Exploiter la structure analogique du lexique construit : une approche computationnelle

1. Introduction

L'analogie joue un rôle central dans la structuration du lexique, en particulier sur le plan morphologique. La morphologie flexionnelle définit dans le lexique des paradigmes qui peuvent être étendus par analogie pour produire les formes fléchies des néologismes. Le même type d'organisation existe sur le plan constructionnel[1] puisque les affixes forment eux aussi des paradigmes. La structure analogique du lexique construit[2] peut être exploitée pour acquérir des connaissances morphologiques à partir de collections de données telles que des corpus de textes ou des dictionnaires. Ces connaissances permettent de constituer des ressources constructionnelles, et en particulier des bases de données destinées à la psycholinguistique, au traitement automatique des langues (TAL) et à la recherche d'information (RI).

La visée de notre travail est essentiellement appliquée. L'approche générale est compatible avec le modèle en réseau proposé par J. BYBEE (1988, 1995). 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[3] qui partagent en même temps des propriétés phonologiques et des propriétés sémantiques. Les affixes constructionnels sont vus comme des sous-graphes qui forment des séries proportionnelles[4] (CRUSE 1986, p. 118 et suivantes) et que l'on peut étendre par analogie.

[INSÉRER ICI LA FIGURE 1]

On peut ainsi représenter graphiquement une portion du sous-graphe qui correspond au suffixe *-able* comme en figure 1. Les lignes fines représentent un partage de son et les lignes épaisses un partage de sens. Dans le schéma, les arcs qui relient *contrôlable* à *contrôler* rendent compte du partage des propriétés phonologiques et sémantiques qui correspondent au radical de l'adjectif qui se trouve aussi être celui du verbe. De même, les arcs qui relient *contrôlable* à *agitable*, *gonflable*,

---

1 Nous adopterons ici la terminologie de D. CORBIN (2001) 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 ne présuppose pas l'existence de règles, de niveaux de dérivation...

2 Nous désignons sous ce terme l'ensemble des lexèmes construits.

3 Les lexèmes peuvent être définis comme des sous-graphes constitués de formes qui ne diffèrent que par leurs marques flexionnelles.

4 Une série proportionnelle est un ensemble de couples $(x_i, y_i)\ 1 \le i \le n$ tels que :
$\forall j, 1 \le j \le n, \forall k, 1 \le k \le n, P(x_j, y_j, x_k, y_k) \wedge P(y_j, x_j, y_k, x_k) \wedge P(x_j, x_k, y_j, y_k) \wedge P(x_k, x_j, y_k, y_j)$
où $P(a,b,c,d)$ ssi *a* est à *b* ce que *c* est à *d*.

activable... correspondent au partage de son et de sens qui peut être associé au suffixe *-able*.

1.1. Bases de données morphologiques

Ce travail s'inscrit dans le cadre du projet MorTAL[5] (acronyme de : « MORphologie pour le TAL » ; (DAL *et al.* 1999 ; HATHOUT *et al.* 2002)). Son objectif général est de constituer de manière semi-automatique une base de données constructionnelles à large couverture pour le français. La base MorTAL est composée de deux parties. La première, basée sur le modèle SILEX, est réalisée au moyen de l'analyseur DériF (« Dérivation en Français »). Elle décrit actuellement les lexèmes construits par les affixes *-able*, *-ité*, *-et*(*te*) *-is*(*er*), *-ifi*(*er*), *re-*, *in-* et *dé-*. DériF produit des analyses morphologiques très fines à partir de règles et de fichiers d'exceptions mis au point manuellement par ses conceptrices (DAL et NAMER 2000). En contrepartie, le nombre de lexèmes traités est assez faible. La seconde partie de la base MorTAL est construite par le système DéCor (« Dérivation pour les Corpus ») présenté en §2. À la différence de DériF, DéCor a été conçu en privilégiant la couverture du lexique construit plutôt que la finesse et à la précision des analyses.

Les bases de données morphologiques sont essentiellement utilisées en psycholinguistique[6], en TAL et en RI (JACQUEMIN et TZOUKERMANN 1999 ; FABRE et JACQUEMIN 2000 ; DAL *et al.* 2004). Elles peuvent par exemple être exploitées pour identifier des variantes morphosyntaxiques dans les documents (JACQUEMIN 1997b). Ainsi, un moteur de recherche sur internet qui utiliserait une base contenant les relations constructionnelles *actif* :*activité* et *activer* :*activable* pourrait 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 un processus* pourrait être proposée en réponse à une requête comprenant *un processus activable*.

1.2 Séparation des outils et des ressources

L'un des principes généraux de ce travail en morphologie computationnelle est de ne pas inclure de connaissances linguistiques « explicites » dans les outils mais d'utiliser des ressources externes (dictionnaires, corpus...) et les informations fournies par l'utilisateur lors de l'exécution des programmes. Les objectifs de cette séparation des outils et des connaissances sont multiples.

1. Les outils construits selon ce principe sont indépendants vis-à-vis des langues particulières.
2. En s'imposant d'utiliser des ressources suffisamment générales, on permet à d'autres d'employer les mêmes techniques[7].

---

5 `http://www.univ-lille3.fr/www/Recherche/silex/mortal/`

6 Par exemple, J. HAY (2000) s'appuie sur la base CELEX (BAAYEN *et al.* 1995) pour construire des expériences visant à déterminer l'incidence de la fréquence lexicale sur la décomposition des mots complexes et sur leur représentation à long terme.

7 On garantit aussi, dans une certaine mesure, la reproductibilité des expériences même si la

3. Les méthodes qui incluent peu ou pas de connaissances linguistiques sont plus faciles à développer et à mettre en œuvre que celles qui en incluent. La tâche la plus difficile dans le développement d'outils de TAL basées sur des descriptions linguistiques est en effet l'explicitation et la formalisation des connaissances linguistiques. Des problèmes de cohérence peuvent également se poser de façon critique au fur et à mesure que la couverture du système s'élargit et que l'ensemble des connaissances grossit.

La séparation des outils et des ressources conduit à privilégier la couverture à la précision des traitements. Les méthodes conformes à ce principe ne peuvent donc être que semi-automatiques. Cette caractéristique est en réalité un avantage : il est plus aisé d'avoir recours à la compétence de personnes chargées de la révision de ressources construites par programme car elles n'ont pas à expliciter leurs intuitions. D'autre part, aucune compétence informatique n'est requise pour les tâches de révision. Il n'est pas non plus nécessaire de maîtriser parfaitement une ou plusieurs théories linguistiques pour décider de la validité de constructions ou de relations. Les intuitions de locuteurs natifs suffisent amplement. Les méthodes semi-automatiques permettent donc de construire des ressources linguistiques de manière plus économique.

1.3 Analogie et lexique construit

Nous nous intéressons dans ce travail à deux types d'exploitation de la structure analogique du lexique construit. La première consiste à identifier les paradigmes définis par les affixes en s'appuyant uniquement sur les formes graphémiques présentes dans un lexique flexionnel. Elle est détaillée en section 2. Nous présentons également en §2.3 une méthode qui utilise les connaissances morphologiques acquises à partir d'un lexique pour constituer des familles constructionnelles. Le deuxième type d'exploitation, auquel est consacrée la section 3, repose sur une technique permettant d'améliorer la qualité des ressources produites en croisant les connaissances morphologiques avec des informations sémantiques issues de dictionnaires de synonymes.

