adjectif *pointed* est le seul élément d'un synset défini comme *direct and obvious in meaning or reference ; often unpleasant ; "a pointed critique" ; "a pointed allusion to what was going on" ; "another pointed look in their direction"*. Ces exemples se traduiraient par trois adjectifs différents en français : *une critique dure*, *une allusion claire* et *un regard appuyé*. Il n'existe pas de solution satisfaisante lors de la traduction d'un tel synset : le synset résultant contiendra soit trop soit trop peu de traductions. Nous avons décidé de ne pas traduire ces synsets dans notre annotation manuelle. Ces problèmes de granularité concernent 3 % des synsets nominaux, 8 % des synsets verbaux et 6 % des synsets adjectivaux. Actuellement, WoNeF n'essaie pas de détecter de tels synsets et les traduit donc comme tous les autres.

L'autre difficulté principale découle de traductions manquantes, ce qui peut être considéré comme un défaut de nos dictionnaires. Les sens rares d'un mot sont parfois absents. Par exemple, le sens *to catch* du jeu du chat (ou du loup) et le sens *coat with beaten egg* du verbe *to egg* ne sont pas présents. Aucun de ces sens n'est dans les synsets les plus polysémiques (définis à la section 3.3.2), ce qui confirme que cela ne se produit que pour les sens rares. Pourtant, WoNeF pourrait être amélioré en utilisant des dictionnaires spécifiques pour, par exemple, les espèces (comme dans Sagot and Fišer [2008a]), les termes médicaux, les entités nommées (en utilisant Wikipédia) et ainsi de suite. Un autre exemple est celui des adjectifs de jugement : il n'y a pas de bonne traduction de *weird* en français. Même si la plupart des dictionnaires fournissent *bizarre*

comme traduction, on ne retrouve pas dans *bizarre* l'aspect *stupide* du mot *weird* : les deux adjectifs ne sont pas substituables dans tous les contextes, ce qui est un problème si l'on considère que le sens d'un synset doit être conservé par la traduction.

3.2.2. Accord inter-annotateurs

Malgré les difficultés mentionnées ci-dessus, l'annotation résultante a été validée par la mesure de l'accord inter-annotateurs, qui montre que l'approche par extension pour la création de nouveaux wordnets est valide et peut produire des ressources utiles. Trois annotateurs humains (deux par partie du discours), soit linguiste informaticien soit informaticien linguiste, ont annoté de façon indépendante les mêmes synsets choisis au hasard pour chaque partie du discours. Ils ont utilisé WordNet pour examiner les synsets voisins, le dictionnaire Merriam-Webster, le TLFi [Pierrel, 2003] et des moteurs de recherche pour attester l'utilisation des divers sens des mots considérés. Après adjudication faite par ces deux annotateurs en confrontant leurs opinions en cas de désaccord, l'annotation de référence a été formée.

	Noms	Verbes	Adjectifs	Adverbes
Kappa de Fleiss	0.72	0.71	0.66	0.57
Synsets non-vides	270	222	267	300
Candidats par synset	6.22	14.50	7.27	5.17

TABLE 3.2. – Accord inter-annotateurs sur l'annotation de référence. Pour les adverbes, nous avons tiré 300 synsets *non-vides*, ce qui explique pourquoi c'est la seule partie du discours avec 3000 synsets non-vides.

La Table 3.2 montre l'accord inter-annotateur évalué par le kappa de Fleiss pour les trois parties du discours annotées. Même s'il s'agit d'une métrique discutée [Powers, 2012], toutes les tables d'évaluation existantes considèrent ces scores comme étant suffisamment élevés pour décrire cet accord inter-annotateurs comme « bon » [Gwet, 2001], ce qui nous permet de dire que notre annotation de référence est de bonne qualité. L'approche par extension pour la traduction de WordNet est elle aussi validée.

3.3. Résultats

Nous présentons dans cette section les résultats de WoNeF. Nous commençons par décrire les résultats après l'application de l'étape des sélecteurs initiaux seulement puis ceux de la ressource complète. Notre annotation de référence est découpée en deux parties : 10 % des littéraux forment l'ensemble de développement utilisé pour choisir les sélecteurs s'appliquant aux différentes versions de WoNeF, tandis que les 90 % restant forment l'ensemble de test servant à

l'évaluation. Précision et rappel sont calculés sur l'intersection des synsets présents dans WoNeF et l'annotation de référence considérée, que ce soit l'ensemble de test de notre propre adjudication (sections 3.3.1 à 3.3.3) ou WOLF (section 3.3.4). Par exemple, la précision est la fraction des paires (littéral, synset) correctes au sein de l'intersection en question.

3.3.1. Sélecteurs initiaux

Pour les noms, les verbes et les adjectifs, nous avons calculé l'efficacité de chaque sélecteur initial sur notre ensemble de développement, et utilisé ces données pour déterminer ceux qui doivent être inclus dans la version ayant une haute précision, celle ayant un F-score élevé et celle présentant une grande couverture. Les scores ci-dessous sont calculés sur l'ensemble de test, plus grand et plus représentatif.

	P	R	F1	C
monosémie	71.5	76.6	74.0	54 499
unicité	91.7	63.0	75.3	9 533
sources multiples	64.5	45.0	53.0	27 316
Levenshtein	61.9	29.0	39.3	20 034
haute précision	**93.8**	50.1	65.3	13 867
haut F-score	71.1	**72.7**	**71.9**	82 730
haute couverture	69.0	69.8	69.4	**90 248**

TABLE 3.3. – Sélecteurs initiaux sur l'ensemble des traductions (noms, verbes et adjectifs). La couverture C est le nombre total de paires (littéral, synset).

La Table 3.3 montre les résultats de cette opération. La couverture donne une idée de la taille des ressources. En fonction des objectifs de chaque ressource, les sélecteurs initiaux choisis seront différents. Différents sélecteurs peuvent choisir plusieurs fois une même traduction, ce qui explique que la somme des couvertures soit supérieure à la couverture de la ressource à haute couverture. Fait intéressant non visible dans la table, le sélecteur le moins efficace pour les verbes est la distance de Levenshtein avec une précision de l'ordre de 25 % : les faux amis semblent être plus nombreux pour les verbes.

3.3.2. Résultats globaux

Nous nous intéressons maintenant aux résultats globaux (Table 3.4). Ils comprennent l'application des sélecteurs initiaux et des sélecteurs syntaxiques. Le mode haute précision applique également un vote (section 3.1.4). Comme pour la table précédente, la couverture C indique le nombre de paires (littéral, synset).

	Tous synsets				Synsets BCS			
	P	R	F1	C	P	R	F1	C
haute précision	**93.3**	51.5	66.4	15 625	**90.4**	36.5	52.0	1 877
haut F-score	68.9	73.0	**70.9**	88 736	56.5	62.8	**59.1**	14 405
haute couverture	60.5	**74.3**	66.7	**109 447**	44.5	**66.9**	53.5	**23 166**

TABLE 3.4. – Résultats globaux : tous les synsets et synsets BCS.

Dans WordNet, les mots sont majoritairement monosémiques, mais c'est une petite minorité de mots polysémiques qui est la plus représentée dans les textes. C'est justement sur cette minorité que nous souhaitons produire une ressource de qualité. Pour l'évaluer, nous utilisons la liste des synsets **BCS** (Basic Concept Set) fournie par le projet BalkaNet [Tufiş et al., 2004]. Cette liste contient les 8 516 synsets lexicalisés dans six traductions différentes de WordNet, et représente les synsets les plus fréquents et ceux qui comportent le plus de mots polysémiques. Alors que les ressources à haut F-score et à haute couverture perdent en précision pour les synsets BCS, ce n'est pas le cas pour la ressource à haute précision. En effet, le mécanisme de vote rend la ressource haute-précision très robuste, et ce même pour les synsets BCS.

3.3.3. Résultats par partie du discours

		P	R	F1	C
haute précision	noms	**96.8**	56.6	71.4	11 294
	verbes	**68.4**	41.9	52.0	1 110
	adjectifs	**90.0**	36.7	52.2	3 221
	adverbes	**92.2**	32.4	48.8	968
haut F-score	noms	71.7	73.2	**72.4**	59 213
	JAWS	70.7	68.5	69.6	55 416
	verbes	48.9	**76.6**	**59.6**	9 138
	adjectifs	69.8	71.9	**70.8**	20 385
	adverbes	82.1	74.6	**78.2**	5 950
haute couverture	noms	61.8	**78.4**	69.1	**70 218**
	verbes	45.4	61.5	52.2	**18 844**
	adjectifs	69.8	**71.9**	70.8	**20 385**
	adverbes	82.1	**74.6**	78.2	**5 950**

TABLE 3.5. – Résultats par partie du discours. JAWS ne contient que des noms : il est comparé à la ressource nominale à haut F-score. Les scores sont les mêmes pour les adjectifs et les adverbes pour les ressources à haut F-score et à haute couverture : les mêmes sélecteurs ont étés appliqués.

La Table 3.5 montre les résultats détaillés pour chaque partie du discours. Concernant les noms,

le mode de haute précision utilise deux sélecteurs, tous deux fondés sur la relation syntaxique de complément du nom : le sélecteur par méronymie décrit à la section 2.2.1, et le sélecteur par hyponymie. La ressource de haute précision pour les noms est notre meilleure ressource. La version avec le F-score optimisé a un F-score de 72,4 %, ce qui garantit que peu de paires (littéral, synset) sont absentes tout en ayant une précision légèrement supérieure à celle de JAWS.

Les résultats des verbes sont moins élevés. L'explication principale est que les verbes sont en moyenne plus polysémiques dans WordNet et nos dictionnaires que les autres parties du discours : les synsets verbaux ont deux fois plus de candidats que les noms et les adjectifs (Table 3.2). Cela montre l'importance du dictionnaire pour limiter le nombre initial de littéraux parmi lesquels les algorithmes doivent choisir.

Le sélecteur par synonymie est le seul sélecteur syntaxique appliqué aux verbes. Il utilise les relations syntaxiques de second ordre pour trois types de dépendances syntaxiques verbales : si deux verbes partagent les mêmes objets, ils sont susceptibles d'être synonymes ou quasi-synonymes. C'est le cas des verbes *dévorer* et *manger* qui acceptent tous deux l'objet *pain*. Les autres sélecteurs syntaxiques n'ont pas été retenus pour les verbes en raison de leurs faibles résultats. En effet, alors que la détection de l'hyponymie en utilisant seulement l'inclusion de contextes a été efficace sur les noms, elle a les performances d'un classifieur aléatoire pour les verbes. Cela met en évidence la complexité de la polysémie des verbes.

Pour les adjectifs, comme pour les verbes, seul le sélecteur de synonymie a été appliqué. Pour les ressources à haut F-score et haute couverture, ce sont les mêmes sélecteurs (initiaux et syntaxiques) qui sont appliqués, ce qui explique que les résultats sont les mêmes. Alors que l'accord inter-annotateurs était plus bas sur les adjectifs que sur les verbes, les résultats eux sont bien meilleurs pour les adjectifs. Cela s'explique principalement par le nombre de candidats parmi lesquels sélectionner : il y en a deux fois moins pour les adjectifs. Cela met en avant l'importance des dictionnaires.

3.3.4. Évaluation par rapport à WOLF

	WOLF 0.1.4			WOLF 1.0b		
	pP	pR	Ajouts	pP	pR	Ajouts
Noms	50.7	40.0	9 646	73.6	46.4	6 842
Verbes	33.0	23.9	1 064	41.7	17.5	1 084
Adjectifs	41.7	46.1	3 009	64.4	53.8	3 172
Adverbes	56.2	44.4	3 061	76.5	41.9	2 835

TABLE 3.6. – Évaluation de la ressource à haute précision en considérant WOLF 0.1.4 et 1.0b comme des références.

Il n'est pas possible de comparer WOLF et WoNeF en utilisant notre annotation de référence :

tout mot correct de WOLF non présent dans les dictionnaires pénalisera WOLF injustement. Nous avons décidé d'évaluer WoNeF en considérant WOLF 0.1.4 et WOLF 1.0b comme des références (Table 3.6). Les mesures ne sont pas de véritables précision et rappel puisque WOLF lui-même n'est pas entièrement validé. Le dernier article donnant des chiffres globaux [Sagot and Fišer, 2012a] dont nous avons connaissance indique un nombre de paires autour de 77 000 pour une précision de 86 %[3]. Nous appelons donc pseudo-précision (pP) le pourcentage des éléments présents dans WoNeF qui sont également présents dans WOLF, et pseudo-rappel le pourcentage d'éléments de WOLF qui sont présents dans WoNeF. Ces chiffres montrent que même si WoNeF est encore plus petit que WOLF, il s'agit d'une ressource complémentaire, surtout quand on se souvient que le WoNeF utilisé pour cette comparaison est celui présentant une précision élevée, avec une précision globale de 93,3 %. Il convient également de noter que la comparaison de la différence entre WOLF 0.1.4 et WOLF 1.0b est instructive puisque elle montre l'étendue des améliorations apportées à WOLF.

La colonne « Ajouts » donne le nombre de traductions qui sont présentes dans WoNeF mais pas dans WOLF. Pour les noms, les verbes et les adjectifs, cela signifie que nous pouvons contribuer 11 098 nouvelles paires (littéral, synset) de haute précision en cas de fusion de WOLF et WoNeF, soit 94 % des paires du WoNeF haute précision ce qui montre la complémentarité des approches : ce sont des littéraux différents qui sont ici choisis. Cela produira un wordnet français 13 % plus grand que WOLF avec une précision améliorée. Une fusion avec la ressource de F-score élevée aurait une précision légèrement inférieure, mais fournirait 57 032 nouvelles paires (littéral, synset) par rapport à WOLF 1.0b, résultant en une fusion contenant 73 712 synsets non vides et 159 705 paires (littéral, synset), augmentant la couverture de WOLF de 56 % et celle de WoNeF de 83 %.

3.4. Fusion entre WOLF et WoNeF

WOLF et WoNeF ne sont pas en compétition. En réalité, les travaux sur WoNeF et WOLF ont commencé indépendamment sans que les auteurs ne soient au courant des travaux de l'autre équipe. Certes, les deux ressources semblent complémentaires (section 3.3.4), mais pourquoi est-ce que les deux traductions existent toujours en 2014 alors que les premières publications datent de 2008 et 2009 ? Une fusion a en réalité été envisagée en 2014, mais elle n'a pas aboutie : les auteurs de WoNeF et WOLF n'ont pas encore une vision commune claire de ce que constitue une traduction correcte dans un synset. Deux problèmes ont étés identifiés pour l'instant.

Le premier problème, et le plus simple, est du à un biais lexicographique particulièrement présent dans WoNeF. Par exemple, WordNet contient un synset pour l'adverbe *unkindly#1* qui est traduit dans nos dictionnaires par *sans aménité*. Ceci pose un problème parce que *sans aménité* n'est pas une unité lexicale qu'on s'attend à trouver dans un dictionnaire monolingue étant donné que le sens est déductible par compositionnalité. Ce problème est accru en traduisant l'anglais, où

3. Les résultats détaillés pour WOLF 1.0b ne sont pas actuellement disponibles.

de nombreux mots sont formés par concaténation. Ceci dit, étant donné qu'il n'y a pas d'espace entre *un* et *kindly* dans *unkindly*, il semble logique de l'inclure dans WordNet, même si on peut aussi déduire son sens par compositionnalité.

Le second problème est du à la finesse des sens proposés par WordNet. Bien que les évaluations autour de WoNeF aient mené à un nombre conséquent de débats passionnés sur le sens des mots et l'adéquation des traductions, ce n'est qu'en évaluant la fusion de WOLF et WoNeF que nous nous sommes rendus compte en comparant l'évaluation que WOLF avait tendance à être plus strict au moment d'accepter un synset.

Ces difficultés mettent en évidence le besoin crucial d'un guide d'annotation basé sur le résultat de différentes adjudications avant de poursuivre les travaux autour de la fusion. En effet, sans cet étalon, il est très difficile de donner du sens aux mesures d'évaluation de WoNeF, WOLF, et des autres traductions en général.

Le lecteur qui considère que ces problèmes ne l'affectent pas peut utiliser la fusion des deux ressources disponible sur `http://wonef.fr/data/`.

Conclusion

Dans ce chapitre, nous avons montré que l'utilisation d'un modèle de langue syntaxique pour identifier des relations lexicales entre des lexèmes est possible dans un environnement contraint et conduit à des résultats ayant une précision au niveau de l'état de l'art pour la tâche de traduction de WordNet. Nous offrons trois ressources différentes, chacune d'elles ayant un objectif différent. Enfin, nous fournissons une annotation de référence validée de haute qualité qui nous a permis de montrer à la fois la validité de l'approche de traduction de WordNet par extension et la validité de notre approche spécifique. Cette annotation de référence peut également être utilisée pour évaluer et développer d'autres traductions françaises de WordNet. WoNeF est disponible librement au format XML DEBVisDic [4] sur `http://wonef.fr/data/` sous la licence CC-BY-SA.

Les travaux futurs sur WoNeF mettront l'accent sur les verbes, les adjectifs et les adverbes, pour lesquels de nouveaux sélecteurs efficaces peuvent être envisagés pour améliorer la couverture. Par exemple, le sélecteur de similarité peut être étendu à la relation de quasi-synonymité que partagent certains adjectifs dans WordNet. En effet, la synonymie entre les adjectifs est limitée par rapport à la quasi-synonymie : alors que *fast* est le seul mot dans son synset, c'est le quasi-synonyme de 20 synsets. Puisque les techniques de sémantique distributionnelle ont plutôt tendance à identifier des quasi-synonymes plutôt que des synonymes, utiliser cette relation de WordNet pour identifier de nouveaux adjectifs fait partie de nos objectifs.

4. `http://nlp.fi.muni.cz/trac/deb2/wiki/WordNetFormat`

Une autre source importante d'amélioration sera l'enrichissement de notre modèle de langue syntaxique qui pourra prendre en compte les verbes pronominaux et les expressions multi-mots. Nous aimerions aussi nous orienter vers un modèle de langue continu [Le et al., 2012] plus performant. Cela sera couplé avec la collecte d'un corpus issu du Web plus récent et plus grand analysé avec une version récente de notre analyseur linguistique LIMA. Cela nous permettra de mesurer l'impact de la qualité du modèle de langue sur la traduction de WordNet.

4. Verb∋Net : une traduction de VerbNet

> Je levai les yeux vers leurs visages effrayés et me rendit compte qu'ils parlaient – non, ils hurlaient presque ; [...] quelqu'un demandait sans cesse : « Que s'est-il passé ? Que s'est-il passé ? » Je fus soudain frappé du manque de grâce de la parole : tous ces mots rattachés les uns aux autres, que chaque bouche prononçait différemment... Et c'était ainsi que nous communiquions ?
>
> *(Robin Hobb, La Voie magique)*

La ressource VerbNet (section 1.3.3) rend possible l'annotation en rôles sémantiques selon nos objectifs présentés à la section 1.2. Or, nous désirons réaliser cette annotation en français mais VerbNet n'est disponible que pour l'anglais. C'est pourquoi nous présentons dans ce chapitre sa traduction vers le français, nommée Verb∋Net [1]. Une partie conséquente de ce chapitre provient d'articles déjà publiés [Danlos et al., 2014, Pradet et al., 2014a]. Nous avons réalisé la première étape de traduction en collaboration avec Laurence Danlos. Ensuite, Laurence Danlos et Takuya Nakamura ont réalisé seuls le travail lexicographique, c'est-à-dire les deuxième et troisième étapes. Pour faciliter ce travail, nous avons développé et activement maintenu une interface d'édition entièrement pensée pour VerbNet afin de faciliter le travail lexicographique. Cette interface (section 4.4) est disponible sur `https://verbenet.inria.fr/`.

La construction de cette ressource se fait en trois étapes.

