Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes

Abstract

This paper presents a work on the extension of a semantic lexical resource for French adverbs, using both the derivation and the synonyms relations. The latter relies on the DicoSyn synonyms database. We also describe the construction of links between this semantic resource and the adverbial entries of the syntactic lexicon Le*fff*, which were mostly extracted from the lexicon-grammar tables from (Molinier & Levrier 2000). The extended semantic resource is carefuly evaluated. Both the WOLF and the Le*fff* are freely available.

Keywords : Adverbs, WordNet, syntactic lexicon.

1. Introduction

La question de la disponibilité des ressources pour le Traitement Automatique du Langage (TAL) reste, encore aujourd'hui, une question cruciale, notamment pour le français. Si les disparités avec l'anglais s'estompent un peu en ce qui concerne les ressources morphologiques et syntaxiques (Sagot *et al.* 2006), l'accès à des ressources sémantiques de qualité reste difficile. Des efforts sont néammoins faits pour mettre à la disposition de la communauté un WordNet du français, librement disponible, le WOLF (cf. section 2.2.). Enfin, le couplage entre ressources syntaxiques et sémantiques reste à faire.

Dans cet article, nous présentons un premier pas dans cette direction. En restreignant notre champ d'investigation aux adverbes, nous avons cherché à compléter le WOLF puis à coupler ceux-ci aux entrées adverbiales du lexique syntaxique Le*fff* (cf. section 2.1.), elles-mêmes construites en grande partie à partir des tables du lexique-grammaire des adverbes en *-ment* de (Molinier & Levrier 2000). Pour ce faire, nous nous sommes également appuyés sur la base de synonymes DicoSyn (Ploux & Victorri 1998).

Cet article est organisé comme suit. La section 2. décrit les trois ressources à la base de ce travail. La section 3. décrit la façon dont nous avons étendu le WOLF à l'aide de deux techniques complémentaires, ainsi qu'une évaluation manuelle des entrées obtenues. Enfin, la section 4. décrit la façon dont nous avons associé à certaines entrées adverbiales du Le*fff* un ou plusieurs identifiants de synsets du WOLF, en exploitant les classes d'adverbes de (Molinier & Levrier 2000).

2. Ressources

2.1. Le Le*fff* et les tables du lexique-grammaire

Le Le*fff* (Lexique des Formes Fléchies du Français) (Sagot *et al.* 2006), est un lexique morphologique et syntaxique du français à large couverture, librement disponible [1]. Le Le*fff* a pour objectif de concilier la pertinence linguistique et l'utilisabilité dans des applications de TAL. Il est notamment utilisé dans divers analyseurs syntaxiques reposant sur différents formalismes

[1] `http://gforge.inria.fr/projects/alexina/`

linguistiques (LFG, TAG). Le Le*fff*, aujourd'hui en version 3, couvre toutes les catégories et est progressivement enrichi d'informations syntaxiques et sémantiques, notamment par comparaison avec d'autres ressources syntaxiques (Sagot & Danlos 2007; Danlos & Sagot 2007). Ainsi, les entrées adverbiales du Le*fff* ont pu être complétées (Sagot & Fort 2007) grâce aux tables du *Lexique-Grammaire* des adverbes en *-ment*, publiées par Molinier (Molinier & Levrier 2000). En particulier, chaque entrée du Le*fff* correspondant à un adverbe en *-ment* est associée à une classe syntaxico-sémantique de Molinier. Naturellement, un même lemme adverbial peut correspondre à plusieurs entrées, chacune appartenant alors nécessairement à une classe différente.

Les adverbes en *-ment* forment une classe numériquement importante d'adverbes, qui, contrairement aux autres adverbes, est une classe ouverte. Ces adverbes forment une classe morphologiquement homogène, puisqu'ils sont majoritairement construits sur un schéma du type adjectif+*ment*. Toutefois, nombre d'autres adverbes existent, et notamment une quantité importante de locutions adverbiales. Un grand nombre d'entre elles ont été décrites dans les tables d'adverbes figés de Maurice Gross. Ces tables comportent également de nombreuses collocations à fonction adverbiale mais dont le statut d'adverbe composé fait débat. C'est pourquoi nous n'avons pas cherché à en tirer parti pour l'instant.

