L'analogie, un moyen de croiser les contraintes et les paradigmes

Acquisition de connaissances morphologiques à partir de dictionnaires de synonymes

Nabil Hathout

ERSS (CNRS - UTM)
Maison de la Recherche
5, allées Antonio Machado
F-31058 Toulouse cedex 9
Nabil.Hathout@univ-tlse2.fr

RÉSUMÉ. *L'article décrit une méthode d'acquisition de connaissances morphologiques constructionnelles (dérivationnelles) à partir de dictionnaires de synonymes. Cette méthode, destinée à la création semi-automatique de bases de données constructionnelles, exploite de différentes manières la structure paradigmatique du lexique. Elle repose sur l'identification de quadruplets analogiques (morpho-synonymiques) qui permettent de croiser des contraintes sémantiques définies au moyen de relations synonymiques et des contraintes morphographiques. La méthode est robuste et indépendante vis-à-vis des langues particulières. Elle a été utilisée avec succès pour des dictionnaire de synonymes français et dictionnaires. Nous proposons en outre un typage des quadruplets morpho-synonymiques qui rend explicite le fait que certains couples de lexèmes sont soumis à davantage de contraintes que d'autres.*

ABSTRACT. *The paper proposes a technique to acquire morphological constructional (derivational) relations from dictionaries of synonyms in order to create morphological databases semi-automatically. It exploites the paradigmatic structure of the lexicons in several ways. The technique is based on the discovery of analogical (morpho-synonymic) quadruplets where semantic contraints defined as synonymic relations are combined with morphographical constraints. It is robust and language independent: it has been successfully applied to French and English dictionnaries of synonyms. The technique can be enhanced by typing the morpho-synonymic quadruplets in order to make explicit the fact that some pairs of lexemes are submitted to more constraints than others.*

MOTS-CLÉS : *lexique, morphologie computationnelle, suffixation, analogie morphographique, synonymie, apprentissage non supervisé.*

KEYWORDS: *lexicon, computational morphology, suffixation, morphographical analogy, synonymy, unsupervised learning.*

1. Introduction

L'analogie tient une place centrale dans l'organisation morphologique du lexique des langues. Elle est à l'origine de la structure paradigmatique du lexique tant sur le plan flexionnel que constructionnel[1]. Par exemple, les formes fléchies d'un verbe récent comme *badger* entrent dans le paradigme flexionnel d'autres verbes ayant la finale *-ger* comme *bridger* ou *diverger*. De même, les affixes constructionnels définissent des paradigmes qui peuvent être complétés par analogie : *écrantage* désigne l'action d'*écranter* comme *blindage* l'action de *blinder*.

Dans cet article, nous présentons une méthode permettant de tirer profit de la structure paradigmatique du lexique en utilisant l'analogie pour définir des contraintes qui opèrent simultanément aux niveaux graphémique et sémantique. La méthode est destinée à acquérir des connaissances morphologiques constructionnelles à partir de ressources lexicales comme des dictionnaires de synonymes ou des bases de données WordNet [MIL 90]. Ces connaissances permettent de créer de manière semi-automatique, des lexiques constitués de couples de lexèmes morphologiquement apparentés qui peuvent être utilisés pour le traitement automatique des langues (TAL), la recherche d'information (RI) et la réalisation d'expériences psycholinguistiques. Nous nous intéressons ici en particulier aux possibilités qu'offre l'analogie de croiser différentes contraintes et d'améliorer ainsi les résultats de l'acquisition.

Sur le plan linguistique, notre approche est compatible avec le « Modèle en réseau » proposé par J. Bybee [BYB 88, BYB 95]. Dans ce modèle, le lexique est conçu comme un graphe de formes attestées connectées les unes aux autres par des relations de partage de forme et/ou de sens. La construction morphologique est définie de manière classique comme une relation entre des lexèmes[2] qui partagent en même temps des propriétés phonologiques et des propriétés sémantiques. Les affixes constructionnels (préfixes ou suffixes) sont vus comme des sous-graphes qui forment des séries proportionnelles [CRU 86] et que l'on peut étendre par analogie. On peut ainsi représenter graphiquement une portion du sous-graphe qui correspond au suffixe *-able* comme en figure 1. Dans ce schéma, les arcs qui relient *contrôlable* à *contrôler* rendent compte du partage des propriétés phonologiques et sémantiques correspondant au radical de l'adjectif (et à celui du verbe). De même, les arcs qui relient *contrôlable* à *agitable*, *gonflable*, *activable*... correspondent au partage de son et de sens qui peut être associé au suffixe *-able*.

