Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes

Nabil Hathout
ERSS – CNRS & Université de Toulouse Le Mirail
Maison de la Recherche. F-31058 Toulouse cedex 1. France
Nabil.Hathout@univ-tlse2.fr

Résumé - Abstract

Cet article présente une méthode de construction automatique de liens morphologiques à partir d'un dictionnaire de synonymes. Une analyse de ces liens met en lumière certains aspects de la structure morphologique du lexique dont on peut tirer partie pour identifier les variations allomorphiques des suffixations extraites.

We present a method to extract morphological links from a synonym dictionary without supervision. The analysis of these links brings somes aspects of the morphological structure of the lexicon to light. We have then taken advantage of this structure to identify allomorphic variations of the extracted suffixations.

Mots-clés : morphologie dérivationnelle ; analogie ; structure du lexique.

Key words: derivational morphology; analogy; structure of the lexicon.

1 Ressources de langue générale pour le TALN

C'est un fait bien établi mais qu'il est toujours bon de rappeler : les ressources de langue générale sont utiles pour le TALN et valent la peine d'être construites. Par exemple, la disponibilité d'un dictionnaire comme WordNet a donné lieu à des travaux sur la construction de concordances sémantiques, sur l'identification des sens ou sur le rattachement des arguments, etc. dont plusieurs sont présentés dans (Fellbaum 1999). De même, la base de données morphologiques CELEX a été utilisée avec succès en recherche d'information (Jing et Tzoukerman 1999) et pour la constitution de ressources terminologiques (Daille et Jacquemin 1998). Pour le français, les recherches menées par (Hamon et al. 1999) ont montré l'intérêt que présente un dictionnaire de synonymes pour la structuration des terminologies. Malheureusement, les ressources de langue générale restent peu nombreuses pour le français et difficiles à acquérir (non disponibles ou relativement onéreuses). Signalons cependant la disponibilité du dictionnaire *EuroWordNet* qui comporte 18 777 lemmes répartis en 22 745 *synsets* (ensembles de synonymes). En revanche il n'existe pas encore d'équivalent de la base de données morphologiques CELEX.

Ce travail, réalisé dans le cadre du projet MorTAL (Hathout et al. 2002), s'inscrit dans le cadre de la constitution d'une telle ressource.

Il existe en fait au moins quatre ressources pour le français qui comportent des descriptions constructionnelles[1]. Il s'agit (1) du *Dictionnaire des radicaux DELAR* constitué par le LADL (Courtois 1990) et dont une version est distribuée par ELRA/ELDA : *Dictionnaire des verbes français*. Ce dictionnaire comporte 25 610 verbes pour lesquels il décrit les lexèmes adjectivaux construits en *-able*, *-ant* et *-é*, et les lexèmes nominaux construits en *-age*, *-ment*, *-tion*, *-oir* et *-ure*. Par ailleurs, certaines tables du Lexique-Grammaire (Gross 1975) comportent des informations sur les affixes nominaux et adjectivaux qui peuvent opérer sur les verbes qu'elles décrivent, mais cette ressource n'est malheureusement pas (plus) disponible. (2) La société Mémodata distribue elle aussi un *Dictionnaire dérivationnel* sous forme d'une bibliothèque logicielle intégrable dans des applications commerciales. (3) C. Gruaz a pour sa part construit le *Dictionnaire synchronique de familles morphologiques de mots français* (DISFA) (Gruaz 1997) mais nous ignorons sa disponibilité. (4) Nous avons nous-même construit, à partir de *TLFnome* (lexique de formes construit à partir de la nomenclature du *Trésor de la Langue Française*) *Verbaction*, un lexique de noms d'actions déverbaux qui comporte 6 472 entrées.