2 Analogie graphémique

La structure analogique la plus simple que l'on peut exploiter pour l'acquisition de connaissances morphologiques constructionnelle est l'analogie graphémique. Elle a été utilisée dans de nombreux travaux portant sur ce thème parmi lesquels on peut citer (LEPAGE et SHIN-ICHI 1996 ; LEPAGE 1998 ; PIRRELLI et YVON 1999 ; GAUSSIER 1999 ; DAL *et al.* 1999 ; GRABAR et ZWEIGENBAUM 1999 ; HATHOUT 2000 ; NEUVEL et FULOP 2002 ; HATHOUT *et al.* 2002). On peut illustrer cette structure analogique en considérant les couples *activer*:*activable* et *agiter*:*agitable*. Les relations entre ces quatre formes peuvent être décrites en termes d'ajout et de suppression de préfixes ou de suffixes graphémiques. Graphémiquement, *activer* est à *activable* ce que *agiter* est à *agitable*. En effet, la même relation s'établit entre les

---

communauté des morphologues informaticiens est relativement petite et que la reproduction et la vérification des expériences ne sont pas des pratiques courantes dans la communauté TAL (en particulier parce que les moyens humains et matériels des groupes de recherche sont trop faibles et que la démarche expérimentale n'y est pas encore bien établie).

éléments du premier couple et du second : suppression du suffixe *-er* et ajout du suffixe *-able*[8]. Parallèlement, 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-*. Ces relations peuvent être représentées schématiquement comme en figure 2. Ce quadruplet de formes fait partie d'une série proportionnelle qui inclut d'autres couples : *contrôler*:*contrôlable* ; *gonfler*:*gonflable* ; *activer*:*activable* ; *négliger*:*négligeable*... Cette structure analogique est par exemple exploitée par le correcteur orthographique ispell[9] lors de l'analyse et du stockage des mots absents de son dictionnaire. Ce programme est ainsi capable de prédire qu'une forme comme *recyclable* peut entrer dans la série de la figure 1 si la forme *recycler* appartient à son dictionnaire.

[INSÉRER ICI LA FIGURE 2]

2.1 Apprentissage de schémas d'affixation graphémique

Les structures analogiques peuvent être utilisées pour acquérir automatiquement des schémas d'affixation graphémique[10] à partir d'un lexique. L'apprentissage repose sur l'hypothèse que les lemmes des lexèmes suffixés sont de la forme *radical*×*suffixe* et que ceux des lexèmes préfixés sont de la forme *préfixe*×*radical*[11]. On peut ainsi abstraire à partir de la série représentée en figure 1 un schéma de suffixation er:able qui permet d'apparier un sous-ensemble des verbes du premier groupe avec les adjectifs en *-able* correspondants (lorsqu'ils existent). L'apprentissage repose sur la technique mise en œuvre dans le programme findaffix[12]. Elle consiste à former l'ensemble des couples $X:Y$ tels que X est la graphie d'un lemme adjectival en *-able* et Y est la graphie d'un lemme verbal. Chaque couple $X:Y$ définit un schéma de suffixation $X':Y'$ tels que $X = Z \times X'$ et $Y = Z \times Y'$ où Z est le préfixe graphémique maximal commun à X et. (Par exemple, pour X = `agiter` et Y = `agitable`, le préfixe commun maximal est Z = `agit` ; il en suit que X'= `er` et Y'= `able`. Le schéma de suffixation correspondant à *agiter*:*agitable* est donc $X':Y'$= `er:able`.) Le schéma $X':Y'$ permet de calculer Y à partir de X en supprimant de ce dernier le suffixe X' puis en lui ajoutant le suffixe Y'. Tous les schémas obtenus ne sont naturellement pas conservés. Plusieurs paramètres peuvent être utilisés pour contrôler cet apprentissage, dont : la taille minimale du radical Z ; la taille maximale des affixes X' et Y'; le nombre minimum de couples $X:Y$ que le schéma permet de connecter. Les valeurs de ces paramètres sont définies par l'utilisateur en fonction de ses objectifs.

---

8 Cette analogie peut également être interprétée à un niveau morphologique et sémantique mais cela implique plusieurs approximations (*cf.* §3).

9 `http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html`

10 Un schéma d'affixation graphémique est un couple de relations de suppression et d'ajout d'affixes.

11 *Préfixe*, *radical* et *suffixe* sont ici des chaînes de caractères arbitraires qui ne correspondent pas nécessairement à des préfixes, des radicaux ou des suffixes linguistiques. Le point est l'opérateur de concaténation.

12 Ce script fait partie du correcteur orthographique ispell.

[INSÉRER LA FIGURE 3 ICI]

Les schémas d'affixation peuvent être affinés pour prendre en compte les catégories morphosyntaxiques des lemmes qu'ils permettent d'apparier. Les lemmes de la figure 1 peuvent par exemple être catégorisés à l'aide du jeu d'étiquette GRACE (RAJMAN *et al.* 1997) comme en figure 3. Le schéma de suffixation correspondant peut alors être représenté sous la forme `er/Vmn----:able/Afpms`. Il permet d'apparier les lemmes X et Y si :

- X porte l'étiquette *Vmn----* ;
- Y est étiqueté *Afpms* ;
- la graphie de Y peut être formée à partir de celle de X en lui retirant le suffixe graphémique *-er* puis en lui ajoutant le suffixe graphémique *-able*.

On peut alors acquérir en même temps les schémas de suffixation pour les adjectifs en *-able* formés sur des verbes (comme *contrôlable*:*contrôler* ou *gonflable*:*gonfler*) et pour ceux qui le sont sur des noms comme *charitable*, *ministrable* ou *cyclable* construits respectivement sur *charité*, *ministre* et *cycle* (PLÉNAT 1988). Les schémas catégorisés permettent ainsi de traiter de façon unifiée les affixes dont les lexèmes-bases ne sont pas catégoriellement homogènes.

Un deuxième affinement possible consiste à exploiter au niveau constructionnel l'ensemble des radicaux des lexèmes qui en ont plusieurs (comme par exemple, les verbes du 3e groupe) ainsi que les allomorphies et allographies flexionnelles. Il suffit pour cela que les schémas d'affixation soient appris à partir de formes fléchies et non pas seulement à partir des lemmes. On peut par exemple identifier la base des adjectifs *observable/Afpms*, *envisageable/Afpms* et *croyable/Afpms* à partir des formes *observant/Vmpp---*, *envisageant/Vmpp---* et *croyant/Vmpp---* (participes présents) des verbes *observer/Vmn----*, *envisager/Vmn----* et *croire/Vmn----* en utilisant un seul schéma, `nt/Vmpp---:ble/Afpms`, au lieu de trois : `er/Vmn----:able/Afpms`, `r/Vmn----:able/Afpms`, `ire/Vmn----:yable/Afpms`.