Ce travail s'inscrit dans le cadre du projet MorTAL (acronyme de « MORphologie pour le TAL ») dont l'objectif est de constituer une base de données morphologiques à large couverture pour le français [HAT 02]. À terme, cette base devrait fournir une

1. Nous adopterons ici la terminologie de D. Corbin [COR 01] sans pour autant nous placer dans le cadre du Modèle SILEX. Nous utiliserons le terme *constructionnel* de préférence à *dérivationnel* car il nous semble plus explicite et plus neutre du point de vue théorique (il ne présuppose pas l'existence de règles, de niveaux de dérivation...).
2. Les lexèmes peuvent être définis comme des sous-graphes constitués de formes qui ne diffèrent que par leurs marques flexionnelles.

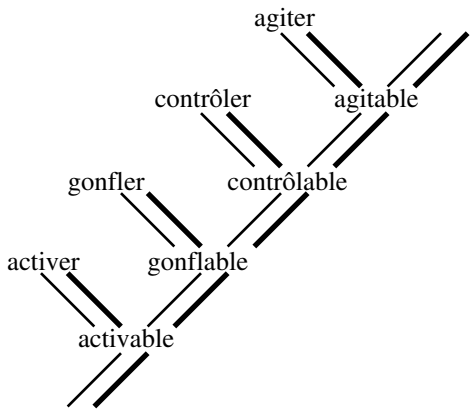

Figure 1. *Portion du sous-graphe correspondant au suffixe* -able. *Les lignes fines représentent un partage de son et les lignes épaisses un partage de sens*

description de la morphologie constructionnelle du français similaire à celles que la base CELEX [BAA 95] offre pour le néerlandais, l'anglais et l'allemand. Les bases de données morphologiques sont utilisées en TAL et en RI pour identifier des variantes morphosyntaxiques dans les documents [JAC 99]. Par exemple, un moteur de recherche sur internet qui dispose d'une base contenant les relations constructionnelles *actif*:*activité* et *activer*:*activable* peut proposer, parmi les réponses à une requête qui inclut *un processus actif*, une page web dans laquelle apparaît le SN *l'activité du processus*. De même une page web contenant *activer le processus* pourrait être proposée en réponse à une requête comprenant *un processus activable*.

2. Apprendre les relations morphographiques en utilisant l'analogie graphémique

Une manière simple d'identifier des couples de lexèmes morphologiquement apparentés consiste à mettre en évidence la structure paradigmatique du lexique écrit ou plus exactement les séries proportionnelles qui la composent. Cette technique est utilisée dans de nombreux travaux sur l'acquisition de connaissances morphologiques parmi lesquels on peut citer [LEP 98, PIR 99, GAU 99, GRA 99, HAT 02]. Les couples de lexèmes qui entrent dans les séries proportionnelles présentent une configuration analogique que l'on peut illustrer par le quadruplet *activer*:*activable*, *agiter*:*agitable*. Si l'on décrit les relations entre les graphies de ces quatre lemmes en termes d'ajout et suppression de préfixes ou de suffixes, *activer* est à *activable* ce que *agiter* est à *agitable*. En effet, la même relation s'établit entre les éléments du premier couple et du second : suppression du suffixe *-er* et ajout du suffixe *-able*. Pa-

rallèlement, la même relation s'établit entre *activer* et *agiter* d'une part, et *activable* et *agitable* d'autre part : suppression du préfixe *activ-* et ajout du préfixe *agit-*.