2 Morphologie constructionnelle

Dans des travaux précédents, nous avons utilisé *TLFnome* pour produire des entrées d'une base de données constructionnelles (Hathout 2000). Nous travaillons également sur la classification en familles constructionnelles. La technique mise en œuvre consiste à apprendre, à partir du lexique, un ensemble de schémas de préfixation et/ou de suffixation puis à les appliquer pour apparier des lexèmes bases et des lexèmes supposés construits. (La même technique est utilisée par (Grabar et Zweigenbaum 1999).) Ces associations sont ensuite filtrées en s'appuyant sur les nombres d'instances des schémas. Les résultats que nous avons obtenus sont de qualité variable, très bons pour le suffixe *-able* (précision de l'ordre de 95%), très insuffisants pour le suffixe *-is-* (*-iser* ; la précision est inférieure à 50%). Cette différence s'explique essentiellement par la plus grande homogénéité de l'ensemble des bases de *-able* (presque toutes des verbes) par rapport à celui des bases de *-is-* (les substantifs et les adjectifs y sont en nombre comparable).

Ces problèmes de précision ont pour origine la nature du lexique utilisé : un lexique de formes graphématiques qui ne comporte pas de description phonétique ni sémantique. Or la construction morphologique est précisément une relation régulière entre des lexèmes qui partagent simultanément des propriétés phonétiques et sémantiques. Les formes graphémiques sont des approximations raisonnables des formes sonores et permettent de calculer avec une précision acceptable les partages de propriétés phonétiques. Elles ne sont en revanche pas suffisamment précises pour décider s'il y a partage de propriétés sémantiques. L'amélioration des performances globales de l'appariement et de la classification en familles passe nécessairement par l'augmentation de la précision de cette seconde approximation. Une solution sure et facile à mettre en œuvre consiste à utiliser un dictionnaire de synonymes qui, précisément, décrit des relations de partage de propriétés sémantiques, même si ces dernières ne coïncident pas avec les relations constructionnelles. Par ailleurs, cette ressource est déjà validée et donc relativement

[1] Nous avons choisi de suivre D. Corbin en préférant le terme *constructionnel* à *dérivationnel* parce qu'il est plus explicite que le second et qu'il est relativement neutre du point de vue théorique (*dérivation* sous-entend l'existence de règles, de niveaux de représentation...).

fiable.

La suite de l'article s'organise comme suit : nous présentons en §3 la technique utilisée pour identifier les relations constructionnelles dans le dictionnaire de synonymes. Les résultats obtenus, analysés en §4, font clairement apparaître une structure morphologique du lexique. §5 est consacré au problème de la variation morphologique, à savoir l'allomorphie. Nous comparons ensuite, en §6, ce travail avec d'autres recherches proches par leurs objectifs ou par les techniques qu'elles proposent.

3 Identification des analogies morpho-synonymiques

L'identification des relations constructionnelles[2] utilisée ici est basée sur la technique que nous proposons dans (Hathout et al. 2002). Elle est cependant plus précise que cette dernière car elle exploite les informations relatives au partage de propriétés sémantiques pour filtrer les relations morphologiques apprises automatiquement. Le filtrage se fait en extrayant du dictionnaire des « analogies morpho-synonymiques », à savoir des quadruplets de lemmes (x_1, x_2, y_1, y_2) qui forment des séries proportionnelles (Cruse 1986) comme celle qui est présentée en figure 1. Les relations de cette figure marquées par des lignes pleines doivent se lire de la manière sui-

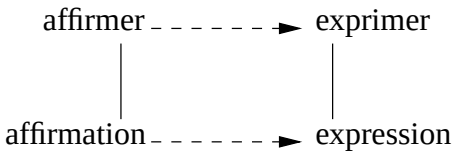

FIG. 1 – Analogie morpho-synonymique concernant le couple *affirmer*:*affirmation*. Les lignes pleines correspondent à des relations constructionnelles et les flèches en pointillé à des relations de synonymie, en fait de proximité sémantique forte.

vante : x_1 = *affirmation* est à x_2 = *affirmer* ce que y_1 = *expression* est à y_2 = *exprimer*. Plus généralement, une analogie (x_1, x_2, y_1, y_2) est correcte :

(1) a. si (x_1, x_2) et (y_1, y_2) sont morphologiquement apparentés et
    b. s'il existe une relation sémantique qui unit à la fois (x_1, x_2) et (y_1, y_2).

La polysémie éventuelle des lexèmes de l'analogie fait qu'il peut exister plusieurs relations sémantiques S telles que $S(u_1, u_2)$ et $S(v_1, v_2)$.