1. Nous avons d'abord traduit vers le français les membres de VerbNet en utilisant deux ressources linguistiques qui encodent des informations syntaxiques et sémantiques sur les verbes du français (section 4.1.1).
2. La deuxième étape (section 4.1.2), toujours en cours, est l'adaptation des frames VerbNet vers le français et la réorganisation des classes VerbNet en fonction de ces frames. Chaque classe est plus ou moins difficile à adapter (section 4.2) : nous présentons certaines difficultés récurrentes à la section 4.3.2.
3. La troisième étape (section 4.1.3), aussi en cours de réalisation, est la validation des verbes français de chaque classe.

Grâce aux mises aux correspondance réalisées lors de la première étape, notre traduction de VerbNet est liée aux deux ressources linguistiques que nous utilisons, Les Verbes Français et

1. Le nom vient de la prononciation à la française qui "rajoute" un *e*. Le ∋ est là pour bien marquer ce *e* et ainsi essayer d'éviter la confusion avec VerbNet.

le Lexique-Grammaire. La ressources est aussi ouverte : nous voulons encourager les contributions externes avec notre outil basé sur le web. Nous souhaitons enfin faciliter l'utilisation de la ressource en utilisant le même format XML que le VerbNet anglais et en rendant Verb∋Net librement réutilisable sous la licence CC-BY-SA, qui autorise notamment les usages commerciaux.

4.1. Étapes de constructions de Verb∋Net

L'idée directrice est que la hiérarchie de Verb∋Net doit être aussi proche que possible de celle de VerbNet. Néanmoins, certaines classes peuvent disparaître dès lors que les critères utilisés pour l'anglais ne s'appliquent pas au français, ce qui est parfois le cas lorsque VerbNet emploie des critères morphologiques. Ainsi, une classe VerbNet ne contenant que des verbes dénominaux n'a pas d'équivalent en français. C'est le cas de :

- `pit-10.7` formé avec des parties constitutives d'animaux ou d'objets avec des verbes tels que :
  - *peel* : enlever la peau (*peel*) de certains fruits et légumes,
  - *bark* : enlever l'écorce (*bark*) d'un arbre ou d'une buche,
  - ou *bone* : enlever les os (*bone*) d'un poisson,
- ou encore de `weekend-56` formée à partir de verbes issus de noms décrivant une période de temps tels que :
  - *vacation* : *Luc Leclerc vacationed with Zambito in Mexico*,
  - ou *summer* : *My family always summered at the seashore*.

Par contre, `debone-10.8` et ses verbes formés par le préfixe *de-* plus un nominal (*debark, debone*) a un équivalent français avec les préfixes *dé-* ou *é* (*déveiner*, *équeuter*) : cette classe est donc conservée dans Verb∋Net

Pour toutes les classes conservées, la construction de Verb∋Net se fait en trois étapes.

4.1.1. Première étape : traduction des verbes

La première étape pour construire Verb∋Net est de déterminer quels verbes français appartiennent à une des 270 classes de VerbNet. Voici comment nous avons procédé.

1. Pour une classe VerbNet donnée C_e, nous assignons manuellement les classes LVF (section 1.3.6) C_{lvf} qui correspondent à sa définition sémantique et les classes LG (section 1.3.6) C_{lg} qui correspondent à sa définition syntaxico-sémantique. Par exemple, pour `put-9.1` (*put an entity at some location*), nous assignons L3b (mettre quelque chose

quelque part) et 38LD (LD pour Location Destination, définie par N_0 V N_1 Prép N_2, Prép étant une préposition de location et N_2 étant un lieu).

2. Nous utilisons deux dictionnaires bilingues (SCI-FRAND-EURADIC et le Wiktionnaire) qui fournissent la liste L_{trad} des traductions françaises des verbes anglais appartenant à C_e et ses sous-classes.
3. Les verbes français de la classe C_e sont alors simplement les verbes appartenant à la fois à L_{trad}, C_{lvf} et C_{lg} [2]. Par exemple, *mettre*, *poser* et *installer* sont des verbes français de `put-9.1`).

Cette étape a été exécutée rapidement et a donné des résultats encourageants : en ne gardant que les verbes à l'intersection de L_{trad}, C_{lvf} et C_{lg}, les résultats sont souvent précis et cohérents d'un point de vue sémantique et syntaxique, ce que nous verrons par exemple à la section 4.2.1. Cette méthode a produit un lexique de 4058 verbes (2128 verbes distincts)

4.1.2. Deuxième étape : adaptation des *frames*

Malgré le succès de la première étape, la deuxième étape de la construction de Verb∋Net a demandé un travail beaucoup plus méticuleux que la première parce qu'il a fallu détailler davantage la définition des classes dans les trois ressources : VerbNet, le Lexique-Grammaire, et Les Verbes Français. Pour chacune des 500 classes et sous-classes, nous déterminons quand cela est possible :

— les frames valides pour le français avec des ajustements possibles pour les rôles thématiques et leurs restrictions de sélection,
— les sous-classes en français en réorganisant si besoin la hiérarchie de l'anglais dans le but d'assigner les verbes obtenus à l'étape 1 à une des sous-classes.

Cette étape a demandé de définir des principes de base sur les frames françaises quand elles diffèrent des frames anglaises (section 4.3.1). Une étude fine au cas par cas révèle que certaines de ces différences entre les deux langues sont difficiles à traiter (section 4.3.2). Il a aussi fallu développer un outil d'édition entièrement spécialisé pour l'édition de VerbNet afin d'assister au maximum le travail lexicographique (section 4.4).

4.1.3. Troisième étape : validation manuelle des verbes

La troisième étape est l'occasion de valider manuellement pour chaque classe les verbes proposés par correspondance de ressources en supprimant les verbes erronés et en rajoutant les verbes manquants afin que la ressource ait été entièrement validée manuellement. La plupart des

2. Quand cette intersection est vide, c'est la liste non-vide (soit C_{lvf} soit C_{lg}) qui est choisie

verbes ne sont pas ajoutés manuellement : les annotateurs cliquent directement sur les verbes inférés pour les valider ou les invalider. Les traductions des dictionnaires étant très productives, les verbes souhaités sont généralement dans les traductions proposées, même s'ils peuvent être absent des correspondances LG, LVF ou les deux. Dans les rares cas où un verbe souhaité n'est pas trouvé (en général parce qu'une classe a été créée sans verbe anglais), il est tout de même possible d'en ajouter un.

4.2. Adaptation pas à pas de deux classes d'exemple

Cette section est l'occasion de présenter l'adaptation de VerbNet par l'exemple, à la fois pour mettre en avant les facilités et difficultés du processus mais aussi pour rendre plus concrètes les discussions de ce chapitre aux lecteurs ne connaissant pas VerbNet.

4.2.1. Une classe ne nécessitant que peu de modifications : `scribble-25.2`

La classe `scribble-25.2` contient 18 verbes en anglais : elle est associée à la classe LVF R3a.1 (*faire quelque chose, un objet*) et à la table LG 32A (A pour Apparition, définie par N_0 V N_1, N_1 devant apparaître comme dans *Max a bâti une maison*), ce qui conduit à une liste de 16 verbes français : *composer*, *couper*, *donner*, *exécuter*, *fabriquer*, *faire*, *forger*, *former*, *imprimer*, *lever*, *produire*, *reproduire*, *sculpter*, *tailler*, *tirer* et *tracer*. Tous ces verbes sont valides pour cette classe.

Cette classe n'ayant pas de sous-classe et la seule frame ayant du sens en français, il n'y a pas eu de travail de réorganisation des frames.

Enfin, la troisième étape de validation des verbes a consisté à valider les verbes corrects présents uniquement dans les tables du Lexique-Grammaire ou Les Verbes Français.

4.2.2. Des classes réorganisées : `run-51.3.2` et dérivées

Les verbes de la classe 51 sont les verbes de *Manner of motion* qui décrivent la façon dont un mouvement a été effectué :

- `run-51.3.2` décrit les façons dont les objets animés se déplacent, avec des propriétés telles que la *Induced Action Alternation* partagée par quelques verbes :
  - The horse jumped over the fence.
  - Tom jumped the horse over the fence.
- Les autres classes que nous considérons ici sont définies sur des critères morphologiques.

— `vehicle-51.4.1` contient des verbes qui sont des dénominaux de moyens de déplacements tels que *skate*, *parachute* ou *taxi*.
— `nonvehicle-51.4.2` contient des verbes qui ne sont pas des dénominaux mais qui décrivent aussi un déplacement avec un type de véhicule particulier, par exemple *row*, *paddle* ou *fly*.
— Enfin, `waltz-51.5` regroupe des verbes issus de noms décrivant des danses (*waltz*, *samba* ou encore *jive*).

Ces autres classes étant définies sur des critères morphologiques en français, tous leurs verbes sont envoyés vers la classe `run-51.3.2`, ce qui amène de nouvelles traductions qui ont été validées ou non lors de la troisième étape.

Enfin, d'autres classes doivent être réorganisées complètement : c'est le cas par exemple des super-classes 13 ou 22. Nous ne décrivons pas en détail le processus : il s'agit d'identifier les constructions communes à certains verbes pour proposer des découpages cohérents.

La section suivante propose des principes généraux adoptés lors de la réalisation de la deuxième étape.

4.3. Principes adoptés pendant l'adaptation

4.3.1. Principes sur les frames

Nous avons jusqu'ici identifié deux différences principales entre le codage des frames anglaises et françaises.

La première différence concerne les sous-structures, c'est-à-dire les frames avec un complément manquant tel que *NP V* (Luc gravait) dans `image-impression-25.1`. C'est en effet ici une sous-structure de *NP V NP.Destination* (Luc gravait les anneaux). Le codage de telles sous-structures est difficile à justifier quand il est basé sur l'introspection linguistique et nécessite une étude de corpus. Nous ne savons pas comment ce codage a été fait dans VerbNet et n'avons pas à notre disposition de corpus français permettant de répondre à la question. Nous avons donc décidé pour le moment de supprimer toutes les sous-structures de Verb∋Net. Par exemple, dans la classe `remove-10.1`, VerbNet encode non seulement NP V NP PP.Source PP.Destination (*Doug removed the smudges from the tabletop*) mais aussi NP V NP (*Doug removed the smudges*). Verb∋Net n'inclut que la première frame, il est implicite que la seconde existe : une application doit donc l'inférer automatiquement à partir de la première sans intervention manuelle. Ce principe ne s'applique pas aux verbes acceptant un seul complément locatif double « from here to there (un seul complément PP.Source PP.Destination) » sans accepter un seul complément source (PP.Source), tout en acceptant un seul complément destination (PP.Destination) : *Fred a transferé le vin de la cruche en pierre vers la cruche en terre cuite, *Fred a transferé le vin de la*

cruche en pierre, Fred a transferé le vin vers la cruche en terre cuite. Dans ce cas exceptionnel, les sous-structures sont codées explicitement.

Une difficulté subsiste. Pour reprendre l'exemple *Doug removed the smudges from the tabletop*, rien n'empêche de produire *Doug removed from the tabletop*. Nous avons cependant préféré laisser cette possibilité (erronée dans cet exemple) plutôt que d'avoir à recourir à des intuitions linguistiques pour l'ensemble des classes de Verb∋Net. Une application pourra noter que pour une frame de type NP V NP PP, la sous-structure qui est *a priori* la plus probable est NP V NP, mais NP V et NP V PP sont aussi possibles.

La deuxième différence concerne l'ordre des compléments. VerbNet encode parfois des frames qui ne diffèrent que par l'ordre des compléments, par exemple dans `bring-11.3` les frames NP V NP PP.Destination NP (*Nora brought to lunch the book*) et NP V NP PP.Destination (*Nora brought the book to the meeting*). En français, l'ordre des compléments dépend d'un certain nombre de facteurs syntaxiques et sémantiques [Thuilier, 2012], mais ne dépend pas à priori d'un facteur lexical : il ne dépend pas du verbe qui gouverne les compléments. C'est pour cette raison que Verb∋Net n'encode qu'une frame dans ces cas, ici seulement *NP V NP PP.Destination* (*Nora a apporté le livre au meeting*) avec l'objet direct avant le syntagme prépositionnel. C'est à l'utilisation de la ressource qu'il faut aussi envisager l'autre option (*NP V PP.Destination NP*, *Nora a apporté au meeting le livre*).

4.3.2. Travail au cas par cas

Dans certains cas, la deuxième étape pose des difficultés, et ce pour deux raisons principales. Premièrement, certaines différences sémantiques entre verbes communes à l'anglais et au français sont prises en compte par VerbNet mais ni par Les Verbes Français ni par le Lexique-Grammaire. Par exemple, dans les verbes de *Sendying and Carrying* (la super-classe 11), les verbes dans les classes `bring-11.3`, `carry-11.4` et `drive-11.5` décrivent un mouvement accompagné (non seulement le Theme mais aussi l'Agent changent de location comme dans *Pamela drove packages to NY*). Au contraire, les autres classes (`send-11.1` et `slide-11.2`) décrivent un mouvement non accompagné (seul le Thème se déplace comme dans *Pamela sent packages to NY*). Dans les ressources françaises, des classes existent pour des verbes avec un changement de location pour un Thème causé par un Agent, mais rien n'est codé pour le mouvement de l'Agent. Face à cette difficulté, deux solutions se présentent :

— soit réaliser une étude complète des verbes français de *Sending and Carrying* pour distinguer les mouvements accompagnés et non-accompagnés ;
— soit ignorer purement et simplement cette différence sémantique.

Dans ce cas, nous avons opté pour la seconde solution étant donné que cette information n'est pas directement utile pour l'annotation en rôles sémantiques [3]. C'est un choix discutable : si pour notre application la position de l'Agent avait eu un intérêt, comme ce pourrait être le cas en implication textuelle [Bobrow et al., 2007], une étude plus complète aurait été souhaitable. Nous préférons nous concentrer sur la finalisation d'une première version de Verb∋Net, quitte à revenir sur certains choix par la suite.

Le fait d'ignorer cette différence nous mène à adopter dans Verb∋Net une hiérarchie différente de VerbNet pour la super-classe 11 : il n'y a pas d'équivalent dans Verb∋Net de la classe `carry-11.4`, les verbes de cette classe étant placés dans les classes `send-11.1` et `slide-11.2`. Par ailleurs, il n'y a pas d'équivalent en français de la classe `bring-11.3` qui contient uniquement les deux verbes *bring* et *take* avec une direction spécifiée déictiquement [Levin, 1993, page 135] parce que les déictifs locatifs français *ici* et *là* n'ont pas le fonctionnement de *here* et *there* en anglais : en français, *Je suis là* peut signifier *Je suis ici*, ce qui n'est pas le cas en anglais.

La seconde source principale de difficultés provient de différences cruciales entre le français et l'anglais. Il existe des problèmes de traductions entre ces deux langues qui sont bien connus et documentés, comme la traduction des verbes de mouvement (par exemple *John swam across the river* → *Jean a traversé la rivière à la nage*). Sans traiter ces cas connus, nous discutons ici de situations plus subtiles, comme par exemple avec les verbes de changement de possession. Dans VerbNet, dix classes sont dédiées à ces verbes. Une telle hiérarchie est impossible en français. Sans tout détailler, insistons sur ces quelques points :

— L'absence d'alternances datif et bénéfactif en français implique que les classes VerbNet `give-13.1` et `contribute-13.2` doivent probablement être fusionnées en français.
— La différence sémantique entre la classe `give-13.1` (HAS-POSSESSION) et la classe `future_having-13.3` (FUTURE-POSESSION) est peut-être trop subtile et pourrait être ignorée.
— La préposition *with* dans la frame correspondant à *Agent V Recipient {with} Theme* utilisée en `fulfilling-13.4-1` et `fulfilling-13.4-2` doit être remplacée par *en* et/ou *de* suivant le verbe (e.g. *Luc livre Max en/*de lait*, *Luc équipe Max en/de téléviseurs*, *Luc dote Max *en/de téléviseurs*), ce qui nécessite une réorganisation en sous-classes pour distinguer ces verbes.

En conclusion, il s'avère que rentrer dans le détail des frames lors de cette deuxième étape nous a mené a faire évoluer la hiérarchie de Verb∋Net. Cependant, nous essayons de limiter les modifications quand il est impossible de faire autrement afin de profiter au maximum de VerbNet et pour pouvoir profiter du lien entre les deux ressources.

3. Il semble par ailleurs que pour certains verbes anglais, le mouvement ou non de l'Agent peut être ambigu : voir la différence entre la classe VerbNet `carry-11.4` et la classe VerbNet `carry-11.4-1`, qui elle n'a aucun membre.

4.4. Outil d'édition de Verb∋Net

Nous avons développé un outil (Figure 4.1) permettant d'éditer collaborativement Verb∋Net [4]. Cet outil est en fait un site web présentant Verb∋Net : ses classes, ses frames, ses verbes, etc. Tous les éléments sont modifiables, ce qui permet aux lexicographes de se concentrer sur les problèmes lexicographiques sans avoir à se préoccuper de la représentation des données.

Le site web est réalisé avec Django, un framework web Python. Verb∋Net est stocké dans une base de donnée PostgreSQL. Le nombre de requêtes SQL est minimisé de deux manières :

— La hiérarchie des classes est stockée à l'aide de django-mptt [5] : de nombreuses opérations telles que l'affichage d'une partie de la hiérarchie ne demandent qu'une requête SQL.

— Le nombre de requêtes SQL réalisées par Django est aussi minimisé à l'aide de prefetch_related [6] afin de récupérer toutes les données nécessaires en une seule requête et ainsi éviter le coût de plusieurs milliers de petites requêtes.

Le but de ces optimisations est que les opérations durent moins d'une seconde, de manière à ne pas interrompre la réflexion lexicographique [Nielsen, 1994]. Pour la même raison, les erreurs sont évitées au maximum grâce aux tests unitaires : les erreurs lors de l'utilisation de l'application érodent la confiance des lexicographes en l'outil. Ils ne devraient pas avoir à se préoccuper de la possibilité de perdre leur travail. Enfin, toutes les modifications du schéma de la base de données mais aussi des données sont versionnées automatiquement, afin de rendre impossible toute perte de données.

Avant de commencer l'édition, nous avons commencé par charger VerbNet et les correspondances avec les lexiques français réalisés lors de la première étape. L'édition elle-même se fait par classe de Levin. Par exemple, nous avons traité ensemble les classes `throw-17.1` et `pelt-17.2` parce qu'elles font toutes les deux partie de la « super-classe » 17 : Throwing. Plusieurs classes faisant partie d'une même super-classe sont souvent liées et il faut les comprendre dans leur ensemble avant de commencer l'édition.

Pour chaque classe, nous pouvons éditer ses correspondances, ses frames, ajouter ou supprimer des sous-classes ou des frames. Par exemple, pour la suppression, toutes les frames impliquant un conatif, un datif ou une alternance bénéfactive peuvent être systématiquement supprimées : ces alternances n'existent pas en français.

Avec l'aide de cet outil, la deuxième étape peut s'avérer très simple. Par exemple, les quatre sous-classes de `image-creation-25` ont des classes directement équivalentes en français, donc les seules choses à faire sont de traduire les exemples en français avec les bonnes prépositions. Par exemple, dans la classe `illustrate-25.3`, il a fallu remplacer *with* par la combinaison de *de* et *avec*.

4. Cet outil est libre et disponible sur GitHub : `https://github.com/aymara/verbenet-editor`.
5. `https://github.com/django-mptt/django-mptt/`
6. `https://docs.djangoproject.com/en/stable/ref/models/querysets/#prefetch-related`

Différents mécanismes ont été rajoutés au site web au fil de l'annotation afin de faciliter l'annotation. Nous en présentons quelques-uns ici.

4.4.1. Édition des correspondances

Il est fréquent que les correspondances avec les ressources linguistiques établies lors de la première étape (section 4.1.1) soient amenées à évoluer, et ce pour trois raisons principales : la compréhension des classes anglaises peut évoluer, la réorganisation peut demander des correspondances différentes, et il peut être souhaitable d'affiner la correspondance en spécifiant certains attributs de la classe française.