L'appartenance des deux couples à une même série peut être formalisée comme un partage de signature suffixale définie comme suit. Soit L un alphabet. Pour tout couple de chaînes de caractères $(m_1, m_2) \in L^* \times L^*$, on définit un ensemble de couples de suffixes $S(m_1, m_2) = \{(t_1, t_2) \in L^* \times L^* / \exists r \in L^*, m_1 = r \cdot t_1 \text{ et } m_2 = r \cdot t_2\}$. La signature suffixale de (m_1, m_2), notée $\sigma(m_1, m_2)$, est l'élément $(s_1, s_2) \in S(m_1, m_2)$ pour lequel r est de longueur maximale. Par exemple, S(`laver`, `lavable`) est {(`laver`, `lavable`), (`aver`, `avable`), (`ver`, `vable`), (`er`, `able`)}, r valant respectivement ϵ, `l`, `la` et `lav` ; σ(`laver`, `lavable`) = (`er`, `able`), ce couple de suffixes correspondant au radical r le plus long. La signature d'un couple caractérise la série proportionnelle morphographique à laquelle il appartient (que l'on peut définir comme l'ensemble des couples qui partagent cette signature). Toutes les séries ne sont naturellement pas intéressantes du point de vue morphologique. Un petit nombre de critères permet de caractériser celles qui sont les plus utiles pour constituer un lexique constructionnel ou flexionnel : la taille minimale du radical r des couples qui forment la série ; la taille maximale des suffixes $s_1$et s_2 ; le nombre de couples qui composent la série. Dans les expérimentations décrites dans la section 3.3, les valeurs utilisées pour ces paramètres sont respectivement 3, 8 et 3.

L'identification des séries peut être affinée en prenant en compte les catégories morphosyntaxiques des éléments qui les composent. On peut en effet obtenir des signatures catégorisées en ajoutant à chaque lemme son étiquette morphosyntaxique. Pour le français, nous utilisons le jeu d'étiquettes GRACE [RAJ 97]. Les verbes de la figure 1 reçoivent alors l'étiquette `Vmn----` (par exemple `agiter/Vmn----`) et les adjectifs l'étiquette `Afpms` (par exemple `agitable/Afpms`). La signature du couple de lemmes catégorisés (`agiter/Vmn----`, `agitable/Afpms`) devient ainsi (`er/Vmn----`, `able/Afpms`). La catégorisation des signatures permet de différentier les séries correspondant à des affixes qui ne sont pas catégoriellement homogènes. Elle améliore le filtrage basé sur la taille des séries (nombre de leurs couples) et rend plus sûre l'utilisation des signatures en prédiction.

3. Croiser les contraintes morphographiques et les contraintes sémantiques

Les séries proportionnelles morphographiques mettent en relation les lexèmes uniquement sur la base de leurs graphies. La construction morphologique étant une relation de partage de forme et de sens, les appariements de lexèmes induits par ces séries reposent sur une double approximation : le partage par deux formes d'un radical commun suffisamment long signale un partage (*i*) de propriétés phonologiques et (*ii*) de propriétés sémantiques par les lexèmes correspondants. Si la première approximation est satisfaisante, la seconde est très insuffisante. Or la construction morphologique est d'abord une affaire de sémantique. Il est donc indispensable d'utiliser des connaissances sémantiques dans l'acquisition de la morphologie constructionnelle.

3.1. *Travaux connexes*

Ce problème est traité dans de nombreux travaux en utilisant des corpus textuels ou des bases documentaires pour caractériser sémantiquement les lexèmes à partir des contextes dans lesquels ils apparaissent. C'est le cas de [JAC 97] qui propose une méthode non statistique qui, à partir d'un corpus de textes médicaux en anglais, extrait des bitermes (termes composés de deux mots) morphologiquement apparentés comme *artificial ventilation* et *artificially ventilated*. La cooccurrence au sein de termes constitue une contrainte forte sur le sens des formes appariées. Parmi les méthodes statistiques d'acquisition de la morphologie citons celle de [SCH 00] dans laquelle le sens des formes est représenté par des vecteurs contextuels sémantiques dont la similarité est calculée en utilisant la technique de l'analyse sémantique latente.

La méthode proposée ici est différente puisqu'elle utilise un dictionnaire de synonymes. Ce type de ressources est bien adapté à l'acquisition morphologique car elles décrivent des relations de partage de propriétés sémantiques entre lexèmes. Or c'est justement l'approximation de ces relations par les graphies qui nous pose problème. Une technique similaire est utilisée par [GRA 99]. Elle consiste à exploiter 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 méthode de [GRA 99] diffère cependant de la nôtre par le fait que ces relations ne sont pas filtrées par des contraintes analogiques en raison de la taille réduite du thésaurus initial (5 801 termes).