Nous avons utilisé le dictionnaire de synonymes constitué par Jean-Yves Hamon (CNRS-INaLF, USR 705). Il s'agit en fait de la compilation de 7 dictionnaires d'époques et de tailles différentes. L'hétérogénéité de ces dictionnaires permet de gommer les spécificités de chacun et garantit une certaine généralité à notre étude et aux résultats obtenus. L'extraction des analogies morpho-synonymiques est réalisée en trois étapes : (1) formatage des entrées ; (2) construction d'un graphe constructionnel ; (3) extraction des analogies. Les étapes 1 et 3 sont destinées au filtrage des relations constructionnelles identifiées par l'étape 2. La ligne directrice pour les premières est donc de privilégier la précision à la couverture ; inversement, l'étape 2 vise à construire un graphe dont la couverture est maximale.

[2] Nous ne considérons dans cet article que la construction suffixale.

Formatage.< Le formatage des entrées consiste essentiellement en la catégorisation des entrées et des synonymes. La catégorisation est réalisée à l'aide du lexique *TLFnome+index* en exploitant le fait que la synonymie s'établit entre des lemmes de même catégorie. Cette contrainte implique une séparation préalable des différents sens et acceptions de l'entrée. Une technique robuste basée sur l'appariement suffixal est utilisée pour les lemmes qui n'appartiennent pas à ce lexique de référence. La séparation des sens maximise le nombre des propriétés sémantiques partagées par l'entrée et chacun de ses synonymes ainsi que par ces derniers entre eux. Cela n'a cependant pas d'incidence sur la précision des analogies car le formatage préserve l'orientation des descriptions synonymiques.

Graphe constructionnel. La deuxième étape consiste en une simple application de la technique présentée dans (Hathout et al. 2002) à l'ensemble des lemmes qui figurent en entrée ou comme synonymes. Dans un premier temps, un ensemble de schémas d'appariement est appris en comparant deux à deux l'ensemble des lemmes. (Comme indiqué *supra*, seuls les suffixes sont considérés.) On conserve les schémas (s_1, s_2) tels qu'il existe au moins trois couples de lemmes (m_1, m_2) tels que $m_1 = r \cdot s_1$ et $m_2 = r \cdot s_2$ partagent un préfixe r d'au moins trois caractères. s_1, s_2, m_1, m_2 sont des chaînes de caractères munies de catégories grammaticales ; r est seulement une chaîne de caractères. (s_1, s_2) est la signature constructionnelle des couples (m_1, m_2)[3]. Les schémas appris sont ensuite appliqués aux lemmes pour reconstruire des couples.

Analogies morpho-synonymiques. L'extraction des analogies morpho-synonymiques consiste en une simple exploration du graphe correspondant aux appariements suffixaux pour trouver l'ensemble des quadruplets (x_1, x_2, y_1, y_2) tels que (x_1, x_2) et (y_1, y_2) sont dans le graphe et que y_1 (resp. y_2) figure dans une entrée de x_1 (resp. x_2). Dans ce qui suit, on appelle signature d'une analogie (x_1, x_2, y_1, y_2) le couple des signatures de (x_1, x_2) et de (y_1, y_2). Signalons que trois filtres supplémentaires ont été utilisés afin d'éliminer (*i*) les analogies dont deux éléments ou plus sont identiques ; (*ii*) les analogies dont l'un des couples peut être construit en préfixant le second, comme appeler/Vmn----:approcher/Vmn---->rappeler/Vmn----:rapprocher/Vmn----[4] (cet exemple illustre bien le type de problème rencontré) ; (*iii*) les analogies dont les couples (x_1, x_2) et (y_1, y_2) appartiennent à la même famille ou plus exactement partagent un même radical graphémique comme bassin/Ncms:bassinage/Ncms>bassinoire/Ncfs:bassinement/Ncms (dans la grande majorité de ces analogies, (x_1, x_2) et (y_1, y_2) sont dans des relations sémantiques différentes).

Les résultats obtenus sont très satisfaisants : la précision évaluée en utilisant un échantillon $\mathcal{E}$ de 200 analogies choisies au hasard parmi les 73 018 produites est de 94%. Parmi les 12 analogies erronées, 7 ne respectent pas la contrainte (1a) comme trait/Ncms:traiter/Vmn----> expression/Ncfs:exposer/Vmn---- et 5 ne respectent pas (1b) comme rigoler/Vmn----:rigolade/Ncfs> blaguer/Vmn----:blague/Ncfs. L'échantillon comporte 386 couples de lexèmes différents dont 376 sont corrects, soit une précision de 97%[5].