2.2 Appariement morphographique

Les schémas d'affixation graphémique permettent d'apparier des lexèmes-construits à leur lexèmes-bases. C'est dans cette perspective qu'a été développé le système DéCor (DAL *et al.* 1999 ; HATHOUT 2000 ; HATHOUT *et al.* 2002). L'appariement est réalisé en quatre étapes :

1. On extrait d'un lexique de référence (par exemple, de TLFnome+index[13]), un

13 TLFnome+index est un lexique de formes fléchies munies de catégories morphosyntaxiques, qui a été construit à l'INaLF (CNRS, USR 705 ; aujourd'hui ATILF, UMR 7118, CNRS, Université de Nancy 2 et Université Henry Poincaré). Il est composé de deux parties : TLFnome96 et TLFindex99. TLFnome96 comporte 556 689 entrées. Il a été réalisé par J. MAUCOURT et M. PAPIN à partir de la nomenclature du *Trésor de la Langue Française* (*T.L.F.*). TLFindex99 a été construit par A. BERCHE, F. MOUGIN et N. HATHOUT à partir de l'index du *T.L.F.* et contient 178 588 entrées.

ensemble *C* de lemmes[14] susceptibles d'être construits par un affixe donné. Cette sélection est basée sur les suffixes (ou les préfixes) graphémiques des lemmes et sur les catégories morphosyntaxiques. Par exemple, si l'on s'intéresse à la suffixation en *-tion*, *C* sera l'ensemble des lemmes de substantifs qui finissent en *-tion*. On extrait d'autre part du lexique de référence un ensemble *B* de lemmes de lexèmes-bases potentiels en fonction de leur catégorie morphosyntaxique. Par exemple, pour le suffixe *-tion*, *B* sera l'ensemble des lemmes de verbes. L'hypothèse est que les lexèmes de *C*[15] sont construits au moyen de l'affixe considéré et que c'est parmi les lexèmes de *B* que se trouvent leurs bases.

2. Le programme trouvaffix effectue l'apprentissage des schémas de préfixation et/ou de suffixation à partir de *C* et de *B* en utilisant la technique présentée en §2.1.

3. Le programme applicaffix applique les schémas de préfixation et/ou de suffixation aux lexèmes de *C* et de *B*. Il s'agit de construire un graphe bi-parties dont les sommets sont les éléments de *C* et de *B*.

4. Le programme apparibase filtre les arcs du graphe en utilisant des critères statistiques simples. DéCor associe à chaque lexème-construit un lexème-base au plus. Le filtrage sélectionne, pour chaque élément de *C*, le lexème de *B* qui a les plus grandes chances d'en être la base. La fréquence constitue le meilleur critère de sélection de ces lexèmes.

DéCor a été testé pour différents suffixes, dont *-able*, *-ité* et *-iser* qui sont aussi traités par le système DériF. Les résultats de ce dernier étant validés par ses conceptrices nous les avons utilisés comme références pour calculer la précision de DéCor[16]. Les valeurs obtenues pour ces suffixes sont respectivement de 94% pour *-able*, 89% pour *-ité* et 20% pour *-iser*. On observe ainsi que la qualité de l'appariement graphémique dépend essentiellement de la différentiation des formes des différentes catégories de bases potentielles. Les résultats sont excellents pour le suffixe *-able* dont les bases sont en très grande majorité des verbes et du fait de la bonne différentiation des formes verbales et nominales. Ils sont moins bons pour *-ité* car même si ses bases sont en majorité des adjectifs, les formes de ces derniers sont assez proches des formes

---

14 Il s'agit ici d'un abus de langage, *C* est en fait un ensemble de graphies de lemmes. Il en va de même pour l'ensemble *B*.

15 Par abus de langage, les lexèmes dont les lemmes appartiennent à *C* sont appelés *lexèmes de C*. *Idem* pour *lexèmes de B*.

16 Rappelons que si les appariements morphologiques de DéCor ne s'appuient sur aucune théorie particulière, ce n'est pas le cas des analyses de DériF qui sont basées sur le modèle SILEX. Un certain nombre de différences existent entre les analyses des deux systèmes. Beaucoup sont dues aux divergences théoriques des deux approches ; d'autres ont pour origine le fait que les deux systèmes sont déterministes, c'est-à-dire qu'ils ne proposent qu'un lexème-base pour chaque lexème candidat. Certains lexèmes construits peuvent cependant avoir plusieurs lexèmes bases lorsqu'ils entrent dans plusieurs des paradigmes définis par l'affixe auquel on s'intéresse. Dans ce cas, il arrive que DéCor et DériF fassent des choix différents, mais que les deux solutions soient correctes.

nominales. Enfin, ils sont très insuffisants pour *-iser* du fait de la répartition plus équilibrée entre les bases adjectivales et nominales que l'on ne peut départager efficacement sans connaissances sémantiques[17].

L'évaluation montre que DéCor apporte une aide réelle pour l'analyse de nombreux suffixes, mais qu'il ne constitue pas un système « ultime » ; il n'est pas destiné à l'analyse la plus fine des lexèmes-construits. En pratique, on obtient des résultats satisfaisants lorsque l'apprentissage est réalisé avec un seuil de fréquence égal à 10% du nombre des lexèmes candidats.

2.3 Classification en familles constructionnelles

Les limites de DéCor ont deux origines. La première est que la méthode utilisée identifie d'abord des relations de parenté morphologique[18] même si elle permet aussi de sélectionner les bases des lexèmes-construits. Le second est que l'appariement est basé exclusivement sur les graphies et que les approximations sous-jacentes sont trop grossières, surtout sur le plan sémantique.

Une solution au premier problème consiste à ne plus chercher à identifier des lexèmes-bases, mais à constituer des familles constructionnelles. En d'autres termes, cela revient à remplacer à l'étape 4 le filtrage par une classification dont le but est de partitionner le lexique en familles.

Un outil DéClique a été développé pour réaliser cette classification. Il reprend les trois premières étapes de DéCor qui permettent de constituer un graphe morphologique qu'il s'agit de découper en sous-graphes connexes de sorte que chacune de ces parties soit une famille constructionnelle. Le découpage repose sur l'hypothèse que les éléments d'une famille sont tous morphologiquement apparentés et donc connectés deux à deux. En d'autres termes, les familles sont des cliques c'est-à-dire des sous-graphes complets. La figure 4 présentent deux exemples de familles constructionnelles.

[INSÉRER LA FIGURE 4 ICI]

L'algorithme de classification est un algorithme glouton contrôlé par deux paramètres : l'ordre dans lequel les sommets sont classés et la distance utilisée pour sélectionner la classe de rattachement des sommets qui appartiennent à plusieurs cliques. Les meilleurs résultats ont été obtenus (1) en ordonnant les lexèmes en fonction de la taille du plus long suffixe que l'une de leurs formes partage avec une forme d'un autre élément du lexique et (2) en basant le rattachement des sommets aux sous-graphes sur la somme des fréquences des schémas[19] de suffixation qui les relient.

---

17 DéCor n'est pas capable d'identifier les régularités qui permettraient de traiter correctement à la fois *automatiser* (formé sur *automatique* et non *automate*) et *catégoriser* (construit sur *catégorie* et non pas sur *catégorique*).

18 Dans ce qui suit, ces relations seront aussi appelées *liens morphologiques*.

19 La fréquence d'un schéma est le nombre de couples d'éléménts du lexique d'apprentissage qu'il permet de connecter.

DéClique a été évalué en comparant ses résultats avec une classification manuelle de référence[20]. L'évaluation repose sur la méthode suivante qui permet de déterminer le rappel et la précision :

1. identifier, pour chaque classe candidate, la classe de référence avec laquelle elle partage une intersection maximale ;
2. considérer comme des silences, l'ensemble des nœuds manquants ;
3. considérer comme du bruit, l'ensemble des nœuds supplémentaires.

Le rappel et la précision ont ainsi pu être estimés sur la tranche *fr-* de la nomenclature du *T.L.F.* respectivement à 60% et 80%. SCHONE et JURAFSKY (2000, 2001) utilisent la même méthode pour comparer leurs résultats avec ceux de *Linguistica* (GOLDSMITH 2001) et avec la base CELEX.