En revanche, le Le*fff* ne dispose pas encore d'informations sémantiques, pas même d'identifiants sémantiques à associer à chaque entrée. Un des objectifs de ce travail est de remédier à ce manque, en ce qui concerne les entrées adverbiales, en établissant des liens entre les entrées adverbiales du Le*fff* et celles du lexique sémantique (WordNet) WOLF.

2.2. WOLF

Le WOLF (WOrdnet Libre du Français) est une ressource lexicale sémantique pour le français, librement disponible [2] (Sagot & Fišer 2008). Comme tout WordNet, le WOLF est une base de données lexicales dans laquelle les mots (lexèmes) sont répartis en catégories et organisés en une hiérarchie de nœuds. Chaque nœud a un identifiant unique, et représente un *concept*, ou *synset* (ensemble de synonymes). Il regroupe un certain nombre de lexèmes synonymes dénotant ce concept. Ainsi, dans le Princeton WordNet (Fellbaum 1998), le premier WordNet a avoir été développé, et qui traite de l'anglais, le synset ENG20-02853224-n comprend les lexèmes {*car, auto, automobile, machine, motorcar*}. Les synsets sont précisés par une brève définition et sont liés à d'autres synsets (ainsi, ce synset est lié au synset {*motor vehicle, automotive vehicle*} par un lien d'hypéronymie, et au synset {*cab, hack, taxi, taxicab*} par un lien d'hyponymie). Les lexèmes peuvent être simples ou composés. Les usages métaphoriques et idiomatiques sont pris en compte.

C'est à partir du Princeton WordNet 2.0 et de diverses ressources multilingues qu'a été construit le WOLF, au moyen de deux approches complémentaires. Les lexèmes polysémiques ont été traités au moyen d'une approche reposant sur l'alignement en mots d'un corpus parallèle en cinq langues. Le lexique multilingue extrait a été désambiguïsé sémantiquement à l'aide des WordNets des langues concernées. Par ailleurs, une approche bilingue a été suffisante pour construire de nouvelles entrées à l'aide des mots monosémiques. Nous avons pour cela extrait des lexiques bilingues à partir de ressources wiki (Wikipedia, Wiktionary) et de thésaurus. Le wordnet obtenu a été évalué par rapport au wordnet français issu du projet EuroWordNet.

Le WOLF contient tous les synsets du Princetown WordNet, y compris ceux pour lesquels aucun lexème français n'est connu. En ce qui concerne les adverbes, la dernière version du WOLF (avant ce travail), la version 0.1.4, dispose de lexèmes français pour seulement 676 des 3 664 synsets adverbiaux [3]. Ceci ne représente que 18,4% des synsets adverbiaux, et seulement 983 liens lexème-synset ne mettant en jeu que 665 lemmes adverbiaux distincts. C'est la raison

[2] http://wolf.gforge.inria.fr/

[3] On notera que le seul autre WordNet pour le français, le WordNet développé dans le cadre du projet EuroWordNet (Vossen, P. 1999), ne comporte que des synsets nominaux et verbaux, mais aucun synset adjectival ni adverbial. De plus, d'importants problèmes de licence en font une ressource très peu utilisée par la communauté. Enfin, et en partie pour cette même raison, elle n'a pas évolué depuis sa création. Ce sont là les trois principales motivations du projet de développement du WOLF.