[3] Pour une présentation formelle de ce type d'analogies, voir (Pirrelli et Yvon 1999).

[4] Les étiquettes morpho-syntaxiques sont celles de l'action GRACE (Rajman et al. 1997).

[5] À titre de comparaison, un filtrage des appariements basé sur une classification en familles constructionnelles, sans recours à aucune information sémantique, permet d'obtenir une précision comprise entre 75% et 85%.

4 Analogies morpho-synonymiques et structuration du lexique

Les analogies obtenues peuvent être réparties en deux groupes : les analogies strictes sont telles que les schémas constructionnels qui unissent les couples de droite et de gauche sont identiques (c'est le cas de adoration/Ncfs:adorer/Vmn---- et vénération/Ncfs:vénérer/Vmn---- qui partagent le schéma ation/Ncfs:er/Vmn) ; et les analogies lâches sont celles dont les schémas sont différents. Ces dernières peuvent être divisées en trois sous-groupes : celles dont les deux schémas apparaissent dans une analogie stricte (lâche-1) comme changer/Vmn----:changement/Ncms et permuter/Vmn----:permutation/Ncfs, dont les schémasr/Vmn----:ment/Ncms et er/Vmn----:ation/Ncfs sont tous deux stricts ; celles dont un seul schéma y apparaît (lâche-2) comme réunion/Ncfs:réunir/Vmn---- etmélange/Ncms:mélanger/Vmn---- dont seul /Ncms:r/Vmn---- est strict ; celles dont aucun des deux schémas n'y apparaît (lâche-3) comme bon/Ncms:bonté/Ncfs et fort/Ncms:force/Ncfs. Le tableau en (2) montre clairement que ces 4 types d'analogies n'ont pas la même importance. La 4e colonne présente le nombre moyen d'analogies par signature du type correspondant. Cette classification doit néanmoins être affinée en prenant en compte les allomorphies (cf. §5).

(2)

type	# sign.	# analogies	moyenne
stricte	978	15 632	15,98
lâche-1	6 794	22 664	3,34
lâche-2	15 018	21 800	1,45
lâche-3	5 780	5 922	1,02

Les spécificités sémantiques et phonologiques des affixes expliquent l'importance particulière des analogies strictes. (Elles ont un nombre d'instances moyen nettement plus élevé que les analogies lâches.) Ces dernières s'imposent ainsi comme les analogies canoniques (en fait prototypiques) et comme le principal axe de structuration paradigmatique du lexique par la morphologie constructionnelle. Les différences qui existent entre les analogies lâches sont en revanche moins attendues : les analogies qui s'établissent entre des schémas stricts sont plus fortes que celles qui s'établissent entre un schéma strict et un schéma lâche ou entre deux schémas lâches[6]. Le tableau en (3) présente la répartition des schémas pour chaque type d'analogie ainsi que le nombre de couples de lemmes qu'ils connectent.

(3)

type d'analogies	# schémas stricts	lâches-2	lâches-3	total	# const.	moyenne
stricte	978	0	0	978	13 976	14,29
lâche-1	782	0	0	782	17 200	22,99
lâche-2	768	4 012	0	4 780	18 118	3,79
lâche-3	0	2 154	2 206	4 360	6 744	1,55

Ces résultats confirment les hypothèses du Modèle en Réseau proposé par J. Bybee (Bybee 1988) à savoir que les schémas constructionnels n'ont pas tous la même force. Les schémas stricts constituent des points (en fait des sous-graphes) de référence par rapport auxquels se définit le sens de la plupart des schémas. Ainsi les analogies lâches-1 définissent le sens des schémas stricts. On constate que cette définition est très complète puisque chaque schéma strict

[6]Un schéma constructionnel est dit strict s'il apparaît dans une analogie stricte et lâche sinon. Parmi ces derniers, ceux qui apparaissent dans une analogie lâche-2 sont dit lâches-2. Les autres sont lâches-3.

apparaît en moyenne dans 8,68 signatures d'analogies lâches-1. Ceci renforce d'avantage la capacité des schémas stricts à servir de référence. Les analogies lâches-2 servent à décrire le sens des schémas lâches en fonction de schémas stricts. On constate que ce mode de définition est plus utilisé que la définition du sens des schémas lâches par rapport à eux-mêmes. En particulier, il y a deux fois plus de schémas lâches-2 que de schémas lâches-3 (3e et 4e lignes du tableau en (3)).