GAUSSIER (1999) propose une méthode d'acquisition automatique de la morphologie constructionnelle à partir de lexiques flexionnels qui repose sur une classification en familles similaire à la nôtre[21]. Les paramètres de l'apprentissage des schémas de suffixation sont différents des nôtres : la taille minimale du radical est 5 au lieu de 3 et la fréquence minimale des schémas est 2 au lieu de 10. L'algorithme de classification est également différent puisqu'il s'agit d'une méthode de groupement agglomératif hiérarchique. Cependant la méthode utilisée pour déterminer la proximité entre les groupes (*clusters*) contraint ces derniers à être des cliques. L'évaluation des résultats est basée sur une méthode légèrement plus stricte que celles que nous avons utilisée puisque un mot *m* n'est considéré comme bien classé que s'il appartient à une classe *c* dont la majorité des éléments font partie de la famille constructionnelle de *m* et si la majorité des éléments de la famille de *m* appartiennent à *c*. Les résultats rapportés sont meilleurs que les nôtres (la précision est de 85%) mais rien n'est dit sur les lexiques flexionnel et constructionnel utilisés. É. GAUSSIER propose d'autre part une méthode statistique d'extraction de suffixes (linguistiques) et de relations de suffixation à partir des familles relationnelles.

3 Analogie morpho-synonymique

Le second handicap de DéCor vient du fait que l'appariement des lexèmes repose uniquement sur les graphies. Ce problème est plus sérieux que le précédent dans la mesure où la construction morphologique est avant tout une affaire de sémantique et qu'on ne peut améliorer significativement les performances des systèmes d'acquisition de la morphologie constructionnelle sans leur fournir de connaissances sémantiques. En effet, le partage d'un radical commun suffisamment long par deux formes est en général une bonne approximation d'un partage de propriétés phonologiques ; c'est en revanche, une approximation très insuffisante d'un partage de propriétés sémantiques par les lexèmes correspondants.

---

20 L'auteur remercie J. LECOMTE (CNRS, ATILF) pour son aide dans la révision de cette classification.

21 Dans la terminologie de É. GAUSSIER, ces familles sont dites *relationnelles*. Il réserve le terme de *familles dérivationnelles* aux familles constituées à partir du lexique de référence utilisé pour évaluer la classification.

3.1 Acquisition de connaissances morphologiques à partir de corpus

Ce problème a été 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. La quasi-totalité des méthodes proposées sont statistiques. Ce n'est cependant pas le cas du travail de JACQUEMIN (1997a) qui utilise un corpus de textes médicaux en anglais pour extraire automatiquement des bi-termes (termes composés de deux mots) morphologiquement apparentés comme *artificial ventilation* et *artificially ventilated*. La co-occurrence au sein de termes constitue une contrainte forte sur le sens de formes appariées. Le travail présenté ici est fortement inspiré par celui de Ch. JACQUEMIN.

Les méthodes statistiques d'acquisition de la morphologie à partir de corpus sont nombreuses et variées. Ainsi, YARKOWSKY et WICENTOWSKI (2000) proposent une méthode quasi-automatique d'analyse morphologique flexionnelle pour l'anglais basée sur des corpus textuels. Il s'agit de déterminer les lemmes et les catégories morphosyntaxiques des formes d'un corpus en croisant des observations statistiques effectuées à différents niveaux : fréquence en corpus, similarité des contextes et similarité graphémique. Pour sa part, GOLDSMITH (2001) présente une méthode de segmentation en morphèmes des formes d'un corpus. Cette méthode, basée sur le principe de « description de longueur minimale » (*minimal description length*), a été appliquée à plusieurs langues européennes dont le français. SCHONE et JURAFSKY (2000, 2001) décrivent quant à eux une méthode d'acquisition de la morphologie constructionnelle dans laquelle le sens des formes est représenté par des vecteurs contextuels sémantiques calculés en utilisant la technique de l' « analyse sémantique latente » (*latent semantics analysis* ; (LANDAUER *et al.* 1998)).

Signalons également le travail de JING et TZOUKERMAN (1999) qui présente une méthode permettant de contrôler l'extension de requêtes en RI au moyen des formes morphologiquement apparentées. Il s'agit de déterminer statistiquement la proximité sémantique des mots de la requête avec les formes apparentées qui apparaissent dans les documents en comparant des vecteurs contextuels calculés pour ces deux groupes de mots. La caractérisation sémantique n'est donc pas utilisée pour acquérir des relations morphologiques mais pour les projeter sur les corpus d'une façon plus précise.

3.2 Apprendre la morphologie constructionnelle dans un dictionnaire de synonymes

La méthode proposée ici est différente de celles qui viennent d'être citées puisqu'elle consiste à utiliser comme ressource de départ non pas un corpus textuel, mais un lexique comportant des descriptions sémantiques, à savoir un dictionnaire de synonymes. Les dictionnaires de synonymes sont adaptés à l'acquisition de la morphologie constructionnelle car ils décrivent des relations de partage de propriétés sémantiques entre lexèmes. Or c'est justement l'approximation de ces relations par les graphies qui, dans DéCor, pose problème.

L'utilisation de dictionnaires de synonymes est compatible avec les principes proposés en §1.2. Les connaissances linguistiques sont maintenues hors de l'outil et l'indépendance vis-à-vis des

langues particulières est préservée. La généralité de la méthode l'est aussi car les dictionnaires de synonymes sont des œuvres courantes qui existent pour un grand nombre de langues (au moins en format papier). Plusieurs langues européennes en disposent aussi dans un format électronique. C'est le cas des langues pour lesquelles des dictionnaires WordNet (anglais) ou EuroWordNet (allemand, néerlandais, espagnol, italien...) ont été réalisés (MILLER *et al.* 1990 ; FELLBAUM 1999 ; VOSSEN 1998) : extraire un dictionnaire de synonymes à partir d'un dictionnaire de type WordNet est une opération relativement simple (*cf.* §3.5.1). Par ailleurs, les descriptions sémantiques produites par les rédacteurs de dictionnaires de synonymes sont de très bonne qualité, et dans tous les cas, bien meilleures que les informations que l'on pourrait acquérir automatiquement à partir de corpus textuels.

L'utilisation des dictionnaires de synonymes n'est cependant pas immédiate car les relations de proximité sémantique qu'ils décrivent concernent rarement des lexèmes morphologiquement apparentés. Cette difficulté peut être contounée en utilisant les relations synonymiques de façon indirecte, comme filtres sur les relations de parenté morphologique prédites à partir des graphies des lexèmes[22]. Le filtrage consiste à former des quadruplets morpho-synonymiques $X_1 : X_2 :: Y_1 : Y_2$ tels que :

(1) $X_1 : X_2$ et $Y_1 : Y_2$ sont des couples de lexèmes morphologiquement apparentés ;

(2) a. X_1 est un synomyne de Y_1 ;
b. X_2 est un synomyne de Y_2.

[INSÉRER LA FIGURE 5 ICI]

On peut illustrer ces quadruplets par l'exemple en figure 5. 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. Le quadruplet morpho-synonymique est donc morphologique dans l'une de ses dimensions et synonymique dans l'autre. Sur le plan sémantique, si les relations morphologiques prédites sont valides, alors le quadruplet $X_1 : X_2 :: Y_1 : Y_2$ est analogique, c'est-à-dire que « X_1 est à X_2 ce que Y_1 est à Y_2 »[23]. 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. Le quadruplet en

---

22 Seule la suffixation est considérée dans ce qui suit.

23 La synonymie de X_1 et Y_1 implique qu'ils partagent l'essentiel de leurs propriétés sémantiques. Si X_1 et X_2 sont effectivement morphologiquement apparentés, alors eux aussi partagent une partie de leurs propriétés sémantiques. Par transitivité, Y_1 partage avec X_2 à peu près les mêmes propriétés que X_1. Comme X_2 et Y_2 sont synonymes, ils partagent eux aussi la plupart sinon toutes leurs propriétés sémantiques. Par conséquent, Y_1 et Y_2 partagent une partie de leurs propriétés sémantiques. 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.

figure 5 permet d'acquérir deux couples de lexèmes morphologiquement apparentés : *décoration*:*décorer* et *embellissement*:*embellir*.