3.2. *Quadruplets morpho-synonymiques*

Le dictionnaire de synonymes (ou plus exactement de relations de proximité sémantique) est utilisé d'une part comme un lexique de lemmes pour constituer des séries morphographiques et d'autre part comme une source d'informations sémantiques pour filtrer les couples qui composent ces séries. Ce filtrage consiste à former des quadruplets morpho-synonymiques $X_1{:}X_2{::}Y_1{:}Y_2$ tels que $X_1{:}X_2$ et $Y_1{:}Y_2$ sont des couples de lexèmes qui appartiennent à des séries morphographiques, que X_1 est un synonyme de Y_1 et que X_2 est un synonyme de Y_2. Dans ce qui suit, nous appellerons signature du quadruplet $X_1{:}X_2{::}Y_1{:}Y_2$ le quadruplet de suffixes (s_1, s_2, s_3, s_4) tel que $(s_1, s_2) = \sigma(X_1, X_2)$ et $(s_3, s_4) = \sigma(Y_1, Y_2)$. On peut illustrer ces quadruplets par l'exemple en figure 2. Comme le montre ce schéma, les quadruplets morpho-synonymiques sont morphologiques dans l'une de leurs dimensions et synonymiques dans l'autre.

Sur le plan sémantique, la synonymie de X_1 et Y_1 implique qu'ils partagent l'essentiel de leurs propriétés sémantiques (ce que l'on peut noter informellement $\Sigma(X_1) \cong \Sigma(Y_1)$ où $\Sigma(X)$ représente l'ensemble des propriétés sémantiques du lexème X). Si X_1 et X_2 sont effectivement morphologiquement apparentés, alors eux aussi partagent une partie de leurs propriétés sémantiques ($\Sigma(X_1) \cap \Sigma(X_2) \neq \emptyset$). Par transitivité, Y_1 partage avec X_2 à peu près les mêmes propriétés que X_1

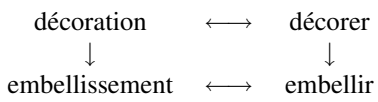

Figure 2. *Exemple de quadruplet morpho-synonymique. Les flèches bi-directionnelles horizontales sont des relations de parenté morphologique prédites qu'il faut filtrer. Les flèches verticales sont des relations de synonymie décrites dans le dictionnaire*

($\Sigma(X_1) \cap \Sigma(X_2) \cong \Sigma(Y_1) \cap \Sigma(X_2)$). Comme X_2 et Y_2 sont synonymes, ils partagent eux aussi la plupart sinon toutes leurs propriétés sémantiques ($\Sigma(X_2) \cong \Sigma(Y_2)$). Par conséquent, Y_1 est Y_2 partagent une partie de leurs propriétés sémantiques ($\Sigma(Y_1) \cap \Sigma(Y_2) \cong \Sigma(Y_1) \cap \Sigma(X_2) \cong \Sigma(X_1) \cap \Sigma(X_2) \neq \emptyset$). Le fait que ces propriétés soient les mêmes que celles que X_1 partage avec X_2 fait que les relations sémantiques qui s'établissent entre X_1 et X_2 d'une part, et d'autre part entre Y_1 et Y_2 sont les mêmes. Par suite X_1:X_2::Y_1:Y_2 forment un quadruplet analogique, c'est-à-dire que « X_1 est à X_2 ce que Y_1 est à Y_2 ». C'est par exemple le cas du quadruplet *décoration*:*décorer*::*embellissement*:*embellir* : la décoration est l'action de décorer exactement comme l'embellissement est l'action d'embellir.

Outre les deux contraintes qui définissent les quadruplets morpho-synonymiques, un filtre supplémentaire est utilisé pour imposer que les graphies des lemmes de X_1 et de X_2 soient différentes et que celles de Y_1 et de Y_2 le soient aussi. Ce filtre élimine par exemple un quadruplet comme *agglutinant/Ncms*:*agglutinatif/Afpms*::*adhésif/Ncms*:*adhésif/Afpms* qui contient une conversion.

3.3. *Expérimentations et évaluation*

Plusieurs expériences d'acquisition de relations de parenté morphologique au moyen de quadruplets morpho-synonymiques ont été réalisées sur des dictionnaires de synonymes en français et en anglais.