Dans le cas des tables du Lexique-Grammaire, les correspondances sont affinées à l'aide des colonnes présentes dans chacune des tables. Par exemple, dans la table 38LD (N_0 V N_1 Prép N_2, Location-Destination) [7], six colonnes encodent les prépositions possibles pour introduire N_2. Ainsi, on sait que le verbe *appuyer* peut s'utiliser avec *sur* et *contre* en raison des + présents aux colonnes *Loc N2 = : sur N2 destination* et *Loc N2 = : contre N2 destination*.

Il est donc possible dans l'interface de VerbNet d'imposer la valeur de certaines des colonnes de la classe, et ce pour filtrer les verbes ne correspondant pas à la classe Verb∋Net en cours d'édition. Donnons deux exemples.

— R3i.1[+T1308 et +P3000] ou R3i.1[+T1306 et +P3000] sélectionne tous les verbes de R3i.1 acceptant une construction pronominale (P3000) et une construction transitive directe à sujet humain, à objet non animé et à circonstant de modalité (T1306) ou instrumental, moyen (T1308).
— 4[+N1 se V de ce Qu P] sélectionne tous les verbes de la classe 4 ayant la valeur '+' à la colonne 'N1 se V de ce Qu P', ce qui signifie qu'ils acceptent cette forme.

Toutes ces correspondances sont validées par un micro-compilateur :

— lors de l'analyse lexicale (par exemple, L3 [L2 n'est pas valable),
— lors de l'analyse syntaxique (par exemple, L3b ou L3a et L3c est invalide)
— et lors de l'interprétation : l'existence des classes et des colonnes est vérifiée.

Les correspondances, une fois validées, sont utilisées pour mettre à jour la liste des verbes français de la classe.

7. `https://verbenet.inria.fr/verbes-html/V_38LD.lgt.html`

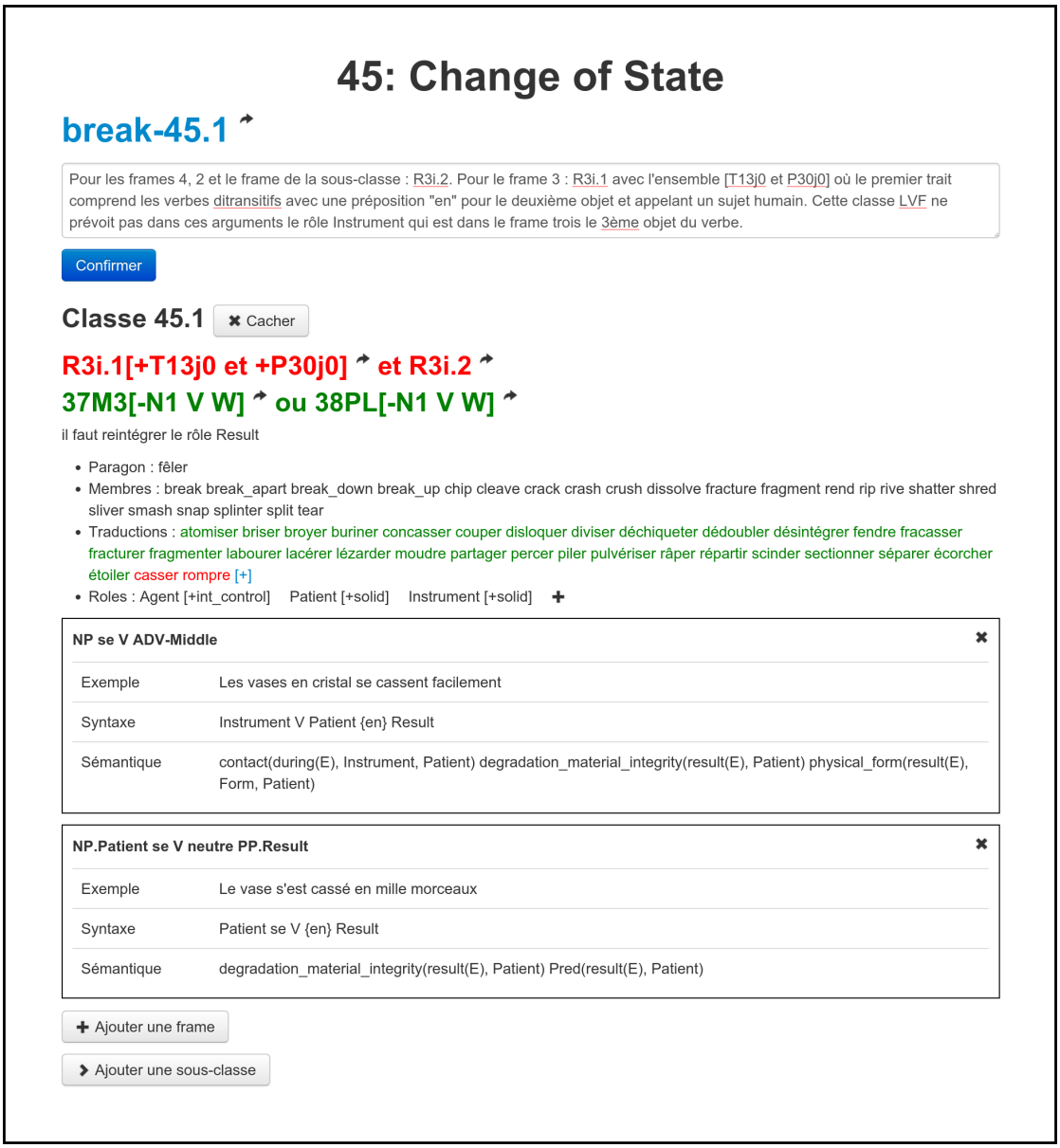


FIGURE 4.1. – Interface web pour analyser et modifier Verb∋Net. Chaque frame peut être modifiée et les classes peuvent être réorganisées. Les traductions en violet apaprtiennent à l'intersection de C_{lvf}, C_{lg} et L_{trad} (section 4.1.1), les traductions en rouge (respectivement en vert) uniquement à C_{lvf} et L_{trad} (respectivement uniquement à C_{lg} et L_{trad}). La description de `break-45.1` est en cours de modification : en cliquant sur 'Confirmer', la modification sera prise en compte rapidement sans avoir à recharger la page pour éviter d'interrompre le travail lexicographique.

4.4.2. Traduction des verbes en temps réel

En effet, la liste des verbes français de chacune des classes n'a pas été seulement calculée à la fin de l'étape 1, mais est recalculée à chaque modification des correspondances ou des verbes anglais. Ainsi, les lexicographes voient immédiatement l'effet de leurs modifications et peuvent vérifier que les verbes attendus sont bien présents dans la liste des verbes français.

À chaque modification, des verbes peuvent disparaître, d'autres peuvent être rajoutés. La validation des verbes est le moyen de stabiliser les listes de verbes.

4.4.3. Validation des verbes français

C'est la troisième étape, en cours parallèlement avec la deuxième, qui concerne la validation des verbes. En cliquant sur un verbe français proposé à partir des correspondances, ce verbe est soit validé, soit invalidé, soit remis à l'état de simple proposition. Contrairement aux verbes proposés par correspondances, les verbes validés ou invalidés ne sont jamais supprimés, même lorsque que les correspondances changent. Par exemple, dans la classe `judgment-33`, le verbe *blâmer* a été validé manuellement : il ne peut plus être supprimé de la liste des verbes valides de la classe.

4.4.4. Vérifications automatiques de la cohérence

Petit à petit, différents outils sont ajoutés pour aider les lexicographes à vérifier la cohérence de leur travail. Deux existent aujourd'hui.

Le premier identifie les verbes dupliqués au sein de la même super-classe de Levin. En effet, bien qu'il soit naturel que le même lemme verbal soit présent à plusieurs endroits de Verb∋Net pour des raisons de polysémie, les classes de Levin regroupent des verbes sémantiquement proches : un verbe donné ne doit être que dans une seule classe. Par ailleurs, quand un verbe peut exister à différents endroits, c'est la classe la plus générale (souvent la première dans la hiérarchie) qu'il faut choisir. Les utilisateurs connectés voient donc un encadré au début de chaque classe de Levin leur indiquant quels verbes sont présents à plusieurs endroits : il devient alors facile de valider/dévalider les verbes concernés.

Le second outil est l'index LADL (`https://verbenet.inria.fr/ladlindex/`). L'objectif est de s'assurer qu'aucune classe des tables du LADL (Lexique-Grammaire) majeure n'a été oubliée, et donc que le lexique couvre bien l'ensemble des verbes du français. C'est en effet le cas : les seules classes sans correspondance avec une classe Verb∋Net sont des classes résiduelles. Cela confirme l'idée que Verb∋Net peut être vue comme une réorganisation des tables du Lexique-Grammaire (section 1.3.6).

4.4.5. Gestion des classes supprimées en anglais

Lors d'une réorganisation, il est fréquent de ne plus avoir besoin d'une classe en anglais. Nous traitons pour le moment deux cas de figures.

Premièrement, lorsqu'une sous-classe est supprimée, tous ses verbes anglais sont automatiquement migrés vers sa classe mère, et les traductions des verbes français sont mises à jour en fonction des nouveaux verbes et des correspondances présentes dans la classe mère. C'est essentiel, étant donné que les verbes les plus importants sont les plus polysémiques et ceux acceptant le plus de constructions : ils sont donc souvent dans les sous-classes, ce qui n'est pas nécessairement intuitif.

Deuxièmement, lorsqu'une classe est simplement supprimée sans classe mère, ses verbes sont théoriquement "perdus". Il est possible de transférer ces verbes vers une autre classe plus appropriée, tout en gardant l'origine des verbes anglais, au cas où cette opération devait être annulée. Ici aussi, les verbes français sont ensuite mis à jour.

Troisièmement, un cas symétrique existe lorsqu'une classe anglaise correspond à plusieurs classes françaises. On ne peut pas alors simplement transférer les verbes anglais vers une classe française : il faut répartir les verbes entre les différentes classes françaises. Enfin, certaines classes de VerbNet disposent de découpages qui n'ont pas de sens en français. Il faut alors réorganiser totalement ces classes. Les sous-classes françaises ne sont pas directement liées aux sous-classes anglaises. Pour ne pas perdre l'information des verbes, il est prévu de pouvoir indiquer que ces verbes appartiennent à la super-classe, avant de les répartir soit automatiquement grâce aux associations avec les ressources françaises, soit manuellement.

Conclusion

Nous avons présenté une méthode pour adapter la ressource syntaxico-sémantique VerbNet vers une nouvelle langue. Cette méthode combine l'automatisation du transfert de frames, la traduction automatique du lexique et une expertise linguistique. Nous avons appliqué cette méthode au français en reconnaissant les différences entre les langues : la structure de Verb∋Net n'est pas exactement celle de VerbNet. Nous avons atteint un point où la méthodologie est validée : le travail systématique sur chaque classe est en cours et devrait être terminé en 2015.

Cette ressource nous permet d'ores-et-déjà de réaliser de l'annotation en rôles sémantiques pour le français : nous montrons que c'est possible au Chapitre 6.

Troisième partie

Méthodes pour l'annotation en rôles sémantiques

Trois raisons expliquent l'importance de l'annotation en rôles sémantiques :

— Elle capture des phénomènes sémantiques de manière utile sous une forme proche de la logique du premier ordre (sans quantifieur ni portée) qu'il est possible d'utiliser dans diverses applications par la suite [Osman et al., 2012, Xie et al., 2013].
— De nombreuses façons d'exprimer un énoncé seront représentées de la même manière : les traitements ultérieurs n'ont plus besoin de considérer cette complexité.
— Enfin, l'annotation en rôles sémantiques est capable d'opérer à grande échelle : sur l'ensemble du vocabulaire et pour tout type de texte. En effet, la définition de l'annotation en rôles sémantiques ne la limite pas à un domaine ou un genre particulier et la méthodologie VerbNet que nous suivons rend possible la classification de la grande majorité des verbes du « domaine général » [8].

Un système d'annotation en rôles sémantiques efficace en général et adaptable avec peu ou pas d'effort vers un nouveau domaine serait donc une avancée majeure. Les systèmes supervisés ne correspondent pas à cet objectif : ils sont par nature dépendants du corpus annoté utilisé. En effet, pour une migration vers un autre domaine, il faut a priori un nouvel effort d'annotation colossal.

La solution que nous explorons dans ce travail est donc de ne pas utiliser de corpus annoté mais une ressource lexicale, indépendante du domaine considéré : VerbNet. VerbNet répertorie des informations syntaxico-sémantiques sur les verbes de l'anglais. Cette ressource correspond à la volonté de réaliser une analyse à large échelle : elle couvre une large partie des occurrences des verbes anglais (tout en ayant le potentiel de couvrir 99 % des verbes [Palmer et al., 2013, partie 1, p .53]) et est adaptée à l'analyse en rôles sémantiques. Pour les cas où un besoin de sémantique plus important se fait sentir, WordNet est une deuxième ressource utile [Shi and Mihalcea, 2005].

L'objectif est d'identifier la capacité de VerbNet à aider l'annotation en rôles sémantiques :
— dans un cadre général sur le corpus FrameNet (Chapitre 5),
— dans des domaines spécifiques en anglais mais aussi en français grâce à Verb∋Net (Chapitre 6),

8. Nous nous concentrons ici sur les verbes, les autres parties du discours devant faire l'objet d'une autre étude.

5. Annotation en rôles sémantiques fondée sur la connaissance

La tâche d'annotation en rôles sémantiques a reçu beaucoup d'attention ces dernières années, à la fois pour les approches supervisées et semi-supervisées. Les approches fondées sur la connaissance, elles, ne se basent pas sur des corpus annotés mais sur des ressources lexicales existantes. Ce type d'approche a été négligé malgré leur complémentarité par rapport aux autres approches.

En nous inspirant de Swier and Stevenson [2004, 2005], nous présentons dans ce chapitre un système d'annotation en rôles sémantiques fondé sur la connaissance qui se veut simple à mettre en place et facile à reproduire. La prise en compte de divers phénomènes linguistiques doit permettre d'améliorer les performances. Par exemple, la prise en compte de la voix passive a réduit le taux d'erreur de 15,7 %, ce qui montre la marge de progrès existante. Malgré des performances moindres par rapport aux approches supervisées quand des données d'entraînement existent, l'approche facilite l'analyse des erreurs, n'a pas besoin d'un corpus annoté manuellement et est *a priori* indépendante du domaine considéré, étant donné qu'elle utilise le lexique VerbNet.

Ce chapitre se concentre sur notre système d'annotation en rôles sémantiques dans un cadre général en l'utilisant sur le corpus FrameNet anglais (qui a été présenté à la section 1.3.4). Le Chapitre 6 montrera la versatilité de ce système dans des contextes différents :

— en domaine spécifique avec les domaines du football, du réchauffement climatique et de l'informatique,
— et dans deux langues différentes : anglais et français.

Le travail présenté dans ce chapitre a été réalisé en collaboration avec Guilhem Pujol durant son stage d'ingénieur. Ce chapitre est une réécriture complète de travaux pour la plupart déjà publiés [Pradet et al., 2013b].

5.1. Tâche

L'objectif de ce chapitre est d'annoter en rôles sémantiques le corpus FrameNet, largement utilisé pour l'évaluation [Baker et al., 2007, Das et al., 2010]. Nous souhaitons cependant l'évaluer en utilisant la ressource VerbNet. Pour ce faire, nous utilisons le projet SemLink [Bonial et al.,

La phrase à annoter est :

> However, in 2002 Russia declared it will eliminate its tactical nuclear weapons by the end of 2004.

L'objectif est d'aboutir à la représentation suivante :

- *declare* déclenche la frame Statement dont les rôles sont remplis ainsi :
  - Speaker : Russia
  - Message : it will eliminate its tactical nuclear weapons by the end of 2004
  - Time : in 2002
- *eliminate* déclenche la frame Removing dont les rôles sont remplis ainsi :
  - Agent : it
  - Theme : its tactical nuclear weapons
  - Source : *non instancié*

FIGURE 5.1. – Exemple d'annotation en rôles sémantiques. Le vocabulaire FrameNet est ici utilisé (voir la section 1.3.5).

2012] qui propose un mapping de rôles sémantiques entre les frames FrameNet et les classes VerbNet ce qui permet d'évaluer l'annotation en rôles sémantiques VerbNet sur FrameNet.

La Figure 5.1 est un exemple d'annotation en rôles sémantiques avec des frames et des rôles FrameNet. Nous utiliserons notamment cette phrase d'exemple par la suite pour illustrer notre propos. Pour une phrase donnée, les prédicats verbaux (*trigger* ou déclencheur dans la terminologie FrameNet) et la frame qu'ils évoquent sont identifiés. Dans chaque phrase, des syntagmes vont remplir un des rôles prévus par la frame dans FrameNet. Ces syntagmes sont nommés « remplisseurs de rôle » (*role filler* en anglais). L'analyse est locale : seule la phrase courante est considérée.

Toute interprétation supplémentaire non présente dans la Figure 5.1 est en dehors du cadre de l'annotation en rôles sémantiques, ce qui est la raison pour laquelle la tâche est aussi connue sous le nom d'analyse sémantique de surface (*shallow semantic parsing*). Il est néanmoins important de garder à l'esprit les applications possibles de tels travaux. Par exemple, un système de question-réponse pourrait utiliser la représentation de ces deux *frames* pour répondre à la question *Does Russia possess tactical nuclear weapons ?* L'annotation de la Figure 5.1 est une information utile, mais elle ne serait pas suffisante pour répondre à la question : il faut aussi comprendre la question, annoter les coréférences (*it* fait référence à *Russia*), comprendre la sémantique de *Removing*, établir la crédibilité du *Speaker*, s'intéresser à d'autres phrases potentiellement contradictoires, etc.

5.2. Système

Le système que nous présentons est une implémentation améliorée et libre du système décrit par Swier and Stevenson [2004, 2005]. Les ressources utilisées ont beaucoup progressé (VerbNet 1.5 contre VerbNet 3.2 notamment). Les corpus sur lesquels s'évaluer ont aussi évolué. En particulier, les systèmes basés sur FrameNet n'utilisent plus le corpus de phrases d'exemples choisis pour leur diversité syntaxique mais s'évaluent sur des annotations de toutes les frames présentes dans des textes issus de différentes sources. Swier and Stevenson [2005] utilisent aussi un mapping non disponible alors que le projet SemLink a fourni un mapping FrameNet-VerbNet « officiel ». Dix ans après, il est donc important d'évaluer à nouveau cette approche pour savoir où elle se situe par rapport à l'état de l'art. Nous proposons par ailleurs certaines améliorations.

Pour commencer, chaque phrase à annoter doit d'abord être étiquetée en morphosyntaxe et analysée syntaxiquement. Nous laissons ici ces analyses à des systèmes externes présentés à la section 5.5.2. Ces analyses sont l'entrée de notre système, la sortie étant l'annotation en rôles sémantiques montrée dans la Figure 5.1.

Ensuite, nous utilisons les informations de VerbNet (présenté à la section 1.3.3) sur l'interface entre la syntaxe et la sémantique. Dans VerbNet, chaque classe regroupe un certain nombre de verbes acceptant tous les mêmes constructions syntaxiques. Les syntagmes participant à ces constructions sont associés à des rôles sémantiques à interpréter dans le contexte de la classe. Ces *frames* sont notées de cette façon : NP.Agent V NP.Theme. Ici, dans cette construction transitive (NP V NP, e.g. *Sally pushed the chair*), le premier syntagme nominal (*Sally*) est Agent alors que le second (*the chair*) est Theme. Bien que des règles précises régissent leur attributions, l'interprétation complète des rôles (ici Theme et Agent) dépend de la classe VerbNet considérée. Par exemple, la classe `resign-10.11` contient la frame NP.Agent V PP.Source (*I resigned from the military*). Dans cette classe, l'Agent est la personne qui démissionne et la Source est le poste qui a été quitté.