Notre méthode est très proche de celle de GRABAR et ZWEIGENBAUM (1999) qui exploitent 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 N. GRABAR et P. ZWEIGENBAUM 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).

3.3 Mise en œuvre et filtres supplémentaires

Les quadruplets morpho-synonymiques sont formés à partir d'un dictionnaire de synonymes. Le dictionnaire est d'abord étiqueté morphosyntaxiquement puis on en extrait le lexique, c'est-à-dire la liste de ses entrées et de ses synonymes. Deux graphes partageant le même ensemble de sommets, à savoir ce lexique, sont ensuite constitués. Les arcs du premier sont les relations de synonymie du dictionnaire. Le second est un graphe de suffixation graphémique construit en utilisant la technique présentée en §2.2 : apprentissage d'un ensemble de schémas de suffixation[24] puis application de ces schémas aux entrées du lexique. Les deux graphes sont ensuite explorés simultanément pour trouver l'ensemble des quadruplets qui vérifient les contraintes (1), (2). En outre, un certain nombre de quadruplets susceptibles d'être erronnés peuvent être éliminés au moyen de trois filtres supplémentaires, partiellement redondants. Le premier impose que :

(3) les formes graphémiques de X_1, X_2, Y_1 et Y_2 doivent toutes être différentes.

Il élimine les quadruplets comme en (4a) qui contiennent des conversions car la caractérisation de la relation sémantique qui s'établit entre leurs éléments est en général difficile.

(4) a. *agglutinant/Ncms*:*agglutinatif/Afpms*::*adhésif/Ncms*:*adhésif/Afpms*
b. *anis/Ncms*:*anisette/Ncfs*::*anisade/Ncfs*:*anis/Ncms*

Ce filtre s'applique également lorsque l'un des deux X a la même forme que l'un des deux Y comme en (4b). Il est en effet peu probable sinon impossible qu'un lexème soit dans la même relation sémantique avec deux éléments différents de sa famille constructionnelle.

Le deuxième filtre stipule que :

(5) seules les relations morphographiques et/ou synonymiques contrôlées par les contraintes (1) et (2) peuvent s'établissent entre les éléments du quadruplet.

Il élimine ainsi les quadruplets tels qu'il existe une relation morphographique entre $X_1 : Y_1$ ou

---

24 Les paramètres de l'apprentissage sont 3 pour la taille minimale du radical et 3 pour le nombre minimal de couples connectés par un schéma. Signalons que la construction de ce graphe a comme objectif de maximiser le rappel ; les erreurs sont filtrées lors de l'identification des quadruplets.

$X_1 : Y_2$ ou $X_2 : Y_1$ ou $X_2 : Y_2$ tels qu'il existe une relation de synonymie entre $X_1 :: X_2$ ou $X_1 :: Y_2$ ou $Y_1 :: X_2$ ou $Y_1 :: Y_2$ ou $X_2 :: X_1$[25] ou $X_2 :: Y_1$ ou $Y_2 :: X_1$ ou $Y_2 :: Y_1$. Il supprime par exemple les quadruplets comme (6a).

(6) a. *forger/Vmn----*:*former/Vmn----*::*constituer/Vmn----*:*construire/Vmn----*
b. *haut/Ncms*:*hauteur/Ncfs*::*profond/Ncms*:*profondeur/Ncfs*

Les relations morphographiques prédites sont ici erronées. Le filtre les détecte grâce à l'existence d'une relation de synonymie entre *forger/Vmn----* et *former/Vmn----* et entre *constituer/Vmn----* et *construire/Vmn----*. Il élimine également des quadruplets comme (6b). Le problème pris en charge par ce filtre vient de ce que les dictionnaires de synonymes décrivent en réalité des relations de proximité sémantique plus ou moins forte.

Le troisième filtre prévoit que :

(7) le radical graphémique de $X_1 : X_2$ ne doit être ni un préfixe ni un suffixe dans celui de $Y_1 : Y_2$ et réciproquement.

Il s'applique d'une part aux quadruplets dont les éléments appartiennent à la même famille morphologique comme (8a).

(8) a. *bassin/Ncms*:*bassinage/Ncms*::*bassinoire/Ncfs*:*bassinement/Ncms*
b. *appeler/Vmn----*:*approcher/Vmn----*::*rappeler/Vmn----*:*rapprocher/Vmn----*

Les radicaux graphémiques de *bassin/Ncms*:*bassinage/Ncms* et de *bassinoire/Ncfs*:*bassinement/Ncms* sont identiques. Dans ce cas, $X_1 : X_2$ et $Y_1 : Y_2$ sont généralement dans des relations sémantiques différentes. Le filtre permet également d'éliminer les quadruplets comme (8b). Dans ce cas, le radical graphémique *app-* est un suffixe du radical *rapp-*. Cet exemple illustre bien le type d'erreur traité. L'apport des trois filtres supplémentaires a été évalué et les résultats sont présentés dans le tableau 1 (§3.4.2).

3.4 Expérience sur le français

Plusieurs expériences d'acquisition de relations de parenté morphologique à partir de quadruplets morpho-synonymiques ont été réalisées sur des dictionnaires de synonymes en français et en anglais. La première expérience a été réalisée sur le français en utilisant DICOSYN, un dictionnaire de synonymes en format électronique constitué à l'INaLF à l'initiative de B. QUÉMADA. Ce dictionnaire a été construit à partir de la fusion de 5 dictionnaires de synonymes (R. BAILLY, H. BÉNAC, H. BERTAUD DU CHAZAUD, M. F. GUIZOT, P.-B. LAFAYE), des synonymes du *Grand Larousse* et des « renvois analogiques » du *Grand Robert*[26].

---

25 Les relations de synonymie ne sont pas nécessairement symétriques (*cf.* §3.4.1).

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

L'hétérogénéité de ces dictionnaires (d'époques et de tailles différentes) permet de gommer les spécificités de chacun d'eux et garantit une certaine généralité à notre étude et aux résultats obtenus.

Plusieurs autres projets de recherche en TAL utilisent ce dictionnaire de synonymes en totalité ou en partie. Ainsi, HAMON *et al.* (1999) présentent un outil d'aide à la structuration et à la mise à jour de terminologie au moyen de liens sémantiques et notamment de la synonymie entre termes complexes. L'outil a été mis au point sur une terminologie construite à partir d'un corpus de documents en langue de spécialité dans le domaine de l'énergie nucléaire. Les liens sémantiques sont inférés en exploitant les renvois analogiques du *Grand Robert* et en faisant « l'hypothèse que la compositionnalité des termes complexes préserve la synonymie ». Ce travail montre clairement que les ressources de langue « générale » et de langue de spécialité (relations sémantiques entre des termes extraites manuellement de corpus techniques ; thésaurus) sont complémentaires. L'inférence de couples de termes se trouvant en relation de synonymie à partir des renvois analogiques du *Grand Robert* s'avère même nettement supérieure, en terme de rappel, à celles qui utilisent des ressources spécialisées. En contrepartie, la précision des couples inférés est faible.

Pour leur part, PLOUX et VICTORRI (1998) utilisent le dictionnaire de synonymes dans sa totalité dans une version corrigée. Leurs objectifs se distinguent des nôtres sur plusieurs points puisqu'ils s'intéressent à la représentation du sens dans un modèle continuiste de la polysémie lexicale. Ce 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 . Leur utilisation du dictionnaire est également relativement différente de la nôtre puisque les synonymes ne sont pas étiquetés et que les relations synonymiques sont rendues symétriques.