3.3.1. *Expérience sur le français*

La première expérience a été réalisée sur le français en utilisant un dictionnaire de synonymes constitué à l'INaLF (CNRS, USR 705 ; aujourd'hui ATILF, UMR 7118, CNRS et Université de Nancy 2) en fusionnant cinq dictionnaires de synonymes (R. BAILLY, H. BÉNAC, H. BERTAUD DU CHAZAUD, M. F. GUIZOT, P.-B. LAFAYE), les synonymes du *Grand Larousse* et les « renvois analogiques » du *Grand Robert*[3]. Ce dictionnaire a également été utilisé par [PLO 98] dans une version corrigée. Les

3. Ce dictionnaire est mis en ligne sur le site de l'ATILF (`http://www.inalf.fr/synonymes/`). Il est également accessible, dans une version corrigée, sur le site du CRISCO (UMR 6170, CNRS et Université de Caen ; `http://www.crisco.unicaen.fr/`)

objectifs de ces auteurs se distinguent des nôtres sur plusieurs points puisqu'ils s'intéressent à la représentation du sens dans un modèle continuiste de la polysémie lexicale. Leur travail porte en particulier sur la caractérisation du sens au moyen de cliques de synonymes et sur la structuration géométrique de l'espace sémantique à l'aide de la distance du χ^2. Leur utilisation du dictionnaire est également différente de la nôtre puisque les synonymes ne sont pas étiquetés et que les relations synonymiques sont rendues symétriques.

Pour l'expérience que nous avons réalisée, le dictionnaire a été formaté en filtrant et en catégorisant morphosyntaxiquement les entrées et les synonymes. Nous avons éliminé les mots grammaticaux, les auxiliaires et les formes multilexicales (locutions et syntagmes). La catégorisation des unités du dictionnaire a été réalisée au moyen du lexique flexionnel TLFnome+index qui a été constitué à l'INaLF à partir de la nomenclature et de l'index du *Trésor de la Langue Française* (*TLF*). Après formatage, le dictionnaire comporte 45 009 lemmes (entrées ou synonymes) et 234 771 couples de synonymes ou plus précisément de lexèmes sémantiquement proches.

Un ensemble de séries proportionnelles morphographiques a été constitué à partir du lexique composé des lemmes du dictionnaire. Les couples de ces séries ont ensuite été réunis en quadruplets en exploitant les relations synonymiques. 43 377 quadruplets ont ainsi été constitués[4]. Ces derniers sont composés de 24 499 couples différents (non orientés) qui ont été évalués en termes de précision et de rappel. La précision a été estimée à 95 % à partir d'un échantillon de 200 couples choisis aléatoirement et révisé par l'auteur. Le rappel a été calculé relativement au lexique Verbaction (6 471 couples *verbe*:*nom* tels que le nom est morphologiquement apparenté au verbe et qu'il dénomme l'action ou l'événement correspondant à ce verbe) et aux familles constructionnelles de la tranche *fr-* de la nomenclature du *TLF* (160 familles regroupant 614 lemmes). Le taux de rappel est de 93,3 % pour Verbaction et de 65,1 % pour les familles de la tranche *fr-* du *TLF*.

3.3.2. *Expérience sur l'anglais*

La deuxième expérience a été réalisée à l'aide de trois dictionnaires anglais extraits de WordNet [MIL 90]. Ces dictionnaires décrivent des relations de proximité sémantique forte (appartenance à un même synset), moyenne (appartenance à un même synset ou à des synsets synonymes ou immédiatement hyperonymes) et faible (appartenance à un même synset ou à des synsets synonymes, immédiatement hyperonymes ou co-hyponymes). Ces trois dictionnaires seront respectivement appelés S-dict, M-dict et L-dict. Leurs entrées sont des lemmes étiquetés par leurs catégories grammaticales, les informations catégorielles ayant été distribuées sur les éléments des synsets. Par ailleurs, les lemmes multilexicaux ont été supprimés. Les deux premières colonnes du tableau 1 présentent le nombre de couples et de lemmes qui composent ces dictionnaires. Le nombre de quadruplets morpho-synonymiques extraits et celui des couples

4. Le nombre de quadruplets effectivement extraits est en fait le double car les relations morphographiques ne sont pas orientées : si X_1:X_2::Y_1:Y_2 est un quadruplet morpho-synonymique alors X_2:X_1::Y_2:Y_1 en est un également.

	dictionnaire		acquisition		évaluation	
	# couples	# lemmes	# quad.	# couples	précis.	rappel
S-dict	127 274	43 055	41 117	24 079	92 %	78,6 %
M-dict	283 422	62 477	77 132	43 349	86 %	77,5 %
L-dict	2 213 331	64 168	334 516	73 376	62 %	72,3 %

Tableau 1. *Taille des dictionnaires anglais. Résultats et évaluation de l'acquisition de couples de lemmes morphologiquement apparentés à partir des dictionnaires anglais*

de lemmes morphologiquement apparentés qui les composent sont indiqués dans les troisième et quatrième colonnes du tableau 1. Celui-ci présente dans ses deux dernières colonnes l'évaluation de ces résultats. Comme pour le dictionnaire français, la précision a été estimée en utilisant des échantillons de 200 couples choisis au hasard et révisés par l'auteur. Le rappel a été calculé relativement à la base CELEX [BAA 95] dont nous avons extrait les couples de lemmes qui sont en relation de suffixation.