Pour une phrase donnée, le système commence par identifier les verbes de cette phrase. Pour chacun de ces verbes, un ensemble de classes VerbNet est identifié. Ainsi, pour notre phrase d'exemple ci-dessus, les classes possibles pour le verbe *declare* sont `declare-29-4-1-1-1`, `say-37.7-1` et `reflexive_appearance-48.1.2`. Le choix correct est `say-37.7-1`, mais il ne nous est pas possible de le déterminer avant de considérer les frames listées par VerbNet dans ces différentes classes.

Par exemple, `reflexive_appearance-48.1.2` contient la frame NP.Agent V NP.Theme. Ainsi, si un des verbes de cette classe, tel que *present* :

— est utilisé dans un sens compatible avec la classe `reflexive_appearance-48.1.2`,
— et est utilisé avec un sujet syntagme nominal et un objet syntagme nominal,

alors le sujet du verbe est l'Agent et l'objet est le Theme, ce que des applications pourront interpréter dans le contexte de la classe VerbNet `reflexive_appearance-48.1.1`.

La phrase d'exemple de la Figure 5.1 ne correspond pas à NP V NP mais à NP V that S (*that S* étant la notation VerbNet pour les complétives introduites par *that*). La seule occurrence de cette frame VerbNet est dans `say-37.7-1` : NP.Agent V that S.Topic (*He ordered that he go*). Les classes `declare-29-4-1-1-1` et `reflexive_appearance-48.1.2` ne listant pas cette frame, il est possible d'établir que :

— la classe VerbNet qui convient est `say-37.7-1`,
— *Russia* est Agent,
— *it will eliminate its tactical nuclear weapons by the end of 2004* est Topic.

Enfin, le mapping de VerbNet vers FrameNet (section 5.5.1) nous informe que :

— la classe VerbNet `say-37.7` correspond à la frame Statement,
— dans cette classe, le rôle VerbNet Agent correspond au rôle FrameNet Speaker,
— dans cette classe, le rôle VerbNet Topic correspond au rôle FrameNet Topic.

Nous aboutissons ainsi à l'annotation en rôles sémantiques voulue pour notre phrase d'exemple.

On ne peut pas toujours identifier la classe VerbNet correcte. C'est le cas par exemple de `reflexive_appearance-48.1.2` et `say-37.7-1` qui contiennent toutes les deux le cadre NP V NP. Ainsi, si notre phrase d'exemple avait été *Russia declared its intentions*, la classe serait restée ambigüe. Sans corpus annoté, ces ambiguïtés ne peuvent pas être résolues. Cependant, une fois qu'une première série de correspondances a été effectuée, il est possible d'utiliser les connaissances du domaine étudié pour annoter de nouveaux syntagmes nominaux (section 5.2.4).

Alors que nous avons jusqu'ici décrit le principe général du système afin d'en donner l'intuition, les sections suivantes expliquent le fonctionnement précis du système qui est découpé en quatre étapes.

5.2.1. Identification du prédicat

Chaque phrase peut contenir un ou plusieurs prédicats : nous nous contentons ici de retenir tous les verbes présents dans VerbNet. Les autres prédicats potentiels, c'est-à-dire ceux qui ne sont pas dans VerbNet et ne sont pas des verbes, sont ignorés. Les parties du discours acceptées sont toutes celles concernant des verbes simples : MD, VB, VBD, VBG, VBN, VBP, VBZ[1]. Les autres formes où le verbe est composé de plusieurs mots (par exemple *He has (VHZ) suffered (VVN) from loneliness*) ne sont pas actuellement prises en compte.

1. FrameNet annonce utiliser simplement les parties du discours des ressources utilisées [Ruppenhofer et al., 2006, section 3.1], mais cela semble correspondre au tagset du TreeTagger.

5.2.2. Identification des arguments

Nous évaluons notre système de deux manières :
— avec les arguments issus de la vérité-terrain.
— et avec les arguments identifiés automatiquement,
En effet, l'objectif est d'évaluer l'apport de VerbNet à la tâche de l'annotation en rôles sémantiques. L'identification des arguments, bien qu'une partie intégrante de tout système complet d'annotation en rôles sémantiques [Das et al., 2010], ne fait pas partie de nos contributions.

Arguments de la vérité-terrain Pour convertir les arguments de la vérité-terrain vers une frame VerbNet, nous utilisons simplement la position de ces arguments dans la phrase. Par exemple, si un syntagme nominal apparaît avant le verbe et un syntagme prépositionnel apparaît après le verbe, la frame VerbNet devient NP V PP. Bien que cela corresponde en général au véritable cadre de sous-catégorisation de la phrase, ce n'est pas toujours le cas. Montrons-le avec deux exemples.

— Pour *Les déchets éliminés polluent* et le verbe *éliminé*, la frame est NP V, alors qu'il aurait fallu V NP.
— *Il dégage le ballon* et *Il le dégage* produisent respectivement les cadres NP V NP et NP NP V. Cependant, *le ballon* et *le* doivent tous les deux être considérés comme un syntagme nominal rattaché au verbe avant de comparer le cadre de sous-catégorisation de la phrase à VerbNet.

Ainsi, même lorsque les arguments de la vérité-terrain sont utilisés, l'absence de prise en compte de la syntaxe est pénalisant, ce qu'il faudra traiter lors de futurs travaux.

Identification automatique Nous utilisons simplement les fils syntaxiques du nœud verbal, en utilisant aussi la position dans la phrase pour la correspondance VerbNet. Il s'est avéré que cette approche simpliste dégrade considérablement les résultats par rapport aux arguments automatiques. Nous avons aussi implémenté la technique à base de règles de [Lang and Lapata, 2011], mais les résultats étaient inférieurs.

5.2.3. Correspondance exacte des frames

Cette étape associe zéro, un ou plusieurs rôles sémantiques à chaque syntagme candidat identifié lors de l'étape précédente. Elle correspond au *frame matching* de Swier and Stevenson [2005].

Nous incluons ici deux étapes traditionnellement séparées dans les systèmes d'annotation en rôles sémantiques : l'identification des frames FrameNet puis l'assignation de rôles aux argu-

ments précédemment identifiés [Gildea and Jurafsky, 2002, Das et al., 2014]. Nous commençons par identifier les frames VerbNet possibles pour restreindre au maximum le nombre de classes VerbNet applicables, le but étant de n'en avoir qu'une. En effet, même si toutes les classes VerbNet réutilisent les mêmes rôles, le sens précis d'un rôle est déterminé par la classe. Il faut donc l'identifier.

Regardons comment se déroule cette étape. Premièrement, les syntagmes candidats sont représentés au format VerbNet. Par exemple, si trois syntagmes nominaux ont été identifiés comme arguments, dont un avant le verbe, la représentation VerbNet de la phrase devient NP V NP NP. Si le troisième syntagme est un syntagme prépositionnel introduit par *in*, la représentation devient NP V NP in NP, et ainsi de suite.

Ensuite, pour comparer la représentation VerbNet de la phrase aux frames VerbNet, nous identifions toutes les classes VerbNet incluant le prédicat. Par exemple, le prédicat *classify* est présent dans deux classes VerbNet : `characterize-29.2` et `classify-29.10`. Les frames VerbNet possibles sont :

- characterize-29.2
  - NP.Agent V NP.Theme (as) S_ING.Attribute
  - NP.Agent V NP.Theme to be ADJ.Attribute
  - NP.Agent V NP.Theme as PP.Attribute
- classify-29.10
  - NP.Agent V NP.Theme
  - NP.Agent V NP.Theme as PP.Goal
  - NP.Agent V NP.Theme in PP.Location

Considérons la phrase *The curator classified the artifacts*. La représentation VerbNet de cette phrase est NP V NP, ce qui correspond au NP.Agent V NP.Theme de la classe `classify-29.10`. On en déduit que *The curator* est Agent, et que *the artifacts* est Theme. Cette phrase est donc correctement annotée en rôles sémantiques VerbNet.

Prenons un autre example, cette fois tiré de FrameNet. La phrase *The company also classifies short and wide radius ruts according to their severity* est transformée en NP V NP according PP. Dans ce cas, seuls les deux premiers syntagmes peuvent être mis en correspondance avec les frames VerbNet listées ci-dessus. Il n'y a pas de correspondance possible pour le troisième syntagme : VerbNet n'encode pas *according* comme une préposition possible alors que *in* et *as* sont acceptées. De tels arguments non prévus pas VerbNet ne sont pas annotés lors de cette étape. C'est ici un problème de couverture. Les auteurs de VerbNet travaillent actuellement sur la couverture de la ressource en ajoutant des informations syntaxiques et lexicales issues de très larges corpus [Bonial et al., 2013]. Le résultat de ces travaux n'est cependant pas disponible dans la version 3.2 de VerbNet que nous utilisons. Ce problème de couverture empêche d'assigner la classe VerbNet qui convient. Par conséquent, même si on sait que quelle que soit la classe, le sujet serait Agent et le premier objet Theme, cette information est difficilement interprétable dans une application.

Dans les cas où différentes frames VerbNet sont possibles, le calcul de scores de correspondances présentés par Swier and Stevenson [2004] peut aider à désambiguïser, notamment en éliminant les *frames* comportant trop d'arguments. Cela dit, de nombreux arguments restent ambigus. En effet, pour un peu plus de 25 % des arguments de notre corpus FrameNet, la correspondance exacte ne permet pas d'identifier le rôle correct, mais seulement de le limiter à quelques rôles possibles. Cette délimitation est très précise : quand VerbNet aboutit à plusieurs rôles possibles, la probabilité pour que le rôle correct soit présent dans la liste est supérieure à 90 % pour des arguments de la vérité-terrain. Nous posons alors la question suivante : comment tirer profit de cette courte liste de rôles possible ? C'est l'objet des sections 5.2.4 et 5.4.

5.2.4. Correspondance probabiliste des frames

Maintenant que les correspondances exactes ont étés identifiées, il reste d'une part les correspondances impossibles et les correspondances ambigües pour lesquelles plusieurs frames sont possibles. Alors que les correspondances impossibles sont à corriger au niveau de la ressource VerbNet ou au niveau de l'analyse syntaxique, nous pouvons utiliser les frames déjà mises en correspondance pour désambiguïser les correspondances ambigües. La méthode, une forme d'induction, reste non supervisée : bien que nous entraînions une forme simple d'algorithme supervisé, nous le faisons sur des données qui sont initialement non annotées. En effet, elles sont simplement obtenues automatiquement sur le corpus existant à l'aide de la correspondance exacte. Nous faisons ici l'hypothèse qu'un corpus entier est à annoter, mais dans le cas où l'annotation se ferait phrase par phrase, il serait possible d'abandonner cette étape ou d'alimenter les classifieurs que nous utilisons au fur et à mesure des annotations. Cette étape correspond aux *probability models* de Swier and Stevenson [2004].

L'apprentissage se fait sous la forme d'un simple classifieur statistique. Le classifieur (issu de Swier and Stevenson [2004]) assigne une probabilité aux différents rôles possibles en s'aidant des rôles déjà identifiés.

$$\text{rôle} = \arg\max_{\text{rôle}} p(\text{rôle}|\text{prédicat}, \text{fonction})$$

La fonction est la fonction grammaticale identifiée d'après l'analyse syntaxique : si un syntagme apparaît avant le verbe, il est sujet, s'il est après le verbe, et il est objet, et ainsi de suite. Les syntagmes prépositionnels sont traités à part : la préposition qui introduit le syntagme est considérée à part.

Ce classifieur utilise l'information du prédicat et la fonction grammaticale détectée. Par exemple, dans notre corpus (section 5.5.2), l'objet direct du verbe *négliger* est le plus souvent *Theme*. La précision pour ce modèle est forte, mais il n'assigne des rôles que pour 40 % des arguments :

dans les autres cas, nous ne disposons pas d'informations pour cette paire (prédicat, fonction grammaticale).

Une procédure de *bootstrap* reposant sur les mêmes principes a aussi été implémentée en suivant Swier and Stevenson [2004].

5.3. Gestion de la voix passive

Une analyse d'erreur a révélé que la voix passive était une source d'erreurs importante dans l'analyse de notre corpus FrameNet. En effet, VerbNet n'encode pas la voix passive qui est un phénomène syntaxique, et non un phénomène lexical, car (pratiquement) tous les verbes anglais permettent le passif (ce qui n'est pas le cas du français). C'est donc au moment de l'analyse syntaxique que les sujets et objets syntaxiques, profonds ou non, doivent être identifiés correctement. Pour annoter la phrase *the artifacts were classified by the curator* en rôles sémantiques, il est donc important de d'abord transformer la phrase en *the curator classified the artifacts* avant d'effectuer la correspondance avec les frames VerbNet.

Les sujets et objets profonds ne sont pas encodés dans les corpus annotés en syntaxe que nous utilisons (le Wall Street Journal pour l'anglais, cf. section 5.5.2). Une étape intermédiaire est donc nécessaire entre l'analyse syntaxique et l'annotation en rôles sémantiques [Bonfante et al., 2011, Ribeyre, 2013]. Cette étape intermédiaire pourra identifier les sujets et objets profonds de tous les verbes considérés, évitant ainsi toute une classe d'erreurs lors de l'annotation en rôles sémantiques.

Afin de valider cette hypothèse, nous nous sommes concentrés sur la voix passive qui était le phénomène de syntaxe profonde le plus présent dans notre corpus. Pour annoter les verbes repérés comme étant utilisés avec la voix passive (et uniquement pour ceux là), nous avons remplacé les frames VerbNet par leur deux équivalents à la voix passive. Les verbes utilisés au passif en anglais sont au participe passé et gouvernés par une forme du verbe *to be*. Étant donné une frame VerbNet telle que *NP.Agent V NP.Theme (eg. the curator classified the artifacts)*, nous la remplaçons par deux nouvelles frames :

— NP.Theme V (the artifacts were classified)
— NP.Theme V by NP.Agent (the artifacts were classified by the curator)

Ce sont ces frames VerbNet transformées qui sont utilisées lorsque qu'une voix passive est detectée, ce qui améliore les résultats (Table 5.2). Cette expérience valide la gestion de tels phénomènes syntaxiques. Pour aller plus loin, il faudra non pas modifier VerbNet mais bien annoter la phrase en syntaxe profonde avant de réaliser les correspondances exactes puis probabilistes. Premièrement, cela limite l'ambiguïté provoquée par la transformation de VerbNet : par exemple, *by* n'est pas limité à l'introduction du sujet dans une construction passive. Deuxièmement, il est

plus naturel d'identifier les sujets profonds avant de les comparer à des représentations conçues pour utiliser la voix active. Troisièmement, cela permet de généraliser à d'autres phénomènes de syntaxe profonde, par exemple en utilisant les systèmes existants [Bonfante et al., 2011, Ribeyre, 2013].

5.4. Restrictions de sélection VerbNet

Nous explorons dans cette section l'apport d'une information de nature plus sémantique pour montrer la complémentarité de la syntaxe et de la sémantique pour notre tâche. Différentes informations sémantiques ont déjà montré leur utilité en annotation en rôles sémantiques supervisée. Citons :

— la présence de relations WordNet entre mots des prédicats pour l'identification des prédicats [Das et al., 2010],
— et l'apport de la représentation de mots (*word embeddings*) pour une meilleure généralisation [Léchelle and Langlais, 2014].

Pour rester dans le cadre de l'annotation en rôles sémantiques fondée sur la connaissance, nous allons dans cette section explorer l'utilisation des restrictions de sélection présentes dans VerbNet.

5.4.1. Restrictions avec WordNet

Nous commençons par essayer d'utiliser la connaissance issue de WordNet. Nous avons d'abord fait correspondre les restrictions de sélection VerbNet à des synsets WordNet qui sont souvent en haut de la hiérarchie à la façon de [Shi and Mihalcea, 2005] (Figure 5.1).

Ensuite, pour chaque syntagme dont la tête est un nom commun, adjectif ou adverbe, nous regardons dans WordNet si le premier synset correspondant à ce syntagmes est un hyponyme du synset correspondant dans la Table 5.1.

Cette première tentative est négative (section 5.6.1). Une potentielle explication est que ces restrictions sont relativement grossières et que les hypéronymes de WordNet n'ont pas été conçus pour représenter des propriétés (animé, organisation).

Restriction de sélection VerbNet	Synset WordNet
abstract	abstraction.n.06
animal	animal.n.01
animate	animate_thing.n.01
body_part	body_part.n.01
comestible	comestible.n.01
communication	communication.n.02
concrete	physical_entity.n.01
currency	currency.n.01
garment	clothing.n.01
human	human.n.01
int_control	animate_thing.n.01
location	location.n.01
machine	device.n.01
organization	organization.n.01
region	region.n.01
scalar	quantity.n.01
solid	matter.n.03
sound	sound.n.01
substance	substance.n.01
time	time_period.n.01
vehicle	transport.n.01

TABLE 5.1. – Correspondances entre restrictions de sélection et synsets WordNet. Toutes les restrictions de sélection ne correspondent pas à des synsets, c'est le cas par exemple de *elongated* ou *nonrigid* qui sont des propriétés qu'on ne peut pas vérifier directement avec WordNet. La plupart des restrictions sont cependant couvertes.

5.4.2. Restrictions en utilisant les syntagmes annotés sans ambiguïté

Une autre expérience que nous avons menée est d'apprendre à partir des correspondances exactes à quels mots sont rattachées le plus souvent les différentes restrictions de sélection dans chaque fichier FrameNet. Au moment de la correspondance exacte, nous enregistrons donc les têtes de syntagmes pour toutes les restrictions de sélection simples et les restrictions rattachées par un opérateur *et*.

Par exemple, dans la classe `tape-22.4`, le rôle Instrument est associé dans VerbNet [2] à la restriction de sélection Concrete & -Animate, ce qu'on peut représenter sous la forme AND(Concrete, NOT(Animate)). Le rôle Patient est lui simplement associé à la restriction Solid.

2. `http://verbs.colorado.edu/verb-index/vn/tape-22.4.php`

Pour chaque rôle annoté avec la correspondance exacte, le syntagme est alors associé à la restriction de sélection. Pour Instrument, ce sera Concrete (la seule restriction directement placé sous une relation *et*. Pour Patient, ce sera Solid, parce que c'est la seule restriction de sélection.

Après avoir réalisé les correspondances sur un fichier entier, nous utilisons ces informations pour donner un score à chaque frame : plus le score est haut, plus les mots étaient proches de ceux vu dans le fichier. Alors que dans l'expérience précédente, les mots tête des syntagmes prépositionnels excluaient les prépositions, elles sont ici conservées comme mot tête pour une meilleure généralisation face à la petite taille des données traitées. Toutes les frames ayant le même score sont conservées.

Cette expérience améliore les résultats de manière conséquente. La prochaine section évalue justement les différentes améliorations présentées jusqu'ici.

5.5. Évaluation

Nous évaluons notre système sur le corpus FrameNet. C'est un corpus équilibré largement utilisé pour l'annotation en rôles sémantiques. Pour ce corpus, diverses méthodes d'apprentissage supervisées ont fait leur preuves, l'état de l'art étant représenté par Das et al. [2014]. L'existence d'un mapping entre VerbNet et FrameNet (section 5.5.1) rend possible dans une certaine mesure l'évaluation de notre système basé sur VerbNet. L'objectif n'est pas de concurrencer les méthodes supervisées qui disposent de plus d'informations pour réaliser l'annotation, mais de savoir où se situe notre système en terme de performance.

5.5.1. Mapping VerbNet - FrameNet

Le projet SemLink a développé un mapping entre VerbNet et FrameNet. Pour chaque frame FrameNet, zéro, une ou plusieurs correspondances vers des classes VerbNet sont proposées. Ensuite, pour chaque paire (VerbNet, FrameNet), les rôles des deux ressources sont associés.