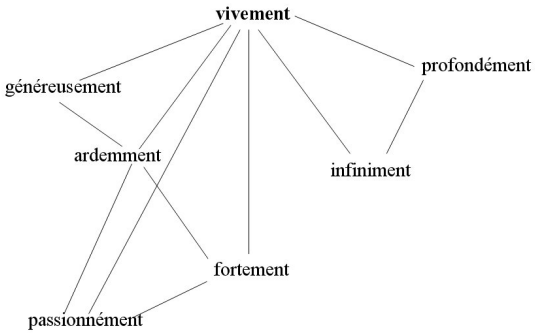

Figure 1. Un extrait du graphe adverbial de synonymie.

pour laquelle, avant de coupler les entrées adverbiales du L*efff* et du WOLF, nous avons décidé de mettre en œuvre deux techniques complémentaires pour en augmenter la couverture. L'une de ces techniques repose sur la relation de dérivation morphologique *et* sémantique, mentionnée précédemment, qui existe souvent entre un synset adverbial et un synset adjectival, tant en anglais qu'en français. L'autre de ces techniques repose sur l'exploitation de la base de synonymes DicoSyn.

2.3. DicoSyn et les cliques de synonymes

DicoSyn est un dictionnaire électronique de synonymes, dont les versions les plus récentes sont consultables en ligne [4]. La base initiale (Ploux & Victorri 1998) est issue de la fusion de sept dictionnaires classiques (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert) dont ont été extraites les relations synonymiques. Ces relations ont été symétrisées (si u1 est synonyme de u2 dans un dictionnaire, alors on considère que u2 est aussi synonyme de u1) Le dictionnaire obtenu contient environ 49 000 entrées et 190 000 relations synonymiques. L'intérêt principal de ce dictionnaire est de donner de manière explicite le graphe de la relation de synonymie [5]. Ploux et Victorri ont mis au point un logiciel, Visusyn, permettant d'explorer le sous-graphe constitué par une unité lexicale et tous ses synonymes. On peut alors visualiser et caractériser, de façon automatique, les propriétés sémantiques de cette unité (François *et al.* 2002) (Venant 2004). Une extension récente de Visusyn (Venant 2007) permet désormais d'explorer des graphes beaucoup plus globaux et d'étudier les caractéristiques sémantiques d'un paradigme lexical dans son ensemble. Nous avons ainsi pu exploiter un graphe de synonymes adverbiaux. DicoSyn ne comportant aucune indication de catégorie, ce graphe adverbial a été construit en croisant DicoSyn avec les entrées adverbiales en -ment du L*efff*. Ce graphe comporte 1597 sommets (les adverbes) et 4344 liens (de synonymie). Parmi les 1597 sommets, certains ne sont pas des adverbes en *-ment*, mais des synonymes de ces adverbes (par exemple *bien* est un sommet du graphe car DicoSyn le signale comme étant synonyme de *amplement* ou *copieusement*). La figure 1 présente un extrait de ce graphe.

L'exploitation de ce graphe repose sur la notion de clique. Une clique est un ensemble, le plus grand possible, de sommets deux à deux connectés. Ainsi le graphe de la figure 1 comporte 3 cliques : la clique *ardemment; fortement; passionnément; vivement* (on ne peut pas ajouter *généreusement* qui n'est pas synonyme de *fortement*, ni de *passionnément*), la clique *ardemment; généreusement; vivement* et la clique *infiniment; profondément; vivement*. Le graphe adverbial que nous avons construit comporte 2247 cliques. L'idée est qu'en première approximation, une clique correspond à un emploi adverbial possible. Ce sont donc les cliques qui vont constituer l'unité structurelle d'analyse sémantique du graphe.

[4] http://dico.isc.cnrs.fr/ et http://elsap1.unicaen.fr/dicosyn.html

[5] Il s'agit bien sûr d'une relation de synonymie partielle, que Ploux et Victorri définissent de la façon suivante : « Deux unités lexicales sont en relation de synonymie si toute occurrence de l'une peut être remplacée par une occurrence de l'autre dans un certain nombre d'environnements sans modifier notablement le sens de l'énoncé dans lequel elle se trouve. »

3. Extension du WOLF

Comme indiqué précédemment, la première étape de notre travail a été d'étendre le WOLF, afin d'augmenter le nombre de synsets adverbiaux non vides (pour lesquels au moins un lexème français est connu), et d'augmenter le nombre de lexèmes dans chaque synset non vide. Pour cela, nous nous sommes appuyés sur deux types de relations entre lexèmes : la relation de dérivation entre un adverbe en *-ment* et l'adjectif sur lequel il est construit, et la relation de synonymie entre adverbes telle que définie par les cliques extraites de DicoSyn.