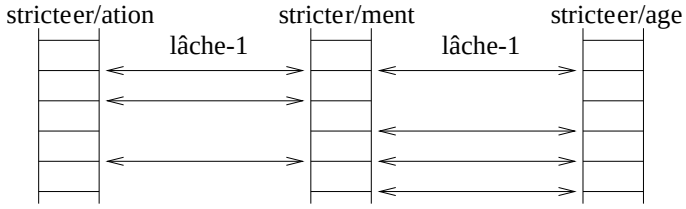

FIG. 2 – Structuration du lexique par les analogies strictes et de type lâches-1.

Il ressort des chiffres des tableaux (2) et (3) qu'il existe une distinction nette entre les analogies lâches du 1er type qui participent fortement à la structuration du lexique (cf. figure 2) et celles des types 2 et 3 sont plus périphériques. On constate également que, dans l'échantillon $\mathcal{E}$, sur les 7 erreurs constructionnelles repérées, quatre analogies sont du type lâche-3 et trois du type lâche-2. On peut donc, dans la perspective de la constitution d'une ressource de langue générale fiable, se limiter aux seules analogies strictes et lâches-1, soit 7 772 analogies sur 28 570 et 24 464 couples de lexèmes sur 39 008.

5 Allomorphies

Le typage des analogies et des schémas est déterminant pour la constitution d'un lexique constructionnel, mais aussi pour son utilisation. Or, les résultats présentés en §4 sont incomplets parce qu'ils ne prennent pas en compte les allomorphies. Par exemple, l'analogie péril/Ncms:périlleux/Afpms>danger/Ncms:dangereux/Afpms est de type lâche-2 alors qu'il s'agit, du point de vue constructionnel, d'une variante de hasard/Ncms:hasardeux/Afpms> danger/Ncms:dangereux/Afpms qui, elle, est stricte. L'identification de cette allomorphie permettrait d'une part de récupérer la 1re analogie comme stricte et sure, mais aussi de réévaluer l'importance du schéma /eux/Ncms/Afpms.

La bonne qualité générale des analogies morpho-synonymiques permet de traiter efficacement ces variations. Comme nous venons de le voir, les analogies s'établissent prioritairement entre des couples de lexèmes connectés par le même schéma constructionnel. Elles constituent de ce fait un contexte privilégié pour identifier les allomorphies.

(4) a. approbation/Ncfs:approuver/Vmn---->adhésion/Ncfs:adhérer/Vmn----
b. stérile/Afpms:stérilisation/Ncfs>aseptique/Afpms:aseptisation/Ncfs
c. brutalement/Rgp:brutalité/Ncfs>durement/Rgp:dureté/Ncfs
d. vision/Ncfs:visiblement/Rgp>sensation/Ncfs:sensiblement/Rgp[7]

Les analogies allomorphiques peuvent être réparties en quatre types selon que les schémas qui les composent sont simples ou composés et qu'ils correspondent à une ou deux relations de

[7] Exemple forgé.

construction (cf. (4)). Par exemple l'analogie (4a) comporte une allomorphie pour les substantifs tandis que (4c) en présente deux, l'une pour les substantifs et l'autre pour les adverbes. Les allomorphies 1-simples sont naturellement les plus intéressantes dans la mesure où ce sont les plus élémentaires et où celles qui correspondent aux trois autres types peuvent être retrouvées à partir de ces dernières.