3.4.1 Formatage du dictionnaire

Le formatage que nous avons effectué sur le dictionnaire consiste à filtrer les unités lexicales (entrées et synonymes) et à les catégoriser morpho-syntaxiquement. Nous avons ainsi éliminé les mots grammaticaux, les auxiliaires, mais aussi les formes multi-lexicales (locutions et syntagmes) puisque notre objectif est de constituer un lexique constructionnel et que nous ne nous intéressons qu'à la suffixation des lexèmes simples. La catégorisation des unités du dictionnaire est basée sur le fait que la synonymie s'établit entre des lexèmes de même catégorie. Cette contrainte implique une séparation préalable des différents sens et acceptions des entrées. Elle est réalisée au moyen du lexique TLFnome+index. Les unités du dictionnaire présentes dans TLFnome+index reçoivent une des catégories qu'elles ont dans le lexique. Une technique de catégorisation robuste, basée sur l'appariement suffixal, est utilisée pour les lemmes qui n'appartiennent pas à TLFnome+index : une entrée absente de TLFnome+index reçoit la catégorie du plus grand nombre de lemmes de TLFnome+index qui partagent avec elle le plus long suffixe. Par exemple, *économétricien* est catégorisé comme *biométricien/Ncms* et *psychométricien/Ncms*. Après formatage, le dictionnaire comporte 45 009 lemmes (entrées ou synonymes) et 234 771 relations de synonymie ou plus précisément de proximité sémantique (chaque entrée a donc en moyenne 5,2 voisins).

La séparation des sens permet de maximiser le nombre des propriétés sémantiques partagées par

---

`http://www.crisco.unicaen.fr/`).

l'entrée et ses synonymes ainsi que par ces derniers entre eux. Cela n'a cependant pas d'incidence sur la précision des quadruplets extraits car le formatage préserve l'orientation des descriptions synonymiques. Cette préservation est selon nous indispensable. En effet, la synonymie absolue étant un phénomène exceptionnel si on prend en compte les valeurs de connotation, les relations synonymiques décrivent en réalité des proximités sémantiques. Ainsi, le fait d'ajouter un lexème Y parmi les synonymes d'un lexème X ouvre la voie à un ensemble d'autres lexèmes dont le sens est très proche de celui de Y. Par exemple, le fait que *cocasse* et *curieux* soient donnés comme synonymes de *amusant* justifie que *bizarre* soit également considéré comme un de ses synonymes. En discours, on peut, dans certains contextes où *amusant* a une connotation péjorative, le remplacer par *bizarre* sans effet de zeugme ou de maladresse. En revanche, les contextes dans lesquels *bizarre* apparaît habituellement ne permettent pas de lui substituer *amusant*[27]. Ce dernier n'est effectivement pas donné comme synonyme de *bizarre*. On peut expliquer cette situation par le fait que l'idée de bizarrerie est présente et suffisamment centrale dans le sens de *amusant* mais que celle d'amusement est relativement périphérique dans celui de *bizarre*. Une justification linguistique de la non-symétrie de la synonymie dépasse les limites du présent article.

3.4.2 Résultats et évaluation

La méthode décrite en §3.2 a été appliquée au dictionnaire de synonymes français avec différentes combinaisons des filtres supplémentaires. La figure 6 présente quelques-uns des quadruplets acquis lorsque l'on utilise simultanément les trois filtres.

[INSÉRER LA FIGURE 6 ICI]

Ces exemples comportent à la fois des quadruplets corrects comme (9a) : un alchimiste est une personne qui pratique l'alchimie ; un archimage est une personne qui pratique l'archimagie. Mais d'autres comme (9b) ou (9c) sont erronées : si blaguer c'est faire des blagues, rigoler n'est pas « faire des rigolades ». De même, un facétieux fait des facéties mais un drôle ne fait pas de drôleries.

(9) a. *alchimie/Ncfs*:*alchimiste/Ncms*::*archimagie/Ncfs*:*archimage/Ncms*
b. *facétie/Ncfs*:*facétieux/Afpms*::*drôlerie/Ncfs*:*drôle/Afpms*
c. *rigoler/Vmn----*:*rigolade/Ncfs*::*blaguer/Vmn----*:*blague/Ncfs*

Dans ce qui suit, nous notons Q_S l'ensemble des quadruplets extraits du dictionnaire et C_S l'ensemble des couples qui les composent. Signalons que Q_S et C_S sont « désymétrisés ». En effet, si X_1 est apparenté à X_2 alors X_2 est apparenté à X_1 (il en va de même pour Y_1 et Y_2). Ainsi, si $X_1 : X_2 :: Y_1 : Y_2$ vérifie les contraintes (1) et (2) qui définissent les quadruplets morpho-synonymiques alors $X_2 : X_1 :: Y_2 : Y_1$ les vérifient aussi. Il en va de même pour la vérification des filtres (3), (5) et (7).

---

27 C'est par exemple le cas dans la phrase « Je rate trois coups de fusil contre des oiseaux bizarres que j'aurais bien voulu voir de près » GIDE, Voyage au Congo, 1927, p. 763. Cet exemple est extrait de la base Frantext (`http://atilf.atilf.fr/frantext.htm`).

Nous avons estimé le rappel et la précision des couples qui composent les quadruplets extraits à partir du dictionnaire de synonymes français pour différentes configurations des filtres (3), (5) et (7) afin d'évaluer leurs contributions respectives. Les résultats sont rassemblés dans le tableau 1. Ils indiquent que la configuration optimale est celle où le filtre (3) est utilisé seul.

[INSÉRER ICI LE TABLEAU 1]

Les deuxième et troisième colonnes du tableau présente les cardinaux de Q_S et de C_S respectivement. La quatrième colonne donne la précision de l'acquisition de relations de parenté morphologique. Elle a été estimée à partir d'échantillons de 200 couples choisis aléatoirement et révisés pas l'auteur. Le rappel relativement au lexique Verbaction[28] est présenté en colonnes cinq et six. Il a été calculé en considérant les couples de Verbaction comme étant non orientés : les membres de chaque couples sont réordonnés lexicographiquement ; soit C_V l'ensemble de ces couples. Nous avons déterminé la restriction R_V de C_V à l'ensemble des lemmes présents dans C_S, puis l'intersection I_V de R_V avec C_S. Le rappel est donné par le rapport entre les cardinaux de I_V et de R_V. Les deux dernières colonnes du tableau 1 présentent le rappel relativement aux familles constructionnelles de la tranche *fr-* de la nomenclature du *T.L.F.*[29]. Ce taux a été calculé à partir de l'ensemble C_{fr} des couples non orientés qu'elles définissent. Rappelons que ces familles sont des cliques. Comme pour Verbaction, nous avons calculé la restriction de R_{fr} à C_{fr} l'ensemble des lemmes présents dans C_S, puis l'intersection I_{fr} de R_{fr} avec C_S.

3.5 Expérience sur l'anglais

La deuxième expérience a été réalisée à l'aide de trois dictionnaires anglais extraits de la base données WordNet[30] (version 1.7). Ces dictionnaires décrivent des relations de proximité sémantique forte (synonymie stricte), moyenne et faible. Trois objectifs sont en vue :

1. vérifier la réalité de l'indépendance de la méthode des quadruplets morpho-synonymiques vis-à-vis des langues particulières ;
2. estimer la robustesse de la méthode en utilisant des dictionnaires correspondant à des relations de proximité sémantique de plus en plus lâches ;
3. montrer que la méthode proposée est également utile pour des langues comme l'anglais qui disposent déjà de base de données morphologiques, en l'occurrence la

---

28 Verbaction est un lexique qui comporte 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. Il a été réalisé à l'INaLF par A. BERCHE, F. MOUGIN, N. HATHOUT et J. LECOMTE.