Le résultat (Figure 5.2) est un mapping n-n : une frame FrameNet peut correspondre à plusieurs classes VerbNet, et vice versa. C'est un problème de granularité dû à la construction différente des deux ressources [Palmer, 2009]. Cela implique qu'il est difficile d'utiliser un tel mapping pour convertir entre annotations VerbNet et FrameNet, et c'est une des limites de notre évaluation. Cela dit, les difficultés à appliquer un tel mapping sont réduites sur le corpus FrameNet : alors que seuls 50 % de tous les rôles sont globalement mis en correspondance, plus de 95 % des occurrences de rôles sont mis en correspondance pour le corpus FrameNet en particulier.

Nous avons modifié ce mapping pour ajouter des correspondances manquantes : en attendant

```
<vncls class='9.1' fnframe='Placing'>
    <roles>
    <role fnrole='Agent' vnrole='Agent'/>
    <role fnrole='Cause' vnrole='Agent'/>
    <role fnrole='Goal' vnrole='Destination'/>
    <role fnrole='Theme' vnrole='Theme'/>
    </roles>
</vncls>
<vncls class='9.1-1' fnframe='Installing'> <!-- added for 'install' -->
    <roles>
    <role fnrole='Agent' vnrole='Agent'/>
    <role fnrole='Fixed_location' vnrole='Destination'/>
    <role fnrole='Component' vnrole='Theme'/>
    </roles>
</vncls>
<vncls class='9.3' fnframe='Cause_impact'>
    <roles>
    <role fnrole='Agent' vnrole='Agent'/>
    <role fnrole='Impactor' vnrole='Theme'/>
    <role fnrole='Impactee' vnrole='Destination'/>
    </roles>
</vncls>
```

FIGURE 5.2. – Example de mapping VerbNet-FrameNet

que ces modifications que nous avons signalées soient prises en compte par SemLink, le mapping utilisé est disponible sur `https://github.com/aymara/knowledgesrl/blob/master/data/vn-fn-roles.xml`.

5.5.2. Détails expérimentaux

FrameNet dispose de deux corpus. Le premier corpus est le corpus d'exemples : des phrases extraites du British National Corpus pour illustrer la diversité de réalisation des frames. Plus tard, les versions 1.3, 1.4 et 1.5 de FrameNet ont introduit puis agrandi un corpus dit full-text, plus adapté à la tâche d'annotation en rôles sémantiques : au lieu d'identifier des exemples diversifiés pour toutes les frames, tous les prédicats présents dans un texte donné sont annotés en rôles sémantiques. L'objectif est de s'approcher au plus près des conditions réelles de l'annotation en rôles sémantiques. Ce corpus full-text est équilibré et inclut des textes de diverses sources : le Wall Street Journal, les corpus AQUAINT et MASC, ainsi que d'autres textes divers. Pour notre évaluation, nous annotons ce corpus full-text de FrameNet 1.5 et pas le corpus d'exemples qui est considéré non seulement plus facile mais aussi difficilement utilisable pour annoter le corpus

full-text [Das et al., 2010, section 2.1].

Nous utilisons VerbNet 3.2 et le mapping VerbNet-FrameNet 1.2.2c [3]. Notre système n'annote que les arguments Core étant donné que ce sont généralement les arguments présents dans VerbNet.

Pour la tâche complète qui inclut l'identification des arguments, nous utilisons le parser MST dans sa version 0.5.0 [McDonald et al., 2006] entraîné sur le Wall Street Journal. Nous avons modifié ce corpus de la façon suivante.

— Nous avons d'abord transformé l'encodage des syntagmes nominaux [4] [Vadas and Curran, 2007].
— Cela nous a permis d'appliquer l'outil de conversion constituants-dépendances du LTH pour une conversion au format CoNLL [5] [Johansson and Nugues, 2007], le parser MST étant un analyseur syntaxique en dépendances.
— FrameNet incluant des fichier du corpus du Wall Street Journal, nous avons supprimé les fichiers 0558, 0089, 0456, 1778, 1286 et 1695 du corpus d'entraînement de l'analyseur syntaxique MST. Ceci évite que l'analyseur syntaxique ait à traiter des phrases déjà observées dans le corpus d'entraînement, ce qui aurait pu améliorer les résultats artificiellement.

Nous avons ensuite utilisé ce modèle avec le script `fnParsedDriver.sh` fourni par SEMAFOR, qui utilise pour l'étiquetage morpho-syntaxique (*part-of-speech tagging*) MXPOST, le modèle étant inclus dans la distribution de SEMAFOR.

5.5.3. Procédure d'évaluation

Les rôles annotés de chaque phrase FrameNet sont d'abord transformés en rôles VerbNet. Pour chaque classe FrameNet, un rôle FrameNet peut correspondre à 0, 1 ou plusieurs rôles VerbNet.

Pour zéro rôle, il n'y pas de correspondance pour les frames FrameNet trop éloignées des classes VerbNet. C'est un problème répandu : seulement 4605 rôles sur les 10052 rôles présents dans FrameNet ont au moins une association VerbNet.

Il y a plusieurs correspondances quand un rôle FrameNet est ambigu par rapport à un rôle VerbNet. Les rôles FrameNet étant définis plus précisément, c'est un problème rare.

Enfin, le reste du temps, un rôle FrameNet correspond à un rôle VerbNet unique, et c'est ce rôle

3. `http://verbs.colorado.edu/semlink/1.2.2c/vn-fn/`
4. `http://sydney.edu.au/engineering/it/~dvadas1/`
5. `http://nlp.cs.lth.se/software/treebank_converter/`

Tâche	F1 (%)	Exactitude (%)
Arguments de la vérité-terrains		
Correspondance exacte	71.11	54.66
Correspondance exacte + passif	73.52	57.20
Correspondance exacte + WordNet	71.64	54.80
Correspondance exacte + restrictions	75.80	68.31
Correspondance exacte + passif + restrictions	78.29	70.99
Correspondance exacte + probabiliste	71.84	58.57
Correspondance exacte + bootstrap	73.63	69.41
Correspondance exacte + bootstrap + passif + restrictions	79.41	75.24
Arguments identifiés automatiquements		
Correspondance exacte	42.92	29.95
Correspondance exacte + passif	44.28	30.98
Correspondance exacte + restrictions	43.81	35.71
Correspondance exacte + restrictions + passif	45.13	36.95
Correspondance exacte + probabiliste	43.25	31.52
Correspondance exacte + bootstrap + passif + restrictions	46.00	39.29

TABLE 5.2. – Résultats pour différentes tâches. L'exactitude est l'*accuracy* et n'évalue donc que les correspondances non ambigües. La correspondance exacte est décrite à la section 5.2.3, *probabiliste* et *bootstrap* indiquent la correspondance probabiliste (section 5.2.4), *passif* indique la détection de la voix passive (section 5.3), *WordNet* indique la prise en compte des restrictions de sélection avec WordNet (section 5.4.1), alors que *restrictions* indique l'utilisation des restrictions de VerbNet après une première correspondance exacte (section 5.4.2). *Identification* est l'identification des arguments (section 5.2.2).

VerbNet que notre système doit déterminer. Nous mesurons la précision, le rappel et l'exactitude (*accuracy*) des associations rôle/syntagmes candidats. Le corpus de test est celui défini par [Das and Smith, 2011], une petite partie du corpus utilisée pour obtenir les résultats, le reste correspond au corpus "d'entraînement" traditionnellement utilisé en apprentissage automatique. En effet, aucun modèle n'a été appris sur ce corpus, mais c'est celui sur lequel nous examiné manuellement les erreurs de notre système.

5.6. Résultats

5.6.1. Analyse des résultats

La Table 5.2 montre les résultats des deux tâches sur lesquelles nous nous évaluons : la correspondance exacte seule d'une part et la correspondance exacte accompagnée de l'identification des arguments en amont d'autre part. La correspondance exacte seule utilise les arguments de la vérité-terrain : on sait que ce syntagme joue un rôle, il faut alors déterminer lequel. L'identification des arguments implique que le texte de départ est une phrase brute : il faut l'analyser syntaxiquement, identifier les arguments, puis réaliser la correspondance exacte à proprement parler.

Le premier enseignement que l'on peut tirer de l'observation de ces résultats est que l'identification des arguments doit être améliorée de manière significative car elle pénalise fortement les étapes ultérieures. En effet :

— sans améliorations, la précision lors de la correspondance exacte n'est que de 60 %, contre plus de 90 % avec les arguments de la vérité-terrain ;
— les différents améliorations ont un impact très faible, contrairement aux expériences menées avec les arguments de la vérité-terrain.

La première difficulté est que seulement 78 % des syntagmes jouant un rôle sont effectivement des sous-arbres de l'analyse syntaxique (section 5.2.2). La seconde raison provient des heuristiques utilisées pour l'identification elle-même : une analyse plus poussée permettrait de mieux comprendre les erreurs qu'elles causent. Des alternatives existent, supervisées ou non supervisées [Abend et al., 2009]. Ainsi, si l'identification des arguments n'était pas parmi nos objectifs pour ce chapitre, l'analyse des résultats nous montre que c'est un travail futur capital.

5.6.2. Absence de comparaison avec SEMAFOR

SEMAFOR [Das et al., 2014] est la référence actuelle en annotation en rôles sémantiques supervisée : c'est le système qui obtient les meilleurs résultats sur le corpus full-text de FrameNet 1.5. Sans réussir à relancer nous-même SEMAFOR [6] pour s'évaluer sur le même sous-ensemble du corpus FrameNet, une comparaison directe n'est pas possible :

— SEMAFOR annote le corpus FrameNet en frames FrameNet et rôles FrameNet alors que nous l'annotons en classes VerbNet et rôles VerbNet *seulement quand une correspondance existe*.

6. L'échec d'un collègue rencontré en conférence à utiliser le code de SEMAFOR et les modèles fournis ne nous a pas encouragé vers cette voie.

— Toutes les parties du discours sont annotées alors que nous nous concentrons sur les verbes.
— Les tâches sont découpées différemment. En effet, SEMAFOR découpe la tâche en trois parties :
  — identification des prédicats déclencheurs ;
  — identification des frames FrameNet ;
  — identification des arguments et annotation ces arguments avec des rôles sémantiques.

Nous considérons par la suite uniquement les résultats de SEMAFOR avec déclencheurs issus de la vérité-terrain : en effet, cette sous-tâche n'est pas pertinente pour une annotation VerbNet ou PropBank [Das et al., 2014, section 4], et ce n'est de toute façon que sur les prédicats de la vérité-terrain que nous nous évaluons.

Il est intéressant de noter l'importance des données d'entraînement pour SEMAFOR : pour l'identification des frames FrameNet les mêmes modèles grimpent de 74.21 % à 90.51 % quand la taille du corpus augmente en passant du corpus SemEval (2198 phrases) au corpus FrameNet 1.5 (3 256 phrases) [Das et al., 2014, section 4]. De la même manière, pour l'identification des arguments, les résultats augmentent de 46.49 % à 64.54 %. C'est très encourageant pour les domaines disposant de très gros corpus, mais suggère que d'autres solutions sont à identifier pour les domaines où de tels corpus ne sont pas disponibles.

5.7. Travaux futurs

Le problème principal est la faible performance de l'identification des arguments : ce n'était pas là-dessus que nous voulions porter nos efforts, mais il devient clair que c'est un problème conséquent (section 5.6.1) qui est la priorité pour nos travaux futurs.

Nous pensons aussi prendre en compte la similarité entre les remplisseurs déjà identifiés et les remplisseurs pour lesquels plusieurs rôles sont possibles afin d'améliorer nos modèles de probabilité. En effet, l'information des cadres de sous-catégorisation est cruciale pour identifier les arguments, mais l'information sémantique concernant le contenu des remplisseurs est aussi utile pour déterminer le rôle correct, comme nous l'avons vu à la section 5.4.2.

Enfin, de la même manière que la prise en compte de la voix passive a amélioré les résultats, d'autres phénomènes de syntaxe profonde doivent être pris en compte. La coordination est une autre source commune d'erreur. En effet, quand deux verbes partagent le même sujet, une analyse syntaxique profonde indique à chaque fois quel est le sujet profond. Voici deux exemples tirés du corpus FrameNet :

— *You belittle Sheik Bin Baz 's blunder and exaggerate the one by Sheik Maqdasi ...*

— *Hostile and even friendly nations routinely steal information from U.S. companies and share it with their own companies*

Dans la première phrase, *you* est détecté comme le sujet du verbe *belittle*, mais c'est aussi le sujet de *exaggerate*, ce qui n'est pas détecté. Le problème se pose aussi pour la deuxième phrase *Hostile and even friendly nations* est le sujet de *share*, ce qui n'est pas détecté.

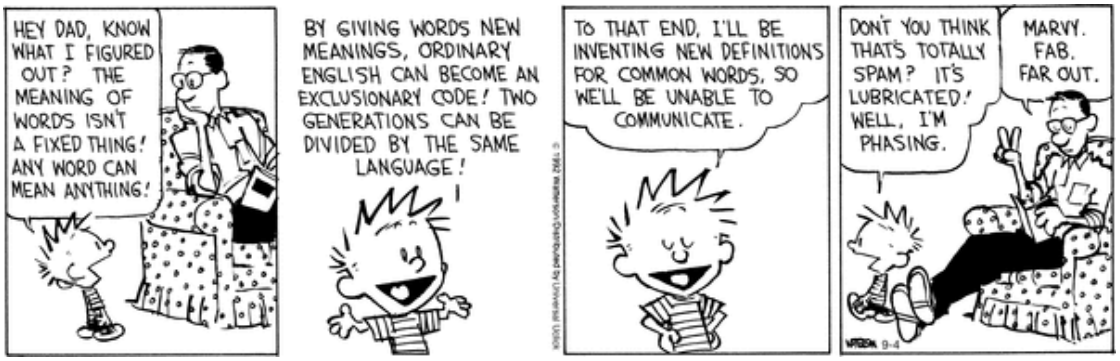
Il serait certes possible d'identifier des règles complexes pour traiter ces cas lors de l'identification des arguments, mais l'objectif est de traiter ces phénomènes de manière plus générale en intégrant par exemple le système de Ribeyre [2013] qui est conçu pour ce genre de problème et permet de prendre en compte de nouveaux phénomènes en ajoutant de nouvelles règles au système. Ainsi, les différents phénomènes seront pris en compte de manière cohérente.

Conclusion

Nous avons implémenté un système d'annotation en rôles sémantiques basé sur la connaissance. Nous avons utilisé des outils et des corpus disponibles publiquement qui rendent notre travail facilement reproductible et facilitent le travail de comparaison, maintenant et dans le futur. Nous avons commencé à améliorer le système initial, montrant son potentiel. L'indépendance de l'approche par rapport au corpus considéré la rend attractive pour annoter des domaines ne disposant que de peu ou pas de corpus annotés en rôles sémantiques.

C'est ce que nous explorons dans le chapitre suivant en étudiant la pertinence de VerbNet face à des domaines spécifiques et variés : football, réchauffement climatique et informatique.

6. Annotation en rôle sémantique en domaine spécifique



(Bill Watterson, Calvin et Hobbes)

Là où le précédent chapitre présentait notre approche générique d'annotation en rôles sémantiques fondée sur la connaissance, celui-là cherche à évaluer son application en domaines spécifiques. À l'heure où les algorithmes supervisés obtiennent d'excellentes performances sur un certain nombre de tâches quand des données annotées pour le domaine considéré existent, l'adaptation au domaine est un défi majeur en Traitement Automatique des Langues, et l'annotation en rôles sémantiques sur de nouveaux domaines reste un problème ouvert.

Dans ce chapitre, nous n'utilisons pas le système du Chapitre 5 et ne développons pas un système, mais nous évaluons la capacité de VerbNet et Verb∋Net à traiter des corpus en domaines spécifiques.

La définition du terme 'domaine' reste assez vague dans le sens où il est difficile de proposer une catégorisation de la connaissance en un ensemble de domaines qui soit à la fois cohérente et efficace [Ma and Fellbaum, 2012]. Il est aussi difficile de séparer clairement le « domaine général » des domaines spécifiques. Néanmoins, c'est un phénomène qui existe et qui est à considérer : les modèles entraînés sur un corpus d'un domaine spécifique (la finance par exemple) risque de produire des performances mauvaises si appliqués à d'autres domaines (le football par exemple). C'est le cas par exemple de la désambiguïsation de classe VerbNet au moment de passer du corpus PropBank à un corpus du domaine biomédical [Abend et al., 2008].

Il est important aussi de différencier le genre et le domaine d'un texte. Le corpus du Wall Street Journal traite du domaine de la finance dans un genre journalistique. D'autres genres existent,

par exemple les genre littéraires que sont la fiction, la poésie, le théâtre, etc. Néanmoins, ces distinctions ne nous concernent pas ici, et l'objectif affiché est de traiter aussi bien un roman qu'une encyclopédie, un email bien écrit qu'un article de journal. C'est une simplification mais les difficultés que nous rencontrons avec les corpus utilisés dépendent plutôt du domaine que du genre. En effet, nous supposons que dans deux domaines différents plus que dans deux genres différents, les changements vont résider dans les verbes utilisés, leur sens et la façon de les utiliser. C'est ce que nous souhaitons prendre en compte ici.

6.1. Corpus considérés

Afin de s'assurer que notre travail reste valide en changeant de domaine, nous considérons ici trois corpus présents dans des domaines différents.

— le Kicktionary [Schmidt, 2006, 2009] rassemblant des dépêches de l'UEFA dans le domaine du football à propos de la Ligue Europa, de la Champions League et de la Coupe du Monde. Ce corpus est disponible en français, anglais et allemand.
— Le DiCoInfo est le corpus Informatique/Internet de l'OLST [OLST, 2014] en anglais, français et espagnol.
— Le DiCoEnviro est le corpus Réchauffement climatique de l'OLST [OLST, 2014] en anglais, français et espagnol.

Les intérêts majeurs communs à ces trois corpus sont :

1. de s'inspirer plus ou moins librement de la méthodologie FrameNet, ce qui permet d'appliquer nos méthodes,
2. d'être disponible à la fois en anglais et en français, ce qui permet de comparer VerbNet à Verb∋Net,
3. et d'être basés sur l'annotation dans un corpus d'un certain nombre de prédicats (ce n'est pas une annotation dite "full-text")

Ce dernier point est plutôt un inconvénient : la manière la plus réaliste de considérer un corpus d'entraînement est de réaliser une annotation dite *full-text* en annotant tous les prédicats rencontrés dans un texte donné. Cependant, les contextes choisis l'ont été à chaque fois sur des critères de diversité syntaxique [Schmidt, 2006, L'Homme, 2012]. Le postulat que nous faisons dans ce travail est que notre méthode est moins affectée par ce problème qu'un modèle statistique se basant exclusivement sur la distribution des différentes constructions.

Les trois corpus ont étés découpés en deux parties de tailles égales : un corpus d'entraînement et un corpus de test. Étant donné qu'il n'y a pas de modèle à entraîner, ce corpus d'entraînement est simplement utilisé pour comprendre les erreurs en l'analysant manuellement. Même

si les trois corpus contiennent des textes proches entre eux, ils sont issus de sources différentes et ont étés écrits par des auteurs différents. Nous avons normalisé ces sources (par exemple, les sources DEBIAN2 et DEBIAN3 dans le corpus Informatique & Internet de l'OLST ont manuellement été transformées vers DEBIAN) et utilisons cette information pour faire le découpage : une source spécifique ne peut être présente que dans un ensemble. Le découpage résultant est disponible au format JSON avec le code source associé à ce chapitre. Ce découpage ne nécessite pas d'avoir normalisé les sources : chaque phrase est simplement associée à l'ensemble de test ou l'ensemble d'entraînement.