L'identification des analogies allomorphiques est basée sur quatre critères. Le premier (critère A) permet de distinguer les allomorphies constructionnelles des allomorphies flexionnelles (qui concernent par exemple les participes passés et présents convertis en adjectifs) et des analogies composées de schémas non allomorphes. Ce critère reprend en partie celui que propose (Jacquemin 1997) qui décrit les allomorphies entre deux suffixes s_1 et s_2, de tailles n_1 et n_2, à l'aide d'un triplet d'entiers (n, n', n'') où n est la taille la partie finale commune à s_1 et s_2 ; soit s cette partie ; $n' = \min(n_1, n_2) - n$ est le nombre de caractères dans le plus petit des deux suffixes qui se trouvent avant s ; $n'' = \max(n_1, n_2) - n$ est le nombre de caractères dans le plus grand des deux suffixes qui se trouvent avant s. Ce découpage est illustré en figure 3 pour le couple

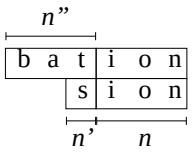

FIG. 3 – Découpage des suffixes allomorphiques bation:sion.

bation:sion auquel il associe le triplet $(3, 1, 3)$. La signature uver:rer est pour sa part caractérisée par le triplet $(2, 1, 2)$. À chaque signature d'analogie correspond ainsi un couple de triplets qui permet de déterminer si elle est susceptible d'être allomorphique. Notre objectif étant différent de celui de (Jacquemin 1997), nous avons défini le critère A directement à partir de ces 6 valeurs et non pas sous la forme d'une distance et d'un seuil. Une signature d'analogie caractérisée par les triplets $\{(n_1, n'_1, n''_1), (n_2, n'_2, n''_2)\}$ vérifie le critère A si :

(5) a. $\max(n_1, n_2) \geq 3$,
b. $|n'_1 - n'_2| \leq 2$, $|n''_1 - n''_2| \leq 2$ et
c. $|n'_1 - n'_2| + |n''_1 - n''_2| \leq 3$.

Les conditions (5a) et (5c) sont destinées à éliminer les analogies non allomorphiques ainsi que les allomorphies flexionnelles comme fanée/Ncfs:faner/Vmn---->flétrie/Ncfs:flétrir/Vmn----. Notre critère se distingue de celui de Ch. Jacquemin par la condition (5b) qui exprime le fait qu'une allomorphie doit concerner les deux schémas constructionnels de manière « comparable ». En effet, les allomorphies sont des variations induites par des contraintes morpho-phonologiques sur la forme construite (output) qui concernent certaines configurations des couples base + suffixe (input). La condition (5b) exprime le fait que les variations de l'output sont induites par les spécificités formelles de l'input. Le critère A est relativement sélectif puisqu'il n'est vérifié que par 3 354 signatures d'analogies, soit 18% de celles qui sont lâches et isocatégorielles. Sa précision, estimée à partir d'un échantillon de 100 signatures satisfaisant ce critère et choisies au hasard est de 62%[8]. Les erreurs se répartissent comme présenté dans la partie droite du tableau en (6). La partie gauche présente une estimation de la proportion des signatures d'analogies allomorphiques pour chacun des trois types. On constate que

[8]Dans la suite de l'article, les précisions des critères sont toutes estimées à partir d'échantillons de 100 signatures choisies au hasard.

la vérification du critère A semble indépendante du type des analogies et que son efficacité est satisfaisante, surtout pour le type lâches-1.

(6)

	témoin			critère A		
type	# sign.	# allom.	prop.	# sign.	# allom.	précision
lâche-1	25	8	32%	22	17	78%
lâche-2	63	14	22%	62	39	63%
lâche-3	12	0	0%	16	6	38%
total	100	22	22%	100	62	62%

L'efficacité du critère A peut être améliorée en lui adjoignant un autre critère permettant d'éliminer certaines des signatures qui posent le plus de problèmes : les 2-simples et les 2-complexes. Ces signatures sont caractérisées par le fait qu'elles entrent généralement dans des configurations « transitives ». Le critère B exploite cette propriété. Il stipule qu'une signature d'analogie (u_1, u_2, v_1, v_2) ou (v_1, v_2, u_1, u_2) doit être éliminée s'il existe un suffixe u_3 et deux schémas constructionnels (u'_1, u'_3) et (u'_3, u'_2) qui apparaissent dans deux signatures tels que :