29 Cette tranche comporte 614 lemmes regroupés en 160 familles. Le graphe correspondant comprend 2 310 couples de lemmes morphologiquement apparentés (*cf.* §2.3).

30 `http://www.cogsci.princeton.edu/wn/`

base CELEX.

3.5.1 Extraire de WordNet des dictionnaires de proximité sémantique

WordNet est une base de données lexicographiques qui décrit les acceptions de lexèmes en les regroupant au sein de *synsets*, c'est-à-dire d'ensembles de synonymes interchangeables dans certains contextes. Par exemple, le verbe *copy* y a quatre acceptions qui correspondent aux quatre synsets présentés en figure 7. Ces acceptions y sont distinguées par des numéros précédés par le caractère '#'[31]. (Les nombres entre parenthèses qui précèdent les deux premiers synsets correspondent aux nombres d'occurrences de ces acceptions dans les corpus annotés sémantiquement utilisés pour construire la base.) La base de données est divisée en quatre sous-bases qui correspondent aux quatre parties du discours : verbe, nom, adjectif et adverbe. WordNet fournit d'autre part une description du sens au moyen d'un ensemble de relations lexicales fondamentales qui s'établissent entre les synsets : synonymie, hyperonymie, troponymie[32], méronymie, antonymie, etc.

[INSÉRER ICI LA FIGURE 7]

WordNet décrit trois relations de proximité sémantique qu'il est possible de lister au moyen de la commande wn en utilisant les options `-over`, `-syns(v|n|a)` et `-coor(v|n)`. Ces relations ont été utilisées pour créer trois dictionnaires qui décrivent des relations de proximité de moins en moins strictes. Ces dictionnaires permettent d'établir la robustesse de la méthode des quadruplets morpho-synonymiques : l'objectif est de vérifier que la qualité des résultats ne s'effondre pas lorsqu'on utilise un dictionnaire dégradé par l'ajout de relations de proximité plus lâches. Les entrées des dictionnaires sont des lemmes étiquetés par leurs catégories grammaticales. Les informations catégorielles ont donc été distribuées sur les éléments des synsets de chaque sous-base. Par ailleurs, comme pour le dictionnaire français, les lemmes multi-lexicaux ont été supprimés (*cf.* §3.4.1).

La première des relations de proximité sémantique est l'appartenance à un même synset comme par exemple *copy*::*imitate* ou *copy*::*simulate* (commande « `wn -over` »). Il s'agit d'une synonymie stricte. Cette relation n'est pas orientée. Elle définit un premier dictionnaire composé des couples de lemmes $X :: Y$ tels que X et Y appartiennent à un même synset. Les synsets qui contiennent un seul lemme ne contribuent pas, de ce fait, à la construction du premier dictionnaire que nous appellerons S-dict.

La deuxième relation est moins stricte que la précédente. Il s'agit de l'appartenance à des synsets synonymes ou immédiatement hyperonymes comme *copy*::*reproduce* ou *copy*::*resemble* (commandes « `wn -synsv` », « `wn -synsn` » et « `wn -synsa` »)[33]. WordNet ne permet pas

---

31 Ces numéros sont omis dans la suite de l'article car ils n'interviennent pas dans l'extraction des quadruplets. En effet, il n'existe pas de relation régulière entre les numéros des acceptions des lexèmes-construits et ceux des acceptions de leurs lexèmes-bases.

32 Dans la terminologie de WordNet, l'hyperonymie verbale est appelée « troponymie ».

33 Pour les adverbes, la commande « `wn -synsr` » ne fournit que les synsets de leurs bases

de distinguer ces deux relations. Leur union n'est pas symétrique. Par exemple, *copy* n'est donné comme synonyme ou hyperonyme immédiat ni de *reproduce* ni de *resemble*. La synonymie/hyperonymie immédiate définit un deuxième dictionnaire, M-dict, composé des couples de lemmes $X::Y$ tels que $X::Y$ appartient à S-dict ou tels que X et Y appartiennent respectivement à des synsets S_X et S_Y et que S_X a pour synonyme/hyperonyme immédiat S_Y.

La troisième relation, la plus faible, est l'appartenance à des synsets co-hyponymes comme *copy*::*duplicate* ou *copy*::*approximate*. Elle est seulement définie pour les verbes et les noms (commandes « `wn -coorv` » et « `wn -coorn` ») et permet de construire un troisième dictionnaire, L-dict, composés des couples de lemmes $X::Y$ tels que $X::Y$ appartient à M-dict ou tels que X et Y appartiennent respectivement à des synsets S_X et S_Y et que S_X a comme co-hyponyme S_Y.

[INSÉRER ICI LE TABLEAU 2]

Le tableau 2 indique le nombre de lemmes et de couples qui composent chacun des dictionnaires. La dernière colonne donne le nombre moyen de voisins dans les graphes correspondants. On observe que M-dict est une fois et demi plus dense que S-dict et que L-dict l'est onze fois plus. L'augmentation du nombre de relations de proximité a pour contrepartie une diminution de la capacité de ces dictionnaires à filtrer efficacement l'analogie graphémique (*cf.* tableau 3).

3.5.2 Résultats et évaluation

Un ensemble de quadruplets morpho-synonymiques a été acquis à partir de chacun des trois dictionnaires anglais en utilisant le même programme que pour l'expérience en français ; les filtres (3), (5) et (7) ont été appliqués lors de l'expérimentation dont nous présentons les résultats. Le nombre de quadruplets et de couples obtenus ainsi que la précision de ces derniers sont présentés dans le tableau 3. La précision est estimée à partir de la révision par l'auteur d'échantillons de 200 couples choisis au hasard. Nous nous sommes appuyés sur les descriptions sémantiques et les gloses de WordNet pour réaliser cette révision.

[INSÉRER ICI LE TABLEAU 3]

La première conclusion que l'on peut tirer de l'expérience est que notre méthode est effectivement indépendante vis-à-vis des langues particulières (pourvu que leur morphologie soit concaténative). On observe d'autre part qu'il n'y a pas d'effondrement brutal des résultats quand on dégrade les relations de proximité. Le nombre de quadruplets et de couples acquis augmente avec l'ajout de nouvelles entrées et de relations supplémentaires. Parallèlement la précision diminue. Cette dernière est très satisfaisante pour S-dict, supérieure même à celle obtenue pour le dictionnaire français. Elle est en revanche assez faible pour L-dict ; il semble indispensable de contraindre davantage les quadruplets issus de ce dictionnaire pour réduire le bruit (par exemple, en utilisant des informations extraites des gloses des synsets).

Nous avons d'autre part calculé le rappel de la méthode relativement à la partie anglaise de la

---

adjectivales. Cette relation n'est donc pas définie pour les éléments de cette catégorie.

base lexicale CELEX qui contient 52 447 lemmes. Nous en avons extrait les descriptions morphologiques et morphosyntaxiques pour l'ensemble des couples de lemmes qui sont en relation de suffixation. Un certain nombre d'opérations de nettoyage ont été réalisées sur les lemmes qui entrent dans ces couples : élimination des lemmes qui contiennent des majuscules ou des chiffres, des lemmes multi-lexicaux et des lemmes dont la catégorie n'est pas nominale, verbale, adjectivale ou adverbiale ; suppression des particules prépositionnelles rattachées aux verbes et aux noms. Le rappel de la méthode a été calculé à partir de l'ensemble C_c des couples ainsi constitué. Les valeurs obtenues pour les trois dictionnaires sont présentées dans le tableau 4. Nous avons calculé pour chacun d'eux la restriction R_c de C_c aux lemmes qui apparaissent dans C_w (l'ensemble des couples extraits du dictionnaire), puis l'intersection I_c de R_c avec C_w.