3.1. Extension par la relation de dérivation

La technique qui repose sur la relation de dérivation part du double constat suivant :

- Le Princeton WordNet inclut une relation de dérivation (*derived*) qui relie certains synsets adverbiaux à un ou plusieurs synsets adjectivaux. Ce lien indique que certains lexèmes adjectivaux présents dans le synset adjectival permettent de construire, par dérivation morphologique (suffixe *-ly*), certains lexèmes adverbiaux présents dans le synset adverbial. Naturellement, ce lien indique également une parenté sémantique entre les deux synsets.

- Le mécanisme de dérivation morphologique et sémantique entre adjectifs et adverbes est souvent parallèle entre l'anglais (adjonction du suffixe *-ly* à l'adjectif) et le français (adjonction du suffixe *-ment* au féminin singulier de l'adjectif [6]).

Nous avons donc récupéré, pour chaque synset adverbial, les adjectifs (français) associés par le WOLF au synset adjectival qui lui est relié par la relation *derived*. Nous avons appliqué l'algorithme de dérivation morphologique à ces adjectifs [7]. Ceux des adverbes ainsi construits qui sont présents dans le Le*fff* ont été conservés, et attribués au synset adverbial de départ (en conservant l'information selon laquelle c'est par dérivation morphologique que ces liens lexème-synset ont été construits).

Par exemple, soit le synset ENG20-00115661-b. Dans le WOLF 0.1.4, seuls les lexèmes *toujours* et *invariablement* y sont présents (et corrects). Or, ce synset est lié par la relation *derived* au synset adjectival ENG20-02417249-a, qui, dans le WOLF, comporte les lexèmes *permanent*, *invariable* et *perpétuel*. Les adverbiaux hypothétiques *permanentement*, *invariablement* et *perpétuellement* sont donc construits. Le premier est éliminé car il n'est pas présent dans le Le*fff*, le second confirme un lexème déjà présent dans le synset adverbial, et le dernier permet la création d'un nouveau lien lexème-synset. Au final, le synset ENG20-00115661-b devient donc {*toujours*, *invariablement*, *perpétuellement*}.

Grâce à l'application de cette technique, le nombre de relations lexème-synset adverbial dans le WOLF est passé de 983 à 1536 (56% d'augmentation). Le nombre de synsets adverbiaux non vides est quant à lui passé de 676 à 969 (43% d'augmentation). Le nombre de lexèmes adverbiaux présents dans le WOLF passe de 665 à 889 (23% d'augmentation).

3.2. Extension par la relation de synonymie

Une fois les synsets adverbiaux du WOLF complétés par l'exploitation de la relation de dérivation entre adverbes en *-ment* et adjectifs, nous avons mis en œuvre une technique qui repose sur la relation de synonymie telle que définie par les cliques extraites de DicoSyn. Pour ce faire, nous avons procédé en trois temps.

1. Nous avons tout d'abord attribué heuristiquement un poids à chaque lien lexème-synset, de la façon suivante. Si un lien a été extrait (entre autres) de ressources wiki, il reçoit un poids de 5. Si un lien a été extrait à partir de corpus multilingues alignés, il reçoit un poids

[6] Ceci n'est naturellement pas toujours exact (cf. courante/couramment et bien d'autres), mais constitue une heuristique raisonnable.

[7] Le féminin singulier de l'adjectif a été récupéré dans le Le*fff*.

de 4 si l'un des corpus comportait au moins 4 langues, et de 3 si tous ne comportaient que 3 langues. Dans les autres cas, y compris pour les liens construits à l'aide de la relation de dérivation, un poids de 2 est attribué au lien.