On observe ainsi que l'acquisition de couples de lemmes morphologiquement apparentés à partir de dictionnaires de synonymes est une méthode robuste : la précision diminue avec le relâchement des relations de proximité sémantique sans pour autant s'effondrer, alors que le rappel est peu affecté par la dégradation de la description sémantique.

4. Typer et croiser les contraintes morpho-synonymiques

Les quadruplets morpho-synonymiques peuvent être constitués soit de couples de lemmes appartenant à une même série morphographique comme *déchirer/Vmn----:déchirure/Ncfs*::*érafler/Vmn----:éraflure/Ncfs* dont la signature est ((er/Vmn----, ure/Ncfs), (er/Vmn----, ure/Ncfs)), soit de couples appartenant à des séries différentes comme *déchirer/Vmn----:déchirure/Ncfs*::*fendre/Vmn----:fente/Ncfs* dont la signature est ((er/Vmn----, ure/Ncfs), (dre/Vmn----, te/Ncfs)). Dans le premier cas, les signatures des deux couples sont identiques. Le quadruplet est analogique à la fois sur le plan morphographique et sur le plan sémantique. Dans le second cas, les couples ont des signatures différentes et le quadruplet est morphographiquement hétérogène. Nous appellerons « exactes » les signatures des quadruplets du premier type et « hétérogènes » celles des quadruplets du second type. Le typage des signatures des quadruplets induit un typage des signatures des couples qui les composent. On dira qu'une signature de couple est de type E si elle apparaît dans une signature de quadruplet exact. Une signature de couple qui n'apparaît que dans des quadruplets hétérogènes sera dite de type H. On peut ensuite distinguer au moins trois groupes parmi les signatures de quadruplets hétérogènes : H0 seront celles qui sont composées de deux signatures de couple E, H1 celles qui sont composées d'une signature de couple E et d'une signature de couple H et H2 celles qui sont composée de deux signatures de couple H. Par abus de langage, on dira qu'un quadruplet est exact (resp. hétérogène,

	français				S-dict			
type	#sign.	#quad.	moy.	#cpl	#sign.	#quad.	moy.	#cpl
E	567	8 160	14,4	7 385	825	8 014	9,7	6 932
H	18 951	35 217	1,9	21 280	16 000	33 103	2,1	19 960
H0	2 231	9 183	4,1	6 652	1 640	4 924	3,0	3 734
H1	9 117	17 222	1,9	13 104	6 962	16 838	2,4	12 712
H2	7 603	8 812	1,2	9 341	7 398	11 341	1,5	9 123
	M-dict				L-dict			
type	#sign.	#quad.	moy.	#cpl	#sign.	#quad.	moy.	#cpl
E	1 300	12 425	9,6	12 280	2 095	38 058	18,2	18 006
H	34 623	64 707	1,9	38 211	158 797	296 458	1,9	68 000
H0	3 720	12 009	3,2	10 500	11 828	50 367	4,6	18 635
H1	15 929	33 393	2,1	24 846	64 981	151 107	2,3	46 546
H2	14 974	19 305	1,3	16 622	81 988	94 984	1,2	40 933

Tableau 2. *Répartition des signatures et des quadruplets par types. La troisième colonne de chaque groupe donne le nombre moyen de quadruplets par signature. La quatrième indique le nombre de couples différents qui composent les quadruplets*

H0, H1 ou H2) si sa signature est exacte (resp. hétérogène, H0, H1 ou H2) et qu'un couple est exact (resp. hétérogène, H0, H1 ou H2) s'il appartient à un quadruplet exact (resp. hétérogène, H0, H1 ou H2). On peut de plus distinguer parmi les couples H1 ceux dont la signature est de type E que l'on notera H1-E et ceux qui ont une signature H que l'on notera H1-H.