(7) a. $(u_1, u_3) \sqsubseteq (u'_1, u'_3)$ et $(u_2, u_3) \sqsubseteq (u'_2, u'_3)$

b. $\text{freq}(u_1, u_2) \leq \text{freq}(u'_1, u'_3)$ et $\text{freq}(u_1, u_2) \leq \text{freq}(u'_3, u'_2)$

où $\sqsubseteq$ dénote une relation de subsomption entre deux schémas constructionnels tels que le premier est une spécialisation du second. Formellement, $(u_1, u_3) \sqsubseteq (u'_1, u'_3)$ si u'_1 est un suffixe de u_1, u'_3 est un suffixe de u_3 et $t(u_1) - t(u'_1) = t(u_3) - t(u'_3)$, où $t(u)$ est le nombre de caractères du suffixe u. $\text{freq}(u, u')$ est la fréquence associée au schéma (u, u') calculée lors de l'apprentissage des schémas à partir du lexique des formes.

Par exemple, le fait qu'il existe un schéma arité/ier/Ncfs/Afpms et un schéma er/èrement/Afpms/Rgp qui subsume ier/ièrement/Afpms/Rgp permet d'éliminer le schéma 2-simple arité/ièrement/Ncfs/Rgp. La partie gauche du tableau en (8) présente une évaluation de la précision des critères A et B lorsqu'ils sont appliqués séparément sur l'ensemble des analogies et que seules les signatures qui vérifient les deux sont conservées.

(8)

	critères A et B			critères A, B, C et D		
type	# sign.	# allom.	préc.	# sign.	# allom.	préc.
lâche-1	38	34	89%	40	39	97%
lâche-2	52	45	84%	53	45	84%
lâche-3	10	6	60%	7	6	85%
total	100	85	85%	100	90	90%

Deux conditions supplémentaires ont été utilisées pour améliorer ces résultats. La première (critère C) consiste à éliminer les signatures dans lesquels apparaissent des schémas constructionnels symétrisables. Un schéma (u_1, u_2) est symétrisable s'il existe un schéma (v_1, v_2) tel que ces quatre suffixes apparaissent dans deux signatures (u_1, u_2, v_1, v_2) et (u_2, u_1, v_1, v_2) ou (v_1, v_2, u_1, u_2) et (v_1, v_2, u_2, u_1). De tels schémas correspondent à une relation sémantique lâche ou difficile à exprimer dans un dictionnaire de synonymes, et sont donc susceptibles de provoquer des erreurs. La seconde (critère D) est une contrainte sur la taille maximale que peut avoir $t(u_1) + t(u_2)$, en l'occurrence $t(u_1) + t(u_2) \leq 10$. Les signatures éliminées par ce critère, généralement 1-complexes, 2-simples ou 2-complexes sont impliquées dans la plupart des erreurs qui échappent aux critères A et B. La partie droite du tableau en (8) présente une estimation de

la précision lorsque les quatre critères sont appliqués (le filtrage par C et D précède celui par B). On constate que les critères C et D sont globalement efficaces et qu'ils améliorent la précision pour les signatures lâches-1 au point où il devient envisageable de les utiliser directement.

6 Travaux connexes

Deux articles, (Grabar et Zweigenbaum 1999) et (Jacquemin 1997), sont à l'origine du travail présenté ici. Le premier, très proche du nôtre, exploite les liens de synonymie présents dans le Microglossaire SNOMED pour identifier les relations de parenté morphologique entre les termes de la terminologie CIM-10. La démarche de ces auteurs diffère cependant de la nôtre par le fait que ces relations ne sont pas filtrées par la contrainte d'analogie du fait de la taille réduite du thésaurus initial (5 801 termes). Le traitement des allomorphies présenté en §5 est directement inspiré du second article. Dans ce travail, Ch. Jacquemin s'intéresse aux variations morpho-syntaxiques de bi-termes. Ces dernières sont similaires à nos analogies à ceci près que la relation entre les deux termes est syntagmatique tandis que celle qui unit les synonymes est paradigmatique.