2. Pour chaque synset adverbial, on lui associe la clique extraite de DicoSyn qui maximise la somme des poids des lexèmes communs entre la clique et le synset.

3. Pour chaque synset, les lexèmes (adverbes) présents dans la clique qui lui est associée, mais absents du synset, lui sont rajoutés.

Par exemple, soit le synset ENG20-00115661-b, le même que précédemment. Après extension par la relation de dérivation, il contenait les adverbes *toujours*, *invariablement* et *perpétuellement*. Les deux premiers ayant été construits grâce au Wiktionary français, ils reçoivent un poids de 5. L'adverbe *perpétuellement*, construit par dérivation, reçoit le poids de 2. La clique qui maximise la somme des poids des lexèmes communs est la clique {*éternellement*, *invariablement*, *perpétuellement*, *sans cesse*, *toujours*}. Ce sont donc deux adverbes qui sont ajoutés au synset ENG20-00115661-b, la locution adverbiale *sans cesse* et l'adverbe en *-ment perpétuellement*.

En procédant ainsi, nous passons de 1536 à 2149 relations lexème-synset adverbial, soit une augmentation de 28,5%.

3.3. Evaluation du WOLF étendu

Enfin, nous avons procédé à une évaluation manuelle d'une partie significative (435 sur 2149 couples lexème/synset, soit 20%) du WOLF étendu. Nous avons pour cela étudié, pour chaque lexème de la sélection, les synsets qui lui sont attribués et avons affecté un code de validation à chaque association lexème-synset. Les codes de validation que nous avons utilisés sont les suivants :

- OK : l'association est correcte (par exemple *abstraitement* dans le synset auquel est associée dans le Princeton Wordnet la définition *in abstract terms*)

- SC (Semantically close) : l'association est sémantiquement proche et forme un hyponyme, hyperonyme, ou pseudo-synonyme (par exemple *abjectement* dans le synset défini par *of a dreadful kind*)

- SR (Semantically related) : l'association est sémantiquement relié, mais fausse (par exemple *acceptablement* dans le synset défini par *to an unacceptable degree*)

- NR (Non Related) : l'association est totalement fausse, sans aucun lien sémantique entre le lexème et le synset (par exemple *abominablement* dans le synset défini par *to an inexpressible degree*)

- CC (Composed Component) : l'association est fausse, mais le lexème pourrait faire partie d'une composante composée qui, elle, serait correcte (par exemple *addition* au lieu de *en addition* dans le synset défini par *by way of addition*)

- ID (Incorrect Derivation) : l'association est fausse, du fait d'un problème de dérivation (par exemple *absolument* dans le synset défini par *in a royal manner*)

- WC (Wrong Category) : l'association est fausse, du fait d'une mauvaise catégorisation grammaticale du lexème (par exemple *bougonnerie*)

Les résultats sont tout a fait encourageants (voir tableau 1). On obtient en effet près de 68% d'associations lexème-synset correctes (OK).

Total	OK	SC	SR	NR	CC	ID	WC
435	295	**83**	14	16	9	10	8
100%	67,81%	**19,08%**	3,21%	3,67%	2,06%	2,29%	1,83%

Table 1. Résultats de la validation manuelle

4. Couplage des entrées syntaxiques du Le*fff* et les entrées sémantiques du WOLF

Le couplage des ressources lexicales syntaxiques et sémantiques est indispensable pour de nombreuses applications de TAL, notamment pour la construction d'analyseurs syntaxico-sémantiques performants. Les entrées du Le*fff* pour les adverbes en *-ment*, issues en grande partie des tables de (Molinier & Levrier 2000), et les synsets adverbiaux du WOLF, après les extensions décrites à la section précédentes, constituent de bons candidats à une telle opération de couplage.