Il existe une corrélation nette entre les types de signatures et leurs nombres moyens d'instances. On observe ainsi dans le tableau 2 que les signatures E ont les fréquences moyennes les plus élevées, de 5 à 10 fois plus importantes que celles des signatures H. De même, les fréquences moyennes des signatures H0 sont toujours supérieures à celles des signatures H1 qui le sont à leur tour par rapport à celles des signatures H2. Ces rapports sont une conséquence du fait que le lexique construit est structuré par les paradigmes. Les quadruplets E forment des paradigmes morphologiques et sont de ce fait les plus importants. Viennent ensuite les quadruplets H0 qui correspondent à des relations entre ces paradigmes, puis les quadruplets H1 qui relient des couples isolés à des couples appartenant à des paradigmes. Les quadruplets H2 qui ne relient que des couples isolés, contribuent le moins à la structuration du lexique.

Le typage des quadruplets et des couples présente également un intérêt pratique comme le montre l'évaluation séparée des couples de chaque type présentée dans le tableau 3. La précision et le rappel ont été calculés comme indiqué en §3.3.1 et §3.3.2. On observe que globalement, les couples H0 et H1-E comportent le moins d'erreurs tandis que les couples H2 en concentrent le plus. Le typage constitue donc un moyen effectif de filtrer les couples de lexèmes morphologiquement apparentés au moyen de critères qualitatifs.

	français				S-dict		
type	#cpl	précision	rappel V	rappel *fr-*	#cpl	précision	rappel
E	7 385	**98,0 %**	97,7 %	73,3 %	6 932	94,5 %	92,1 %
H0	6 652	**98,5 %**	96,5 %	76,3 %	3 734	**95,5 %**	87,8 %
H1-E	10 822	**99,0 %**	94,9 %	59,7 %	9 398	**99,5 %**	80,2 %
H1-H	14 876	78,5 %	84,5 %	59,1 %	10 539	**96,0 %**	77,5 %
H2	9 341	81,0 %	76,8 %	52,9 %	9 123	85,0 %	65,9 %

	M-dict			L-dict		
type	#cpl	précision	rappel	#cpl	précision	rappel
E	12 280	93,0 %	89,2 %	18 006	81,0 %	84,3 %
H0	10 500	**97,5 %**	83,7 %	18 635	90,5 %	78,4 %
H1-E	16 383	**97,0 %**	79,2 %	27 177	82,5 %	73,7 %
H1-H	17 868	91,0 %	72,3 %	35 275	68,5 %	62,9 %
H2	16 622	72,5 %	55,2 %	40 933	46,0 %	49,7 %

Tableau 3. *Évaluation des couples pour chaque type de quadruplets. La première colonne indique le nombre de couples de chaque type. Les précisions supérieures à 95 % sont signalées en gras*

De plus, L'évaluation permet de constater que pour les quatre dictionnaires, la précision des couples H0 et H1-E est supérieure à celle des couples E. Ce résultat est à première vue étonnant car les couples E font partie de quadruplets qui vérifient une contrainte plus stricte que les autres. On s'attendrait par conséquent à ce qu'ils aient une meilleure précision. Dans les faits, la diversité des contraintes semble plus déterminante que leurs forces que l'on peut estimer par la taille des paradigmes. Ainsi, les couples d'un quadruplet exact ne sont contraints que dans un paradigme alors que ceux d'un quadruplet H0 appartiennent à deux paradigmes distincts composés chacun d'au moins deux couples. À cela s'ajoute la contrainte correspondant au quadruplet lui-même, à savoir que les couples sont synonymes. La situation est similaire pour les couples H1-E même si l'autre couple du quadruplet est isolé.

5. Conclusion

La méthode d'acquisition de connaissances morphologiques constructionnelles présentée dans cet article exploite de manière systématique la structure paradigmatique du lexique en s'appuyant sur l'analogie morphographique et sur la synonymie qui n'est autre qu'une analogie sémantique. La méthode repose sur l'identification de quadruplets morpho-synonymiques qui permettent de croiser des contraintes sémantiques définies au moyen de relations synonymiques et des contraintes morphographiques. Elle est robuste et indépendante vis-à-vis des langues particulières : elle a été utilisée avec succès pour un dictionnaire de synonymes français et trois dictionnaires

anglais composés de descriptions de proximité sémantique de plus en plus dégradées. Nous avons en outre proposé un typage des quadruplets morpho-synonymiques et des couples qui les composent permettant d'expliciter le fait que certains couples de lexèmes sont soumis à un plus grand nombre de contraintes que d'autres. L'évaluation des différents types de couples indique que la diversité des contraintes est plus déterminante que leurs forces.