Signalons par ailleurs qu'au moins deux autres projets de recherche utilisent comme ressource des dictionnaires de synonymes. (Hamon et al. 1999) se servent des renvois synonymiques du Robert pour enrichir la structure d'un thésaurus. Leur travail montre clairement que les ressources de langue générale et de langues de spécialité sont complémentaires. La première s'avère même supérieure aux secondes en terme de rappel. En contrepartie, la précision des liens inférés est faible. Pour leur part, (Ploux et Victorri 1998) utilisent le même dictionnaire de synonymes que nous et s'intéressent eux aussi à la structure du lexique. Leur travail se distingue néanmoins du nôtre sur plusieurs points : il ne vise pas à construire ou à enrichir une ressource pour le TALN ; l'étude porte uniquement sur la polysémie dans le cadre d'une théorie « continuiste » du sens ; les synonymes ne sont pas étiquetés ; les relations synonymiques sont symétrisées ; les données sont utilisées telles qu'elles, sans aucun filtrage.

7 Conclusion

Nous avons présenté une méthode permettant d'extraire automatiquement un lexique constructionnel à partir d'un dictionnaire de synonymes. Cette méthode est indépendante des langues particulières puisqu'elle ne met en œuvre aucune connaissance linguistique. Elle est également indépendante des dictionnaires particuliers et devrait être utilisable directement avec d'autres ressources comme les dictionnaires WordNet et EuroWordNet. Ce travail a également permis de mettre au jour certains aspects de la structure morphologique du lexique et de proposer un ensemble de critères qui permettent d'identifier les variations allomorphiques de manière très précise.

Remerciements

Ce travail a bénéficié d'un financement du Ministère de l'Éducation Nationale, de la Recherche et de la Technologie dans le cadre de l'Action concertée incitative « jeunes chercheurs » (1999-

2002). Je voudrais remercier, à titre posthume, Jean-Yves Hamon pour le dictionnaire qu'il a mis à ma disposition. Je remercie également Ludovic Tanguy ainsi que les deux relecteurs anonymes pour leurs commentaires constructifs.

Références

Bybee, J. L. (1988). Morphology as Lexical Organization. In Hammond, M. et Noonan, M., éditeurs, *Theoretical Morphology. Approaches in Modern Linguistics*, chapitre 7, pp. 119–141. Academic Press, San Diego, CA.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, 87:11–22.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Daille, B. et Jacquemin, C. (1998). Lexical Database and Information Access: A Fruitful Association. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pp. 669–673, Granada, Espagne. ELRA.

Fellbaum, C., éditeur (1999). *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Grabar, N. et Zweigenbaum, P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Actes de la 6^{e}Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-98)*, pp. 175–184, Cargèse.

Gross, M. (1975). *Méthodes en syntaxe : Régime des constructions complétives*. Hermann, Paris.

Gruaz, C. (1997). La stratification dérivationnelle dans les familles synchroniques des mots français comtemporain. In *III Coloquio Internacional de Lingüística Francesa*, Salamanca, Espagne.

Hamon, T., Garcia, D., et Nazarenko, A. (1999). Détection de liens de synonymie : complémentarité des ressources générales et spécialisées. In *Actes de Terminologie et Intelligence Artificielle*, pp. 45–58, Nantes, France.

Hathout, N. (2000). Morphological Pairing based on the Network Model. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, pp. 35–38, Pyrgos, Grèce.

Hathout, N., Namer, F., et Dal, G. (2002). An Experimental Constructional Database : The MorTAL Project. In Boucher, P., éditeur, *Many Morphologies*. Cascadilla, Cambridge, Mass. À paraître.

Jacquemin, C. (1997). Guessing Morphology from Terms and Corpora. In *Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pp. 156–167, Philadelphia, PA. ACM.

Jing, H. et Tzoukerman, E. (1999). Information Retrieval based on Context Distance and Morphology. In *Proceedings of 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 90–96, Berkeley, CA. ACM.

Pirrelli, V. et Yvon, F. (1999). The hidden dimension: a paradigmatic view of data-driven NLP. *Journal of Experimental & Theoretical Artificial Intelligence*, 1999(11):391–408.

Ploux, S. et Victorri, B. (1998). Constructions d'espaces sémantiques à l'aide de dictionnaires de synonymes. *T.A.L.*, 39(1):161–182.

Rajman, M., Lecomte, J., et Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Rapport GRACE GTR-3-2.1, EPFL & INaLF.