En effet, si chaque lemme adverbial en *-ment* peut avoir différentes entrées dans les tables de (Molinier & Levrier 2000) et donc dans le Le*fff*, il n'a au plus qu'une seule entrée dans chacune des classes d'adverbes définies par (Molinier & Levrier 2000). Associer un ou plusieurs synsets du WOLF à une entrée du Le*fff* pour un adverbe en *-ment* revient donc à associer à chaque synset une classe d'adverbes de Molinier. Pour ce faire, nous avons simplement associé à chaque synset *S* du WOLF comportant au moins un adverbe en *-ment* la classe de Molinier qui est le plus souvent associée aux lexèmes de *S*. Nous avons alors attribué à chaque entrée du Le*fff* pour le lemme *l* et de classe de Molinier *c* les synsets comportant le lexème *l* et auxquels a été associée la classe *c*.

5. Conclusion et perspectives

A l'heure où l'abscence de ressource lexicale à grande échelle pour le français se fait cruellement sentir, nous avons montré l'importance de faire collaborer différentes ressources existantes, de façon à enrichir ou diversifier les informations qu'elles contiennent. L'interaction Le*fff*-WOLF, *via* l'utilisation de DicoSyn, a permis de faire évoluer chacune de ces ressources vers une plus grande complétude qualitative et quantitative. Ce travail nous a en effet permis d'obtenir une augmentation de près de 55% du nombre de relations lexème-synset adverbial dans WOLF, et d'attribuer des identifiants de synset à de nombreuses entrées adverbiales du Le*fff*.

Ces résultats encourageants montrent également la pertinence de l'exploitation d'un lexique sous forme de graphe, au moins en ce qui concerne l'accès automatique aux informations sémantiques qu'il contient. La synonymie et les adverbes en *-ment* ont constitué un terrain d'expérimentation idéal, et nous incitent à l'exploration d'autres relations, paradigmatiques (hyperonymie, antonymie...) ou syntagmatiques (via l'analyse de corpus), ainsi qu'à d'autres parties du discours, comme par exemple les noms en *-ité* ou les verbes en *-ifier* et *-iser*.

References

DANLOS L. et SAGOT B. (2007), "Comparaison du Lexique-Grammaire et de Dicovalence: vers une intégration dans le Le*fff*", in *Actes de TALN 07*,Toulouse, France.

FELLBAUM C. (1998), *WordNet: An Electronic Lexical Database*, MIT Press.

FRANÇOIS J., VICTORRI B. et MANGUIN J.-L. (2002), "Polysémie adjectivale et synonymie : l'éventail des sens de curieux", in *La polysémie*.

MOLINIER C. et LEVRIER F. (2000), *Grammaire des adverbes. Description des formes en* -ment, Droz, Genève, Suisse.

PLOUX S. et VICTORRI B. (1998), "Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes", in *Traitement Automatique des Langues (T.A.L.)*, n° 1, vol. 39.

SAGOT B., CLÉMENT L., VILLEMONTE DE LA CLERGERIE E. et BOULLIER P. (2006), "The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use", in *Proc. of LREC'06*.

SAGOT B. et DANLOS L. (2007), "Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – Constructions impersonnelles et expressions verbales figées", in *Cahiers du Cental*.

SAGOT B. et FIŠER D. (2008), "Building a free French wordnet from multilingual resources", in *Actes de Ontolex 2008*,Marrakech, Maroc, (à paraître).

SAGOT B. et FORT K. (2007), "Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire - Adverbes en -ment.", in *Actes du Colloque Lexique et Grammaire*,Bonifacio, France.

VENANT F. (2004), "Polysémie et calcul du sens", in *Le poids des mots, Actes de JADT 2004*,Louvain-la-Neuve, Belgique.

VENANT F. (2007), "Une exploration géométrique de la structure sémantique du lexique adjectival français", in *Traitement Automatique des Langues (T.A.L.)*, n° 2, vol. 47.

VOSSEN, P. (1999), *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer, Dordrecht.