Croiser des contraintes différentes permet donc d'améliorer l'efficacité de l'acquisition de connaissances morphologiques. Cette technique peut être facilement étendue en utilisant des couples de lexèmes morphologiquement apparentés, acquis à partir d'autres ressources comme des dictionnaires électroniques (en utilisant par exemple une méthode adaptée de [GAU 02]), des dictionnaires bilingues, des corpus, des bi-textes [KRA 01], etc. De par leur mode d'acquisition, ces couples vérifient des contraintes de natures diverses et sont parfaitement indiqués pour constituer d'autres types de quadruplets analogiques.

6. Bibliographie

[BAA 95] BAAYEN R. H., PIEPENBROCK R., GULIKERS L., « The CELEX Lexical Database (Release 2) », CD-ROM, 1995, Linguistic Data Consortium, University of Pennsylvania, Pennsylvania, USA.

[BYB 88] BYBEE J. L., « Morphology as Lexical Organization », HAMMOND M., NOONAN M., Eds., *Theoretical Morphology. Approaches in Modern Linguistics*, chapitre 7, p. 119-141, Academic Press, San Diego, CA, 1988.

[BYB 95] BYBEE J. L., « Regular Morphology and the Lexicon », *Language and cognitive processes*, vol. 10, n° 5, 1995, p. 425-455.

[COR 01] CORBIN D., « Préfixe et suffixes : du sens aux catégories », *Journal of French Language Studies*, vol. 11, n° 1, 2001, p. 41-69.

[CRU 86] CRUSE D. A., *Lexical Semantics*, Cambridge University Press, Cambridge, UK, 1986.

[GAU 99] GAUSSIER É., « Unsupervised Learning of Derivational Morphology from Inflectional Lexicons », *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*, University of Mariland, USA, 1999, Association for Computational Linguistics, ACL'99.

[GAU 02] GAUME B., DUVIGNEAU K., GASQUET O., GINESTE M.-D., « Forms of Meaning, Meaning of Forms », *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 14, n° 1, 2002, p. 61-74.

[GRA 99] GRABAR N., ZWEIGENBAUM P., « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical », *Actes de la 6e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-99)*, Cargèse, juillet 1999, p. 175-184.

[HAT 02] HATHOUT N., NAMER F., DAL G., « An Experimental Constructional Database: The MorTAL Project », BOUCHER P., Ed., *Many Morphologies*, p. 178-209, Cascadilla, Somerville, Mass., 2002.

[JAC 97] JACQUEMIN C., « Guessing Morphology from Terms and Corpora », *Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, Philadelphia, PA., 1997, ACM, p. 156-167.

[JAC 99] JACQUEMIN C., TZOUKERMANN E., « NLP for term variant extraction: synergy between morphology, lexicon, ans syntax », STRZALKOWSKI T., Ed., *Natural Language Information Retrieval*, p. 25-74, Kluwer Academic Publishers, Dordrecht, 1999.

[KRA 01] KRAIF O., « Exploitation des cognats pour l'alignement. Architecture et évaluation », *Traitement automatique des langues*, vol. 42, n° 3, 2001, p. 833-867.

[LEP 98] LEPAGE Y., « Solving analogies on words: an algorithm », *Proceedings of the of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 2, Montréal, Canada, août 1998, p. 728-735.

[MIL 90] MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K. J., « Introduction to Wordnet: An On-line Lexical Database », *International Journal of Lexicography*, vol. 3, n° 4, 1990, p. 335-391, Oxford University Press.

[PIR 99] PIRRELLI V., YVON F., « The hidden dimension: a paradigmatic view of data-driven NLP », *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 11, n° 3, 1999, p. 391-408.

[PLO 98] PLOUX S., VICTORRI B., « Constructions d'espaces sémantiques à l'aide de dictionnaires de synonymes », *T.A.L.*, vol. 39, n° 1, 1998, p. 161-182.

[RAJ 97] RAJMAN M., LECOMTE J., PAROUBEK P., « Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique », Rapport grace gtr-3-2.1, 1997, EPFL & INaLF.

[SCH 00] SCHONE P., JURAFSKY D. S., « Knowledge-Free Induction of Morphology Using Latent Semantic Analysis », *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, Lisbon, Portugal, 2000, p. 67-72.