6.2. Mappings de rôles

Ces trois corpus utilisent des rôles spécifiques. Les corpus DiCoInfo et DiCoEnviro utilisent les conventions VerbNet alors que le Kicktionary définit un nouvel ensemble de rôles pour chaque frame à la façon de FrameNet, par exemple `Passer`, `Moving_Ball` ou `Shot`. Dans les trois cas, les rôles ne correspondent pas directement aux rôles VerbNet, et il faut donc établir une correspondance entre les rôles VerbNet et les rôles des trois corpus. Nous avons assigné manuellement une classe VerbNet à chaque classe des trois corpus, tout en faisant correspondre les rôles. Le résultat de ce mapping est disponible aux URLs suivantes :

— DiCoInfo anglais : `https://github.com/aymara/knowledgesrl/blob/master/data/domain/info/vnroles_info_en.xml`
— DiCoEnviro anglais : `https://github.com/aymara/knowledgesrl/blob/master/data/domain/enviro/vnroles_enviro_en.xml`
— Kicktionary : `https://github.com/aymara/knowledgesrl/blob/master/data/domain/kicktionary/kicktionary-vn-roles.xml`

Dans le cas de DiCoInfo et DiCoEnviro, les noms des rôles sont proches des noms des rôles employés dans VerbNet et LIRICS [Bonial et al., 2011] : Agent, Patient, Destination, Instrument, etc. Cependant, même si les noms sont les mêmes, la définition de ces rôles est spécifique à DiCoInfo et DiCoEnviro.

Cependant, même si les noms sont les mêmes, la définition de ces rôles est spécifique à DiCoInfo et DiCoEnviro. C'est parfois systématique. Par exemple, DiCoInfo et DiCoEnviro ne distinguent pas Theme de Patient mais n'utilisent que Patient, ce qui facilite l'annotation. [1]

C'est parfois spécifique à une situation. Prenons deux phrases d'exemple dans DiCoInfo et DiCoEnviro pour illustrer les associations.

1. Ce choix est le bienvenu, étant donné que la distinction entre Theme et Patient est souvent difficile à établir. Dans *Le chaton a léché mes doigts*, est-ce que mes doigts ont changé d'état ? Si oui, ils devraient être Patient, et sinon, Theme [Palmer et al., 2010, p. 5].

— *In the interest of fair competition you should ALLOCATE the same amount of memory to both engines*.
— *Techniques and tools exist to MEASURE carbon stocks in project areas relatively precisely depending on the carbon pool.*

Dans la première phrase, le sens du verbe *allocate* au sens *allouer de la mémoire vive* est très précis et très spécifique au domaine de l'informatique. Pourtant, il se comporte syntaxiquement de la même manière que le sens plus général considéré par WordNet et OntoNotes : *distribute or set aside according to plan*. Par conséquent, un mapping manuel a été réalisé de la lexie allocate.1 (qui correspond à la phrase ci-dessus) vers la classe VerbNet *future_having-13.3*. Enfin, les actants définis par DiCoInfo (qui correspondent aux roles *Core* de FrameNet et aux rôles de VerbNet) ont étés mis en correspondance avec les rôles de VerbNet :

— Patient devient Theme ;
— Recipient devient Goal ;
— Agent reste Agent.

Une fois que ce mapping est réalisé, la tâche de notre algorithme d'annotation en rôles sémantiques devient de détecter que la classe `future_having-13.3` est utilisée ici, que *You* est Agent, que *the same amount of memory* est Theme, et que *to both engines* est Goal.

Pour la seconde phrase, la démarche est la même : il s'agit d'identifier que la classe Verbnet est `register-54.1`, que *Techniques and tools* est Agent, et que *carbon stocks* est Theme.

DiCoInfo et DiCoEnviro font une distinction entre les rôles core et non core : nous n'avons annoté que les rôles core étant donné que VerbNet ne considère que ces rôles, même si la distinction peut différer entre VerbNet et DiCoInfo/DiCoEnviro. Étant donné qu'une frame DiCoInfo/DiCoEnviro est un sens spécifique d'un verbe n'acceptant que des constructions spécifiques, nous avons toujours associé de tels sens à une seule classe VerbNet.

Le Kicktionary a été plus difficile à associer : ses frames considèrent un grand nombre de verbes au comportement parfois assez différent. Les règles de FrameNet indiquent que de telles frames auraient du êtres découpées [Ruppenhofer et al., 2006], mais le Kicktionary ne suit pas toujours ces règles, ayant été développé à part de FrameNet, couvrant un domaine spécifique au lieu du domaine général, et étant multilingue [Schmidt, 2006]. Certaines frames sont définies correctement : c'est le cas de `Receive_Card` et `Give_Card` qui correspondent respectivement à `obtain-13.5.2` et `give-13.1-1`. Par contre, la frame `Goal` a par exemple été définie pour un but marqué, mais les différentes façons de l'exprimer n'ont pas étés séparées : ouvrir la marque, égaliser, frapper, et d'autres utiliseront différentes constructions syntaxiques, évoqueront des rôles différents et correspondront donc à des classes VerbNet différentes. Nous n'évaluons pas ces frames.

6.3. Comparaison à VerbNet

De simples transformations permettent de gérer l'encodage spécifique de frames dans DiCoInfo et DiCoEnviro pour l'adapter à VerbNet. Premièrement, des rôles répétés sont supprimés : `NP.Agent V NP.Theme NP.Theme` devient `NP.Agent V NP.Theme`. En effet, DiCoInfo et DiCoEnviro répètent le même rôle deux fois quand deux syntagmes nominaux liés par exemple avec la conjonction « et » partagent le même rôle dans la même frame. Deuxièmement, au moment de rencontrer une forme `V NP.Theme`, le syntagme devant le verbe est aussi supprimé dans VerbNet avant comparaison. Ces formes sont en général des verbes à l'impératif où le sujet n'est pas exprimé, et doivent dont êtres transformées avant traitement par VerbNet, VerbNet ne décrivant les frames qu'à l'indicatif.

Dans le cas de Verb∋Net, pour prendre en compte les différences d'encodage (expliquées à la section 4.3.1), nous générons les frames qu'il faut inférer. Pour ce faire, nous prenons l'ensemble des arguments (syntagmes après le verbe), et calculons l'ensemble des permutations des parties de cet ensemble, avec produire des frames complètes à partir de ces permutations.

Pour chaque occurrence de phrase dans notre corpus convertie en rôles VerbNet, nous :
- vérifions si le lemme existe dans VerbNet ou Verb∋Net,
- observons si la frame syntaxique du corpus est exprimée exactement telle quelle dans VerbNet,
- regardons si la classe souhaitée d'après le mapping est effectivement parmi la liste des classes VerbNet possibles après avoir filtré les correspondances syntaxiques,
- et déterminons enfin si les rôles sont corrects en considérant la classe correcte.

Ces quatre observations nous permettent d'évaluer à la fois la couverture de VerbNet et Verb∋Net (nombre de lemmes présents, nombres de constructions présentes) mais aussi en terme de précision (est-ce que les rôles sont corrects ?).

La méthode décrite ci-dessus est implémentée dans le dossier `src/domain` de `https://github.com/aymara/knowledgesrl`.

6.4. Résultats

Nous considérons la couverture de VerbNet dans ces trois ressources, en mesurant la couverture des lemmes, la couverture des classes, la couverture syntaxique et l'exactitude des rôles (section 6.3). Nous calculons ces scores sur les occurrences et non pas les types.

La Table 6.1 présente les résultats pour les trois domaines considérés. Les résultats du Kicktionary sont à analyser avec prudence : d'une part, le mapping n'est pas encore satisfaisant (section 6.2) d'où le score de classe extrêmement bas, et d'autre part aucune analyse d'erreur

	Info	Enviro	Kicktionary
Lemme verbal présent	80	89	69
Frame présente	80	82	69
Classe correcte incluse	57	79	11
Rôle correct	94	95	67

TABLE 6.1. – Couverture VerbNet en anglais pour les corpus DiCoInfo, DiCoEnviro et Kicktionary. Le score de lemme est le pourcentage d'occurrences de lemmes verbaux présents dans VerbNet. Le score de frame est le pourcentage de correspondances exactes entre les cadres de sous-catégorisation VerbNet et cadres identifiés dans les corpus. Le score de classe est le pourcentage de classes correctes qui sont présentes d'après notre mapping quand la frame était dans VerbNet, indépendemment de l'ambiguïté. Enfin, le score de rôle est le pourcentage de rôles correctement identifiés.

n'a pour le moment été effectuée. Nous pouvons tout de même tirer des conclusions de ces résultats.

Premièrement, VerbNet couvre entre 69% et 89% des occurrences de lemmes, ce qui est très encourageant et correspond à notre objectif de couvrir l'ensemble du vocabulaire. Deuxièmement, les résultats varient par domaines : en particulier, le Kicktionary est le corpus le plus loin de VerbNet. Cependant, l'écart surtout flagrant au moment de voir si la classe correcte est incluse et est dû à la difficulté d'associer des classes VerbNet aux situations finalement très générales du Kicktionary. Au-delà de ça, les erreurs sont principalement dues à des mots « du domaine général » qui ne sont pas spécifiques au football. Par exemple, le verbe *celebrate* est absent de VerbNet mais il serait tout de même utile dans le domaine général. Troisièmement, une fois que la classe a été identifiée correctement, les résultats sont bons pour les corpus DiCoInfo et DiCoEnviro (plus de 90 %).

	Info	Enviro	Kicktionary
Lemme verbal présent	52	37	42
Frame présente	78	84	59
Classe correcte incluse	47	46	28
Rôle correct	78	69	68

TABLE 6.2. – Couverture Verb∋Net en français pour les corpus DiCoInfo, DiCoEnviro et Kicktionary. Verb∋Net a été développé de manière complètement indépendante : aucune instruction n'a été donnée pour obtenir de meilleurs résultat ici, à part de traiter la classe 45, présente dans beaucoup de cas dans DiCoEnviro.

De manière attendue, Verb∋Net, encore en développement, a une couverture plus faible en termes de lemmes présents et de classes incluses, mais les résultats sont prometteurs (Table 6.2).

Plus particulièrement, ces résultats seront l'occasion d'améliorer la couverture de Verb∋Net en utilisant les informations manquantes pour continuer à améliorer Verb∋Net dans le but de répondre aux besoins du Traitement Automatique des Langues.

Conclusion

Nous avons montré qu'il est possible de réaliser de l'annotation en rôles sémantique en domaine spécifique en utilisant VerbNet. Le système complet d'annotation en rôles sémantiques utilisera les techniques décrites au Chapitre 5.

Notre approche rend possible l'annotation d'un nouveau domaine avant de passer à une annotation manuelle. Il est aussi possible de l'utiliser comme une simple baseline contre des approches plus sophistiquées. Enfin, utiliser cet outil sur de nouveaux domaines est un moyen efficace d'obtenir de meilleures performances mais aussi de guider de nouveaux développements VerbNet étant donné que les lemmes, classes et frames manquants sont montrées.

Remerciements

Merci à Marie-Claude L'Homme et Thomas Schmidt pour nous avoir fourni les corpus DiCoInfo, DiCoEnviro et Kicktionary.

7. Conclusion

Nous avons dans cette thèse contribué au Traitement Automatique des Langues du français de deux manières : en améliorant l'état des ressources du français avec nos traductions de WordNet et de VerbNet : WoNeF et Verb∋Net. Nous avons aussi eu la volonté d'utiliser ces ressources en annotation en rôles sémantiques afin d'apporter au français des méthodes faciles à mettre en œuvre utilisant nos ressources et ne demandant pas de corpus annoté.

WordNet dispose de différentes traductions pour le français que nous espérons voir s'unifier dans les années qui viennent. Nos travaux montrent en tout cas qu'il est encore pertinent de travailler sur la traduction de WordNet : il y a encore un besoin important pour gagner en couverture et en précision, chose qui sera facilitée en coordonnant l'ensemble des travaux autour de la traduction de WordNet.

Il n'existait pas de traduction de VerbNet pour le français, et nous espérons que cette ressource continuera à vivre au-delà de la thèse qui a permis de la créer. Dans un premier temps, le développement du site web permettant l'édition va continuer afin que les lexicographes puissent aboutir à une première traduction complète de VerbNet dans les mois qui viennent. Dans un deuxième temps, nous avons souhaité assurer un futur à cette ressource en facilitant au maximum son utilisation et son amélioration. Pour ceci, nous avons réutilisé le format VerbNet, utilisé une licence permissive et encouragé un développement ouvert où chacun peut participer à élaborer et à améliorer la ressource, soit en remontant les erreurs aux auteurs soit en utilisant directement l'outil d'édition, accessible à tous.

Nous avons aussi pu mettre en valeur VerbNet et Verb∋Net en utilisant la ressource pour l'annotation en rôles sémantiques : c'est une méthode prometteuse en l'absence de corpus annoté qui peut s'appliquer à toute langue disposant d'un VerbNet et d'un analyseur syntaxique. En effet, VerbNet et Verb∋Net affichent des résultats encourageants sur les corpus en domaines spécifiques.

Nos perspectives sont de porter plus d'attention à la syntaxe, à la manière des évaluations autour du *broad-coverage semantic parsing* [Oepen et al., 2014]. En effet, bien que ce n'était pas l'objet de nos travaux, c'est en réalité la syntaxe plus que la sémantique qui fait défaut à notre système, et il est désormais indispensable de s'appuyer sur des techniques solides d'analyse syntaxique profonde afin de mieux traiter toute la diversité syntaxique présente dans les corpus que nous étudions.

Il faudra ensuite être capable de mieux traiter divers phénomènes tels que la négation et les phrases conditionnelles. C'est un problème reconnu aussi par FrameNet [Baker, 2014]. La compositionnalité de l'annotation en rôles sémantiques est un problème majeur aussi : les phrases n'utilisant qu'un seul verbe sont rares, et extraire du sens quand plusieurs prédicats sont impliqués n'est pas un problème résolu, même si on sait qu'annoter conjointement les différents prédicats d'une phrase améliore l'annotation en rôles sémantiques [Yang and Zong, 2014].

Enfin, et surtout, nous souhaitons utiliser notre approche dans diverses applications. Notre système d'annotation en rôles sémantiques a été intégré dans LIMA : les premiers tests porteront sur l'intérêt de l'utilisation de l'annotation en rôles sémantiques dans le moteur de recherche AMOSE qui se base sur LIMA.

Bibliographie

Omri Abend, Roi Reichart, and Ari Rappoport. A supervised algorithm for verb disambiguation into VerbNet classes. In COLING 2008, 2008. URL http://www.aclweb.org/anthology/C08-1002.

Omri Abend, Roi Reichart, and Ari Rappoport. Unsupervised argument identification for semantic role labeling. In ACL-IJNCLP 2009, Singapore, August 2009.

E. Agirre and A. Soroa. Personalizing PageRank for Word Sense Disambiguation. In EACL 2009, 2009.

Eneko Agirre, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. Semeval-2010 task 17 : All-words word sense disambiguation on a specific domain. In SemEval-2010, Uppsala, Sweden, July 2010. URL http://www.aclweb.org/anthology/S10-1013.

Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. Towards building kurdnet, the kurdish wordnet. In GWC 2014, Tartu, Estonia, January 2014. URL http://www.aclweb.org/anthology/W14-0101.

Marianna Apidianaki and Benoît Sagot. Applying cross-lingual WSD to wordnet development. In LREC'12, May 2012. ISBN 978-2-9517408-7-7.

C. Baker, M. Ellsworth, and K. Erk. SemEval'07 task 19 : Frame Semantic Structure Extraction. In SemEval'07, page 99–104, Prague, Czech Republic, July 2007.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. The Berkeley FrameNet project. In ACL-COLING 98, Montréal, Canada, August 1998.

Collin Baker. Framenet : A knowledge base for natural language processing. In Proceedings of Frame Semantics in NLP : A Workshop in Honor of Chuck Fillmore (1929-2014), Baltimore, MD, USA, June 2014. URL http://www.aclweb.org/anthology/W14-3001.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In 7th Linguistic Annotation Workshop and Interoperability

with Discourse (LAW VII & ID), Sofia, Bulgaria, August 2013. URL http://www.aclweb.org/anthology/W13-2322.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict ! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL 2014, page 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1023.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. Developing a large semantically annotated corpus. In LREC'12, Istanbul, Turkey, may 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/534_Paper.pdf.

Marzieh Bazrafshan and Daniel Gildea. Semantic Roles for String to Tree Machine Translation. In ACL 2013, Sofia, Bulgaria, August 2013. URL http://www.aclweb.org/anthology/P13-2074.

Marzieh Bazrafshan and Daniel Gildea. Comparing representations of semantic roles for string-to-tree decoding. In EMNLP 2014, Doha, Qatar, October 2014. URL http://www.aclweb.org/anthology/D14-1188.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In NIPS 2000, 2001. URL http://www.iro.umontreal.ca/~lisa/pointeurs/nips00_lm.ps.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3 : 1137–1155, 2003. URL http://www.iro.umontreal.ca/~lisa/pointeurs/BengioDucharmeVincentJauvin_jmlr.pdf.

Anderson Bertoldi and Rove Chishman. Frame semantics and legal corpora annotation : theoretical and applied challenges. Linguistic Issues in Language Technology, 7(1), 2012.

Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Faiza Gara, and Nasredine Semmar. LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. In LREC 2010, May 2010.

Hans C Boas. Multilingual FrameNets in computational lexicography : Methods and applications, volume 200. Walter de Gruyter, 2009.

Daniel G Bobrow, Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price, and Annie Zaenen. Parc's bridge and question answering system. In Proc. of the GEAF 2007 Workshop. CSLI Studies in Computational Linguistics Online, 2007.

Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), Sofia, Bulgaria, August 2013. URL http://www.aclweb.org/anthology/P13-1133.

Francis Bond and Kyonghee Paik. A Survey of WordNets and their Licenses. In GWC 2012, page 64–71, 2012.

Guillaume Bonfante, Bruno Guillaume, Mathieu Morey, and Guy Perrier. Modular graph rewriting to compute semantics. In IWCS 2011, Oxford, UK, February 2011. URL http://aclweb.org/anthology/W11-0108.

Claire Bonial, William Corvey, Martha Palmer, VolhaV Petukhova, and Harry Bunt. A hierarchical unification of LIRICS and VerbNet semantic roles. In Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on, page 483–489. IEEE, 2011.

Claire Bonial, Weston Feely, Jena D Hwang, and Martha Palmer. Empirically Validating VerbNet Using SemLink. In Seventh Joint ACL-ISO Workshop on Interoperable Semantic Annotation, Istanbul, Turkey, May 2012.

Claire Bonial, Orin Hargraves, and Martha Palmer. Expanding VerbNet with Sketch Engine. In Conference on Generative Approaches to the Lexicon (GL2013), Pisa, Italy, September 2013.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. PropBank : Semantics of New Predicate Types. In LREC'14, Reykjavik, Iceland, may 2014.

Jean Paul Boons, Alain Guillet, and Christian Leclère. La structure des phrases simples en français : constructions intransitives. Librairie Droz, 1976.

Antoine Bordes, Nicolas Usunier, Ronan Collobert, and Jason Weston. Towards understanding situated natural language. In AISTATS 2010, Sardinia, Italy, May 2010.

Emanuela Boros, Romaric Besançon, Olivier Ferret, and Brigitte Grau. Étiquetage en rôles événementiels fondé sur l'utilisation d'un modèle neuronal. In TALN 2014, page 25–35, Marseille, France, July 2014. URL http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-003.

Johan Bos, Kilian Evang, and Malvina Nissim. Annotating semantic roles in a lexicalised grammar environment. In Workshop on Interoperable Semantic Annotation (isa-8), Pisa, Italy, 2012. URL http://sigsem.uvt.nl/isa8/isa8_submission_12-1.pdf.

Thorsten Brants and Alex Franz. Web 1T 5-gram Version 1. Linguistic Data Consortium, 2006. ISBN 1-58563-397-6.

Susan Windisch Brown and Martha Palmer. Semantic Annotation of Metaphorical Verbs with VerbNet : A Case Study of 'Climb'and 'Poison'. In Workshop on Interoperable Semantic Annotation, page 72, 2012.

Paul Buitelaar. CoreLex : systematic polysemy and underspecification. PhD thesis, Brandeis University, 1998.

Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël De Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot, and Laure Vieu. Developing a french framenet : Methodology and first results. In LREC'14, Reykjavik, Iceland, May 2014.

Xavier Carreras and Lluís Màrquez. Introduction to the CoNLL-2005 shared task : Semantic role labeling. In CoNLL-2005, Ann Arbor, Michigan, June 2005. URL `http://www.aclweb.org/anthology/W05-0620`.

Angel Chang, Manolis Savva, and Christopher D. Manning. Learning Spatial Knowledge for Text to 3D Scene Generation. In EMNLP 2014, Doha, Qatar, October 2014. URL `http://www.aclweb.org/anthology/D14-1217`.

David L Chen and Raymond J Mooney. Learning to sportscast : a test of grounded language acquisition. In ICML 2008, Helsinki, Finland, July 2008. ACM.

Alexander Chuchunkov, Alexander Tarelkin, and Irina Galinskaya. Applying HMEANT to English-Russian Translations. In SSST-8, Doha, Qatar, October 2014. URL `http://www.aclweb.org/anthology/W14-4005`.

A. Copestake and A. Herbelot. Lexicalised Compositionality. Unpublished draft, 2013.

Fred J. Damerau. A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3) :171–176, March 1964. ISSN 0001-0782. doi : 10.1145/363958.363994. URL `http://doi.acm.org/10.1145/363958.363994`.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. Investigating regular sense extensions based on intersective Levin classes. In ACL-COLING 98, 1998.

Laurence Danlos, Takuya Nakamura, and Quentin Pradet. Vers la création d'un VerbeNet du français. In Atelier FondamenTAL, TALN 2014, July 2014.

Amitava Das, Sivaji Bandyaopadhyay, and Björn Gambäck. The 5W Structure for Sentiment Summarization-Visualization-Tracking. In CICLING 2012, New Delhi, India, March 2012.

D. Das, N. Schneider, D. Chen, and N.A. Smith. Probabilistic frame-semantic parsing. In HLT-NAACL 2010, page 948–956, Los Angeles, California, USA, June 2010.

Dipanjan Das and Noah A. Smith. Semi-Supervised Frame-Semantic Parsing for Unknown Predicates. In ACL 2011, Portland, Oregon, USA, 2011.

Dipanjan Das, Desai Chen, André F.T. Martins, Nathan Schneider, and Noah A. Smith. Frame-Semantic Parsing. Computational Linguistics, 40(1) :9–56, March 2014.

Orphée De Clercq, Michael Schuhmacher, Simone Paolo Ponzetto, and Véronique Hoste. Exploiting FrameNet for Content-Based Book Recommendation. In CBRecSys 2014 at ACM RecSys, Foster City, California, USA, October 2014. URL `http://ceur-ws.org/Vol-1245/cbrecsys2014-paper03.pdf`.

Éric Villemonte De La Clergerie. Jouer avec des analyseurs syntaxiques. In TALN 2014, 2014.

Gerard de Melo and Gerhard Weikum. On the Utility of Automatically Generated Wordnets. In GWC 2008, January 2008. ISBN 978-963-482-854-9.

Gerard de Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In CIKM 2009, November 2009.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. JASIS, 41(6) :391–407, 1990.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In ACL 2014, page 1370–1380, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-1129`.

Quoc-Khanh Do, Alexandre Allauzen, and François Yvon. Modèles de langue neuronaux : une comparaison de plusieurs stratégies d'apprentissage. In TALN 2014, Marseille, France, July 2014. URL `http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-023`.

Bonnie J Dorr, Mari Olsen, Nizar Habash, and Scott Thomas. LCS verb database. Online Software Database of Lexical, 2001.

Jean Dubois and Claude Dubois. Introduction à la lexicographie : le dictionnaire. Larousse, Paris, 1971.

Jean Dubois and Françoise Dubois-Charlier. Éléments de linguistique française : syntaxe. Larousse, Paris, 1970.

Jean Dubois and Françoise Dubois-Charlier. Les verbes français. Larousse, 1997.

Helge Dyvik. Translations as Semantic Mirrors : From Parallel Corpus to WordNet. In ICAME 23, May 2002.

P. Edmonds and A. Kilgarriff. Introduction to the special issue on evaluating word sense disambiguation systems. Natural Language Engineering, 8(4) :279–291, 2002.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. Measuring Word Meaning in Context. Computational Linguistics, 39(3), september 2013.

Peter Exner and Pierre Nugues. Using Semantic Role Labeling to Extract Events from Wikipedia. In DeRiVE 2011, Bonn, Germany, October 2011.

Peter Exner, Marcus Klang, and Pierre Nugues. Using distant supervision to build a proposition bank. In SLTC 2014, Uppsala, Sweden, November 2014. URL `http://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_4.pdf`.

Cécile Fabre, Nabil Hathout, Lydia-Mai Ho-Dac, François Morlane-Hondère, Philippe Muller, Franck Sajous, Ludovic Tanguy, and Tim Van De Cruys. Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In TALN 2014, atelier SemDis, Marseille, France, June 2014. URL `https://hal.archives-ouvertes.fr/hal-01022216`.

Ingrid Falk, Claire Gardent, and Jean-Charles Lamirel. Classifying French Verbs Using French and English Lexical Resources. In ACL 2012, July 2012. URL `http://www.aclweb.org/anthology/P12-1090`.

Christiane Fellbaum, editor. WordNet : An Electronic Lexical Database. MIT Press, Cambridge, MA, May 1998.

Christiane Fellbaum and Piek Vossen. Connecting the Universal to the Specific : Towards the Global Grid. In IWIC 2007, January 2007.

Eva Esteve Ferrer. Towards a Semantic Classification of Spanish Verbs Based on Subcategorisation Information. In ACL 2004 : Student Research Workshop, Barcelona, Spain, July 2004. URL `http://aclweb.org/anthology//P/P04/P04-2007`.

Charles J. Fillmore. The Case for Case. In Emmon Bach and R. Harms, editors, Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York, 1968.

Charles J. Fillmore. Frame semantics. In Linguistics in the Morning Calm, pages 111–137. Hanshin Publishing Co., Hanshin, Seoul, 1982.

Thomas Finkenstaedt and Dieter Wolff. Ordered profusion : Studies in dictionaries and the English lexicon, volume 13 of Annales Universitatis Saraviensis. C. Winter, 1973.

John R. Firth. A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special volume of the Philological Society), 1952-59 :1–32, 1957. URL http://www.bibsonomy.org/bibtex/25e3d6c72cdd123a638f71886d78f3c1e/jil.

Darja Fišer, Aleš Tavčar, and Tomaž Erjavec. slowcrowd : a crowdsourcing tool for lexicographic tasks. In LREC'14, Reykjavik, Iceland, may 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1106_Paper.pdf.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. Outsourcing framenet to the crowd. In ACL 2013, Sofia, Bulgaria, August 2013. URL http://www.aclweb.org/anthology/P13-2130.

Nabil Gader, Sandrine Ollinger, and Alain Polguère. One Lexicon, Two Structures : So What Gives ? In GWC'2014, Tartu, Estonia, January 2014. URL http://www.aclweb.org/anthology/W14-0122.

William A. Gale, Kenneth W. Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. Computers and the Humanities, 26(5-6) :415–439, 1992. ISSN 0010-4817. doi : 10.1007/BF00136984. URL http://dx.doi.org/10.1007/BF00136984.

N. Garg and J. Henderson. Unsupervised Semantic Role Induction with Global Role Ordering. In ACL 2012, Jeju Island, Korea, July 2012.

Matthew Gerber and Joice Y. Chai. Beyond NomBank : A study of implicit arguments for nominal predicates. In ACL 2010, july 2010. URL http://www.aclweb.org/anthology/P10-1160.

Daniel Gildea and Dan Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28(3) :245–288, 2002.

Yoav Goldberg and Jon Orwant. A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books. In *SEM 2013, Atlanta, Georgia, USA, June 2013. URL http://www.aclweb.org/anthology/S13-1035.

Matthew R. Gormley, Margaret Mitchell, Benjamin Van Durme, and Mark Dredze. Low-Resource Semantic Role Labeling. In ACL 2014, Baltimore, Maryland, June 2014.

Gregory Grefenstette. Conquering language : Using NLP on a massive scale to build high dimensional language models from the web. In CICLing 2007, February 2007.

Maurice Gross. Méthodes en syntaxe. Régime des constructions complétives. Hermann, Paris, 1975.

Kilem L. Gwet. Handbook of inter-rater reliability. Advanced Analytics, LLC, September 2001. URL `http://www.agreestat.com/book_excerpts.html`.

Fadila Hadouche. Annotation syntaxico-sémantique des actants en corpus spécialisé. PhD thesis, Université de Montréal, 2011.

Valérie Hanoka and Benoît Sagot. Wordnet extension made simple : A multilingual lexicon-based approach using wiki resources. In LREC'12, may 2012. ISBN 978-2-9517408-7-7.

Zellig S. Harris. Distributional structure. Word, 10(2-3) :146–162, 1954.

Silvana Hartmann and Iryna Gurevych. Framenet on the way to babel : Creating a bilingual framenet using wiktionary as interlingual connection. In ACL 2013, Sofia, Bulgaria, August 2013. URL `http://www.aclweb.org/anthology/P13-1134`.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. The verbcorner project : Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. In ACL 2014, Baltimore, Maryland, June 2014. URL `http://aclweb.org/anthology/P14-2065`.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics-Volume 2, page 539–545. Association for Computational Linguistics, 1992.

Karin Friberg Heppin and Maria Toporowska Gronostaj. The rocky road towards a swedish framenet-creating swefn. In LREC'12, May 2012.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. Semantic frame identification with distributed word representations. In ACL 2014, Baltimore, Maryland, 2014.

Geoffrey E Hinton. Learning distributed representations of concepts. In Proceedings of the Eighth Annual Conference of the Cognitive Science Society, 1986.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes : the 90% solution. In HLT-NAACL 2006 : Short Papers, page 57–60, 2006.

Nancy Ide and Yorick Wilks. Making sense about sense. In Word sense disambiguation, chapter 3, page 47–73. Springer, 2006.

Sabine Schulte im Walde. Experiments on the automatic induction of German semantic verb

classes. Computational Linguistics, 32(2), 2006. URL http://aclweb.org/anthology/J/J06/J06-2001.

Christine Jacquin, Laura Monceaux, and Emmanuel Desmontils. Systèmes question-réponse et EuroWordNet. In TALN 2006, 2006.

Indrek Jentson. VerbNet Workbench. In GWC 2014, January 2014.

Richard Johansson and Pierre Nugues. Extended constituent-to-dependency conversion for English. In NoDaLiDa 2007. University of Tartu, 2007.

Mark Johnston and Frederica Busa. Qualia structure and the compositional interpretation of compounds. In ACL SIGLEX workshop on breadth and depth of semantic lexicons, 1996.

Daniel Jurafsky and James H. Martin. Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech, Second Edition. Pearson Education, 2008.

David Jurgens and Ioannis Klapaftis. SemEval-2013 Task 13 : Word Sense Induction for Graded and Non-Graded Senses. In *SEM 2013, june 2013. URL http://aclweb.org/anthology/S13-2049.

Daisuke Kawahara, Daniel W. Peterson, and Martha Palmer. A step-wise usage-based method for inducing polysemy-aware verb classes. In ACL 2014, Baltimore, Maryland, June 2014. URL http://www.aclweb.org/anthology/P14-1097.

Wiltrud Kessler and Jonas Kuhn. Detection of Product Comparisons - How Far Does an Out-of-the-Box Semantic Role Labeling System Take You ? In EMNLP 2013, Seattle, Washington, USA, October 2013. URL http://www.aclweb.org/anthology/D13-1194.

Adam Kilgarriff. I don't believe in word senses. Computers and the Humanities, 31(2) :91–113, 1997.

Adam Kilgarriff. A detailed, accurate, extensive, available English lexical database. In HLT-NAACL 2010 : Demonstration Session, page 21–24, 2010.

Karin Kipper-Schuler. VerbNet : A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania, 2005.

Anna Korhonen and Ted Briscoe. Extended lexical-semantic classification of English verbs. In Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, 2004.

Mikhail Kozhevnikov and Ivan Titov. Cross-lingual transfer of semantic role labeling models.

In ACL 2013, Sofia, Bulgaria, August 2013. URL http://www.aclweb.org/anthology/P13-1117.

Joel Lang and Mirella Lapata. Unsupervised Semantic Role Induction via Split-Merge Clustering. In ACL 2011, Portland, Oregon, USA, June 2011.

Mirella Lapata and Chris Brew. Verb Class Disambiguation Using Informative Priors. Computational Linguistics, 30(1), 2004. URL http://www.aclweb.org/anthology/J04-1003.

Hai-Son Le, Alexandre Allauzen, and François Yvon. Continuous Space Translation Models with Neural Networks. In NAACL-HLT 2012, June 2012. URL http://www.aclweb.org/anthology/N12-1005.

Els Lefever and Véronique Hoste. Semeval-2010 task 3 : Cross-lingual word sense disambiguation. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, July 2010. URL http://www.aclweb.org/anthology/S10-1003.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. From context to meaning : distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics, 1(20) :1–31, 2008.

Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *SEM 2012, June 2012. URL http://www.aclweb.org/anthology/S12-1012.

Beth Levin. English verb classes and alternations : a preliminary investigation. University Of Chicago Press, 1993.

Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In NIPS 2014, 2014.

J Lighthill et al. Artificial intelligence : a paper symposium. Science Research Council, London, 1973.

Dekang Lin. Automatic retrieval and clustering of similar words. In COLING 1998, August 1998.

Thomas Lippincott, Anna Korhonen, and Diarmuid Ó Séaghdha. Learning syntactic verb frames using graphical models. In ACL 2012, Jeju Island, Korea, July 2012. URL http://www.aclweb.org/anthology/P12-1044.

Ken Litkowski. The framenet frame element taxonomy. 2014. URL http://www.clres.com/online-papers/FETaxonomy.pdf.

Xavier Lluís, Xavier Carreras, and Lluís Màrquez. A Shortest-path Method for Arc-factored Semantic Role Labeling. In EMNLP 2014, Doha, Qatar, October 2014. URL http://www.aclweb.org/anthology/D14-1049.

Chi-kiu Lo and Dekai Wu. MEANT : An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In ACL 2011, Portland, Oregon, USA, June 2011. URL http://www.aclweb.org/anthology/P11-1023.

William Léchelle and Philippe Langlais. Utilisation de représentations de mots pour l'étiquetage de rôles sémantiques suivant FrameNet. In TALN 2014, Marseille, France, July 2014. URL http://www.atala.org/taln_archives/TALN/TALN-2014/taln-2014-long-004.

Marie-Claude L'Homme. Adding syntactico-semantic information to specialized dictionaries : an application of the FrameNet methodology. Lexicographica, 28(1) :233–252, December 2012.

Xiaojuan Ma and Christiane Fellbaum. Rethinking WordNet's Domains. In GWC 2012, January 2012.

Jon Malmaud, Earl J. Wagner, Nancy Chang, and Kevin Murphy. Cooking with Semantics. In ACL 2014 Workshop on Semantic Parsing, Baltimore, Maryland, 2014.

Christopher D Manning. Part-of-speech tagging from 97% to 100% : is it time for some linguistics ? In Computational Linguistics and Intelligent Text Processing, pages 171–189. Springer, 2011.

Morgane Marchand, Romaric Besançon, Olivier Mesnard, and Anne Vilnat. Influence des marqueurs multi-polaires dépendant du domaine pour la fouille d'opinion au niveau du texte (study of domain dependant multi-polarity words for document level opinion mining). In TALN 2014, Marseille, France, July 2014. URL http://www.aclweb.org/anthology/F14-1001.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English : The Penn Treebank. Computational linguistics, 19(2) :313–330, 1993.

Jiří Materna. Parameter estimation for lda-frames. In HLT-NAACL 2013, Atlanta, Georgia, June 2013. URL http://www.aclweb.org/anthology/N13-1051.

Jiří Materna. LDA-Frames : An Unsupervised Approach to Generating Semantic Frames. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 7181 of Lecture Notes in Computer Science, page 376–387. Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-28603-2. URL http://dx.doi.org/10.1007/978-3-642-28604-9_31. 10.1007/978-3-642-28604-9_31.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. Approaching plwordnet 2.0. In Proceedings of the 6th Global Wordnet Conference, Matsue, Japan, 2012.

Diana McCarthy. DANTE : a new resource for research at the syntax-semantics interface. In Interdiciplinary Workshop on Verbs, Pisa, Italy, 2010.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. Multilingual dependency analysis with a two-stage discriminative parser. In CONLL 2006, 2006.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. A Multilingual Paradigm for Automatic Verb Classification. In ACL 2002, Philadelphia, Pennsylvania, USA, July 2002. URL `http://aclweb.org/anthology/P02-1027`.

Cédric Messiant, Kata Gábor, and Thierry Poibeau. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. Traitement automatique des langues, 51(1) :65–96, 2010.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank project : An interim report. In HLT-NAACL 2004 Workshop : Frontiers in Corpus Annotation, 2004.

Olivier Michalon. Modélisation probabiliste de l'interface syntaxe sémantique à l'aide de grammaires hors contexte probabilistes expériences avec framenet. In TALN 2014 (RECITAL), Marseille, France, July 2014. URL `http://www.aclweb.org/anthology/F14-4001`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In NIPS 2013, 2013.

George A. Miller. Empirical methods in the study of semantics. In Journeys in Science : Small Steps – Great Strides, pages 51–73. University of New Mexico Press, Albuquerque, 1967.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet : An on-line lexical database*. International journal of lexicography, 3(4) :235–244, 1990.

Claire Mouton. Ressources et méthodes semi-supervisées pour l'analyse sémantique de texte en français. PhD thesis, Université Paris-Sud, 2010.

Claire Mouton and Gaël de Chalendar. JAWS : Just Another WordNet Subset. In TALN 2010, June 2010.

Claire Mouton, Gaël de Chalendar, and Benoît Richert. Framenet translation using bilingual dictionaries with evaluation on the english-french pair. In LREC'10, Valletta, Malta, may 2010. URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/485_Paper.pdf`.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. Semantic role labeling : an introduction to the special issue. Computational linguistics, 34(2) :145–159, 2008.

Roberto Navigli. Word sense disambiguation : A survey. ACM Computing Surveys, 2009.

Roberto Navigli. A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. SOFSEM 2012 : Theory and Practice of Computer Science, page 115–129, 2012.

Roberto Navigli. Babelnet and friends : A manifesto for multilingual semantic processing. Intelligenza Artificiale, 7(2) :165–181, 2013.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet : Building a very large multilingual semantic network. In ACL 2010, July 2010.

Roberto Navigli and Daniele Vannella. Semeval-2013 task 11 : Word sense induction and disambiguation within an end-user application. In SemEval-2013, Atlanta, Georgia, USA, June 2013. URL `http://www.aclweb.org/anthology/S13-2035`.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. Semeval-2007 task 07 : Coarse-grained english all-words task. In SemEval-2007, 2007.

Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 task 12 : Multilingual word sense disambiguation. In SemEval-2013, Atlanta, Georgia, USA, June 2013. URL `http://www.aclweb.org/anthology/S13-2040`.

Jakob Nielsen. Response times : the three important limits. Usability Engineering, 1994. URL `http://www.nngroup.com/articles/response-times-3-important-limits/`.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. Semeval 2014 task 8 : Broad-coverage semantic dependency parsing. In SemEval 2014, pages 63–72, Dublin, Ireland, August 2014. URL `http://www.aclweb.org/anthology/S14-2008`.

Kyoko Hirose Ohara, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. The japanese framenet project : An introduction. In LREC'04, May 2004.

Antoni Oliver. Wn-toolkit : Automatic generation of wordnets following the expand model. In GWC 2014, Tartu, Estonia, January 2014. URL `http://www.aclweb.org/anthology/W14-0102`.

OLST. Corpus spécialisés de l'Observatoire de linguistique Sens-Texte (OLST), 2014.

Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Alba-

raa Abuobieda. An improved plagiarism detection scheme based on semantic role labeling. Applied Soft Computing, 12(5), 2012.

Martha Palmer. Semlink : Linking PropBank, VerbNet and FrameNet. In Proceedings of the Generative Lexicon Conference, page 9–15, 2009.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. Different sense granularities for different applications. In Workshop on Scalable Natural Language Understanding, 2004.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank : An annotated corpus of semantic roles. Computational Linguistics, 31(1) :71–106, 2005.

Martha Palmer, Daniel Gildea, and Nianwen Xue. Semantic role labeling. Synthesis Lectures on Human Language Technologies, 3(1) :1–103, 2010.

Martha Palmer, Ivan Titov, and Shumin Wu. Semantic Role Labeling. In NAACL HLT 2013 Tutorial Abstracts, page 10–12, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-4004`.

Patrick Pantel and Dekang Lin. Discovering word senses from text. In KDD 2002, 2002.

Ted Pedersen. Duluth-WSI : SenseClusters Applied to the Sense Induction Task of SemEval-2. In SemEval-2010, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S10-1081`.

Tommaso Petrolito and Francis Bond. A Survey of WordNet Annotated Corpora. In GWC 2014, January 2014. URL `http://www.aclweb.org/anthology/W14-0132`.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet : developing an aligned multilingual database. In GWC 2002, 2002.

John R Pierce and John B Carroll. Language and machines : Computers in translation and linguistics. 1966.

Jean-Marie Pierrel. Un ensemble de ressources de référence pour l'étude du français : TLFi, Frantext et le logiciel Stella. Revue québécoise de linguistique, 32(1) :155–176, 2003.

Alain Polguère. Principles of lexical network systemic modeling (principes de modélisation systémique des réseaux lexicaux) [in french]. In TALN 2014, Marseille, France, July 2014. URL `http://www.aclweb.org/anthology/F14-1008`.

Alain Polguère. Tissage du Réseau Lexical du Français (RLF) : buts et méthodes. In 27e Congrès International de Linguistique et de Philologie Romanes (CILPR 2013), Nancy, France, July 2013. URL `http://hal.archives-ouvertes.fr/hal-00905207`.

David M W Powers. The Problem with Kappa. In EACL 2012, April 2012.

Quentin Pradet, Jeanne Baguenier-Desormeaux, Gaël de Chalendar, and Laurence Danlos. WoNeF : amélioration, extension et évaluation d'une traduction française automatique de WordNet. In TALN 2013, 2013a.

Quentin Pradet, Gaël de Chalendar, and Guilhem Pujol. Revisiting knowledge-based Semantic Role Labeling. In LTC'13, December 2013b.

Quentin Pradet, Laurence Danlos, and Gaël de Chalendar. Adapting VerbNet to French using existing resources. In LREC'14, May 2014a.

Quentin Pradet, Gaël de Chalendar, and Jeanne Desormeaux Baguenier. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. In Proceedings of the Seventh Global Wordnet Conference (GWC2014), page 32–39, January 2014b.

James Pustejovsky. The generative lexicon. Computational linguistics, 17(4) :409–441, 1991.

James Pustejovsky, Adam Meyers, Martha Palmer, and Massimo Poesio. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In Workshop on Frontiers in Corpus Annotations II : Pie in the Sky, 2005.

Owen Rambow. The simple truth about dependency and phrase structure representations : An opinion piece. In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 337–340, Los Angeles, CA, USA, June 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N10-1049`.

François Rastier. Sémantique interprétative. Presses universitaires de France, 1987.

Corentin Ribeyre. Vers un système générique de réécriture de graphes pour l'enrichissement de structures syntaxiques. In RECITAL 2013, June 2013. URL `http://www.taln2013.org/actes/www/RECITAL-2013/actes/recital-2013-long-014.pdf`.

Kyle Richardson and Jonas Kuhn. Towards Semantic Parsing in Dynamic Domains. In SemDial 2012 (SeineDial), 2012.

Nick Riemer. La conception syntaxique de la polysémie : une critique. CogniTextes. Revue de l'Association française de linguistique cognitive, 6, 2011. URL `http://cognitextes.revues.org/404`.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan

Scheffczyk. FrameNet II : Extended Theory and Practice. International Computer Science Institute, Berkeley, California, 2006. Distributed with the FrameNet data.

Stuart Russell and Peter Norvig. Inlligence artificielle : A Modern Approach. Pearson Education France, 3 edition, 2010.

Benoit Sagot and Darja Fišer. Construction d'un wordnet libre du français à partir de ressources multilingues. In TALN 2008, 2008a.

Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In Ontolex 2008, May 2008b.

Benoît Sagot and Darja Fišer. Automatic Extension of WOLF. In GWC 2012, January 2012a.

Benoît Sagot and Darja Fišer. Cleaning noisy wordnets. In LREC'12, Istanbul, Turkey, may 2012b. URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/1127_Paper.pdf`.

Magnus Sahlgren. The distributional hypothesis. Italian Journal of Linguistics, 2008.

Patrick Saint-Dizier. Constructing Verb Semantic Classes for French : Methods and Evaluation. In COLING 1996, aug 1996. URL `http://aclweb.org/anthology/C96-2204`.

Carolina Scarton and Sandra Aluısio. Towards a cross-linguistic VerbNet-style lexicon for Brazilian Portuguese. In CREDISLAS (LREC'12), 2012.

Carolina Scarton, Magali Sanches Duran, and SandraMaria Aluísio. Using Cross-Linguistic Knowledge to Build VerbNet-Style Lexicons : Results for a (Brazilian) Portuguese VerbNet. In PROPOR 2014 - Computational Processing of the Portuguese Language, 2014. URL `http://dx.doi.org/10.1007/978-3-319-09761-9_15`.

Thomas Schmidt. Interfacing Lexical and Ontological Information In a Multilingual Soccer FrameNet. In OntoLex 2006, Genoa, Italy, May 2006. URL `http://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/2263/file/schmidt_interfacing_lexical_2006.pdf`.

Thomas Schmidt. The Kicktionary – A Multilingual Lexical Resource of Football Language. In Hans C. Boas, editor, Multilingual FrameNets in Computational Lexicography, page 101–134. de Gruyter, New York, 2009.

Dan Shen and Mirella Lapata. Using Semantic Roles to Improve Question Answering. In EMNLP-CoNLL 2007, June 2007. URL `http://www.aclweb.org/anthology/D07-1002`.

L. Shi and R. Mihalcea. Putting pieces together : Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In CICLing 2005, page 100–111. Springer, 2005.

James R Slagle. A heuristic program that solves symbolic integration problems in freshman calculus. Journal of the ACM (JACM), 10(4) :507–520, 1963.

Kelly H. Smith. The Rumble In The Disambiguation Jungle, 2011. Bachelor's thesis.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. Learning to Merge Word Senses. In EMNLP-CoNLL 2007, June 2007. URL `http://www.aclweb.org/anthology/D07-1107`.

Benjamin Snyder and Martha Palmer. The English All-Words Task. In SENSEVAL-3, Barcelona, July 2004. URL `http://www.aclweb.org/anthology/W04-0811`.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. Semeval-2012 task 1 : English lexical simplification. In SemEval-2012, Montréal, Canada, 7-8 June 2012. URL `http://www.aclweb.org/anthology/S12-1046`.

Suzanne Stevenson and Eric Joanis. Semi-supervised verb class discovery using noisy features. In CoNLL 2003, Edmonton, Canada, May 2003. URL `http://www.aclweb.org/anthology/W03-0410`.

Carlos Subirats and Miriam Petruck. Surprise : Spanish framenet. In Computer in Libraries (CIL), March 2003.

Lin Sun, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. Investigating the cross-linguistic potential of VerbNet : style classification. In COLING 2010, 2010.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In CoNLL 2008, 2008. URL `http://www.aclweb.org/anthology/W08-2121`.

Yoshimi Suzuki and Fumiyo Fukumoto. Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering. In TextGraphs-4, page 32–40, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `http://aclweb.org/anthology//W/W09/W09-3205`.

Robert Swier and Suzanne Stevenson. Unsupervised semantic role labelling. In EMNLP 2004, page 95–102, Barcelona, Spain, July 2004.

Robert Swier and Suzanne Stevenson. Exploiting a Verb Lexicon in Automatic Semantic Role Labelling. In HLT-EMNLP 2005, Vancouver, Canada, October 2005. URL `http://www.aclweb.org/anthology/H05-1111`.

Juliette Thuilier. Contraintes préférentielles et ordre des mots en français. PhD thesis, Université Paris-Diderot, 2012.

Ivan Titov and Alexandre Klementiev. A Bayesian Approach to Unsupervised Semantic Role Induction. In EACL 2012, page 12–22, Avignon, France, April 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E12-1003.

Tiago Torrent, Maria Margarida Salomão, Fernanda Campos, Regina Braga, Ely Matos, Maucha Gamonal, Julia Gonçalves, Bruno Souza, Daniela Gomes, and Simone Peron. Copa 2014 framenet brasil : a frame-based trilingual electronic dictionary for the football world cup. In COLING 2014 : System Demonstrations, Dublin, Ireland, August 2014. URL http://www.aclweb.org/anthology/C14-2003.

Dan Tufiş, Dan Cristea, and Sofia Stamou. BalkaNet : Aims, methods, results and perspectives. a general overview. Romanian Journal of Information Science and Technology, 7(1-2) :9–43, 2004.

David Vadas and James R Curran. Adding noun phrase structure to the Penn Treebank. In ACL 2007, 2007.

Lonneke van der Plas and Marianna Apidianaki. Cross-lingual word sense disambiguation for predicate labelling of french. In TALN 2014, Marseille, France, July 2014. URL http://www.aclweb.org/anthology/F14-1005.

Maarten van Gompel, Iris Hendrickx, Antal van den Bosch, Els Lefever, and Veronique Hoste. Semeval 2014 task 5 - l2 writing assistant. In SemEval-2014, pages 36–44, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/S14-2005.

Giulia Venturi, Alessandro Lenci, Simonetta Montemagni, Eva Maria Vecchi, Maria Teresa Sagri, Daniela Tiscornia, and Tommaso Agnoloni. Towards a framenet resource for the legal domain. LOAIT, pages 67–76, 2009.

Marc Verhagen, Amber Stubbs, and James Pustejovsky. Combining Independent Syntactic and Semantic Annotation Schemes. In Linguistic Annotation Workshop, june 2007. URL http://www.aclweb.org/anthology/W07-1517.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In Proceedings of the workshop on geometrical models of natural language semantics, page 74–82. Association for Computational Linguistics, 2009. URL http://eprints.pascal-network.org/archive/00006249/01/W09-0210.pdf.

Piek Vossen. EuroWordNet : a multilingual database with lexical semantic networks. Kluwer Academic, October 1998.

Wikipédia. Lexicographie — Wikipédia, l'encyclopédie libre, 2014. URL `http://fr.wikipedia.org/w/index.php?title=Lexicographie&oldid=102891440`. [En ligne ; Page disponible le 24-juin-2014].

Terry Winograd. Understanding natural language. Cognitive psychology, 3(1) :1–191, 1972.

Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. Semantic Frames to Predict Stock Price Movement. In ACL 2013, Sofia, Bulgaria, August 2013. URL `http://aclweb.org/anthology/P13-1086`.

Haitong Yang and Chengqing Zong. Multi-Predicate Semantic Role Labeling. In EMNLP 2014, page 363–373, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1041`.

David Yarowsky. One sense per collocation. In Proceedings of the workshop on Human Language Technology, page 266–271, 1993.

Daniel Zeman and Philip Resnik. Cross-language parser adaptation between related languages. In IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, January 2008. URL `http://aclweb.org/anthology/I08-3008`.

Liste des publications

Quentin Pradet, Jeanne Baguenier-Desormeaux, Gaël de Chalendar et Laurence Danlos. Juin 2013. WoNeF : amélioration, extension et évaluation d'une traduction française automatique de WordNet. TALN 2013, Les Sables d'Olonne, France.

Quentin Pradet, Gaël de Chalendar and Guilhem Pujol. December 2013. Revisiting knowledge-based Semantic Role Labeling. LTC'13, Poznań, Poland.

Quentin Pradet, Gaël de Chalendar and Jeanne Baguenier Desormeaux. January 2014. WoNeF, an improved, expanded and evaluated automatic French translation of WordNet. GWC 2014, Tartu, Estonia.

Quentin Pradet, Laurence Danlos and Gaël de Chalendar. May 2014. Adapting VerbNet to French using existing resources. LREC 2014, Reykjavik, Iceland.

Laurence Danlos, Takuya Nakamura and Quentin Pradet. July 2014. Vers la création d'un Verb∋Net du français. Atelier FondamenTAL, TALN 2014, Marseille, France.

Quatrième partie

Annexes

Reproduction des systèmes utilisés

Ce chapitre traite de divers détails destinés à aider le lecteur à comprendre, voire reproduire notre traduction de WordNet.

1. Relations syntaxiques identifiées par LIMA présentes dans notre modèle de langue syntaxique

Cette table décrit les identifiants des relations syntaxiques utilisés dans LIMA et donc présents dans le modèle de langue syntaxique que nous utilisons pour traduire WordNet. Ces relations ont été définies à partir des relations définies lors des campagnes d'évaluation Easy et Passage. La définition initiale d'un certain nombre d'entre elles se trouve dans le guide d'annotation de la campagne Easy (`http://perso.limsi.fr/Individu/anne/Guide/PEAS_reference_annotations_v2.2.html`). Il faut noter que dans ces campagnes les relations sont indiquées entre constituants, alors que LIMA fournit des relations en dépendances (les relations sont alors entre les têtes des syntagmes correspondants).

Identifiant	Description	Exemples fréquents
ADJPRENSUB	adjectif épithète pré-nominal	premier fois, présent loi, bon état
ADVADJ	adverbe modifiant un adjectif	tout autre, plus grand
ADVADV	adverbe modifiant un adverbe	tout simplement, très bien
AdvSub	adverbe modifiant un substantif	que j, beaucoup plus, non pas
AdvVerbe	adverbe modifiant un verbe	haut voir, ici envoyer
APPOS	apposition (cf. APP dans PEAS)	aristocrate royaliste, abri piscine
ATB_O	attribut de l'objet (cf. ATB_SO dans PEAS)	site site, álbum annuaire
ATB_S	attribut du sujet en relation avec le verbe (cf. ATB_SO dans PEAS)	disponible être, possible être
ATB_SG	attribut du sujet en relation avec le sujet	nouveau message

COD_V	complément d'objet direct du verbe	être pouvoir, profil voir
COMPADJ	complément de l'adjectif	site plan, page haut
COMPADV	complément de l'adverbe	page haut, forum uniquement
COMPDUNOM	complément du nom	page numéro, page pas, page haut
COMPL	complémenteur (cf. COMP dans PEAS)	que être, que avoir, si être
CPL_V	complément indirect ou circonstanciel du verbe	être être, être avoir
CPLV_V	groupe prépositionnel infinitif après le verbe	connecter cliquer, voir revenir
DetIntSub	déterminant interrogatif	quelle période, quelle manière
DetSubNum	déterminant numéral cardinal	neuf appartement, trente an
MOD_A	modificateur de l'adjectif	être vrai, être même, être autre
MOD_N	modificateur du nom	commander libraire, être personne
MOD_V	modificateur du verbe	que être, être aller, matière tabler
NePas	négation	ne pas, ne rien, ne jamais
Prefixe	préfixe	il- pas, il- être, anti-criminalité
PrepDetInt	relation entre préposition et déterminant interrogatif	de quelle, dans quelle, pour quelle
PrepPronCliv	relation entre préposition et conjonction de subordination considérée comme un pronom clivé	de que, em que, une que
PrepPron	relation entre préposition et pronom personnel	de tézigue, pour mézigue
PrepPronRelCa	relation entre préposition et pronom relatif complément d'attribution	dans lequel, pour qui, sur lequel
PrepPronRel	relation entre préposition et pronom relatif	de quoi, en quoi, comme quoi
SUBADJPOST	adjectif épithète post-nominal	posté profil, référencé page
SUBSUBJUX	substantif juxtaposé à un substantif	top thé, article présent, web site
SUJ_V	sujet du verbe	j avoir, thé pager, j être
SUJ_V_REL	pronom sujet du verbe de la proposition relative	qui être, que avoir

SUJ_V_RELG	antécédent sujet du verbe de la proposition relative	définition suivre, personne avoir

TABLE 1. – Relations syntaxiques présentes dans le modèle de langue syntaxique utilisé pour la traduction de WordNet vers WoNeF. La troisième colonne présente plusieurs exemples de relations entres lemmes fréquentes dans les corpus. Le modèle de langue est disponible sur `http://www.kalisteo.fr/demo/semanticmap/`.

2. Sélecteurs employés pour produire WoNeF

2.1. Combinaisons de sélecteurs initiaux

Noms	haute précision	monosémie, unicité, Levenshtein
Noms	haut F-score	monosémie, unicité, sources multiples, Levenshtein
Noms	haute couverture	monosémie, unicité, sources multiples, Levenshtein
Verbes	haute précision	unicité
Verbes	haut F-score	unicité, monosémie
Verbes	haute couverture	monosémie, unicité, sources multiples
Adjectifs	haute précision	monosémie, unicité, Levenshtein
Adjectifs	haut F-score	monosémie, unicité, sources multiples, Levenshtein
Adjectifs	haute couverture	monosémie, unicité, sources multiples, Levenshtein
Adverbes	haute précision	monosémie, unicité, sources multiples, Levenshtein
Adverbes	haut F-score	monosémie, unicité, sources multiples, Levenshtein
Adverbes	haute couverture	monosémie, unicité, sources multiples, Levenshtein

TABLE 2. – Combinaison de sélecteurs initiaux utilisée pour chaque couple (partie du discours, version de WoNeF). Ces sélecteurs sont décrits aux sections 2.2.1 et 3.1.2.

2.2. Combinaisons de sélecteurs syntaxiques

[Mouton, 2010, section 3.1.1.3] décrit les sélecteurs syntaxiques utilisés dans WoNeF. Les relations utilisées sont décrites à la section 1.

— Le seul sélecteur syntaxique utilisé pour les verbes, adjectifs et adverbes est le sélecteur par synonymie avec diverses relations syntaxiques pouvant refléter la synonymie. Par exemple, avec la relation COD_V inverse, on exprime le fait que les verbes qui acceptent les mêmes objets sont potentiellement synonymes.
— Les verbes utilisent les relations COD_V inverse, CPL_V inverse et CPLV_V inverse.

  - Les adjectifs utilisent les relations SUBADJPOST ADVADJ inverse, et et ATBSG.
  - Les adverbes utilisent les relations AdvSub inverse et ADVADJ inverse.
- Concernant les noms, la configuration change suivant la version de WoNeF.
  - La version haute précision utilise les sélecteurs de méronymie avec la relation COMPDUNOM (Figure 2.2) et d'holonymie avec la relation COMPDUNOM inverse.
  - La version à haut F-score rajoute le sélecteur par hyperonymie avec la relation syntaxique COD.
  - La version à haute couverture rajoute elle les sélecteurs par synonymie (COMPDUNOM, COD_V et APPOS), hyperonymie (COMPDUNOM et SUJ_V) et hyponymie (COMPDUNOM).