[INSÉRER ICI LE TABLEAU 4]

La dernière colonne du tableau 4 donne le nombre des couples de qui n'appartiennent pas à la fermeture symétrique et transitive de . En tenant compte de la précision de la méthode pour chacun des dictionnaires, on peut estimer le nombre de couples susceptibles de compléter respectivement à 5 670, 11 288 et 24 437. On constate ainsi que pour les trois dictionnaires, la moitié environ des couples acquis sont absents de la base CELEX. La méthode que nous proposons serait donc utile pour compléter semi-automatiquement ses descriptions morphologiques.

4 Conclusion

Nous avons décrit dans cet article deux méthodes d'acquisition de connaissances morphologiques permettant de construire des bases de données lexicales destinées à l'expérimentation psycholinguistique, au TAL et à la RI. Ces méthodes exploitent la structure analogique du lexique construit à différents niveaux. La première utilise une ressource de départ assez pauvre, à savoir un lexique flexionnel. Elle s'appuie sur les graphies des lemmes ou des formes fléchies pour acquérir des schémas d'affixation graphémique qui permettent de constituer des couples de lexèmes morphologiquement apparentés. Les couples sont filtrés statistiquement en comparant la fréquence des schémas dont ils sont des instances. La seconde méthode exploite des ressources lexicales beaucoup plus riches puisqu'il s'agit de dictionnaires de synonymes. Les informations sémantiques fournies par les dictionnaires sont croisées avec les connaissances morphologiques acquises au moyen de la première méthode pour réaliser un filtrage qualitatif des couples de lexèmes morphologiquement apparentés.

Les résultats de la première méthode ne peuvent pas être améliorés de manière significative car les informations disponibles sont très limitées. En revanche, ceux de la seconde méthode peuvent l'être, par exemple en typant les quadruplets morpho-synonymiques et les couples qui les composent pour éliminer les moins surs (HATHOUT 2003). Parallèlement, les quadruplets morpho-synonymiques constituent des contextes particulièrement bien adaptés à l'identification des allomorphies et des allographies. Par ailleurs, la méthode des quadruplets morpho-synonymiques peut être adaptée à d'autres ressources comme les bi-textes multilingues, c'est-à-dire des textes alignés qui sont la traduction l'un de l'autre (KRAIF 2001). La synonymie au sein d'une même langue serait alors remplacée par une correspondance sémantique entre lexèmes de deux langues différentes.

Références

BAAYEN, R. Harald, PIEPENBROCK, Richard et GULIKERS, Leon (1995) : « The CELEX Lexical Database (Release 2) ». CD-ROM. Linguistic Data Consortium, University of Pennsylvania, Pennsylvania, USA.

BYBEE, Joan L. (1988) : « Morphology as Lexical Organization ». In HAMMOND, Micheal NOONAN, Michael (éds.), *Theoretical Morphology. Approaches in Modern Linguistics*, chap. 7, p. 119–141. Academic Press, San Diego, CA.

BYBEE, Joan L. (1995) : « Regular Morphology and the Lexicon ». *Language and cognitive processes*, 10(5), p. 425–455.

CORBIN, Danielle (2001) : « Préfixe et suffixes : du sens aux catégories ». *Journal of French Language Studies*, 11(1), p. 41–69.

CRUSE, D. Alan (1986) : *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

DAL, Georgette, HATHOUT, Nabil et NAMER, Fiammetta (1999) : « Construire un lexique dérivationnel : théorie et réalisation ». In AMSILI, Pascal (éd.), *Actes de la VI conférence sur le Traitement Automatique des Langues Naturelle (TALN'99)*, p. 115–124. ATALA, Cargèse, Corse.

DAL, Georgette et NAMER, Fiammetta (2000) : « Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations ». *T.A.L.*, 41(2), p. 423–446.

FABRE, Cécile et JACQUEMIN, Christian (2000) : « Boosting Variant Recognition with Light Semantics ». *Proceedingsof the 18th International Conference on Computational Linguistics (COLING'2000)*, p. 264–270. Luxemburg.

FELLBAUM, Christiane (éd.) (1999) : *WordNet : an Electronic Lexical Database*. MIT Press, Cambridge, Mass.

GAUSSIER, Éric (1999) : « Unsupervised Learning of Derivational Morphology from Inflectional Lexicons ». In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*. Association for Computational Linguistics, ACL'99, University of Mariland, USA.

GOLDSMITH, John (2001) : « Unsupervised Learning of the Morphology of Natural Language ». *Computational Linguistics*, 27(2), p. 153–198.

GRABAR, Natalia et ZWEIGENBAUM, Pierre (1999) : « Acquisition automatique de connaissances morphologiques sur le vocabulaire médical ». In *Actes de la 6e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-99)*, p. 175–184. Cargèse.

HAMON, Thierry, GARCIA, Daniela et NAZARENKO, Adeline (1999) : « Détection de liens de synonymie : complémentarité des ressources générales et spécialisées ». In *Actes de Terminologie*

et Intelligence Artificielle, p. 45–58. Nantes, France.

HATHOUT, Nabil (2000) : « Morphological Pairing based on the Network Model ». In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, p. 35–38. Pyrgos, Grèce.

HATHOUT, Nabil (2003) : « L'analogie, un moyen de croiser les contraintes et les paradigmes. Acquisition de connaissances à partir de dictionnaires de synonymes ». *Revue d'Intelligence Articielle*, 17(5-6), p. 923–934.

HATHOUT, Nabil, NAMER, Fiammetta et DAL, Georgette (2002) : « An Experimental Constructional Database : The MorTAL Project ». In BOUCHER, Paul (éd.), *Many Morphologies*, p. 178–209. Cascadilla, Somerville, Mass.

HAY, Jennifer B. (2000) : *Causes and Consequences of Word Structure*. Thèse de doctorat, Northwestern University, Evanston, IL.

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

JACQUEMIN, Christian (1997b) : *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches, Université de Nantes.

JACQUEMIN, Christian et TZOUKERMANN, Evelyne (1999) : « NLP for term variant extraction : synergy between morphology, lexicon, ans syntax ». In STRZALKOWSKI, Tomek (éd.), *Natural Language Information Retrieval*, p. 25–74. Kluwer Academic Publishers, Dordrecht.

JING, Hongyan et TZOUKERMAN, Evelyne (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)*, p. 90–96. ACM, Berkeley, CA.

KRAIF, Olivier (2001) : « Exploitation des cognats pour l'alignement. Architecture et évaluation ». *Traitement automatique des langues*, 42(3), p. 833–867.

LANDAUER, Thomas K., FOLTZ, Peter W. et LAHAM, Darrell (1998) : « Introduction to Latent Semantic Analysis ». *Discourse Processes*, 25, p. 259–284.

LEPAGE, Yves (1998) : « Solving analogies on words : an algorithm ». In *Proceedings of the of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 2, p. 728–735. Montréal, Canada.

LEPAGE, Yves et SHIN-ICHI, Ando (1996) : « Saussurian analogy : a theoretical account and its application ». In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, vol. 2, p. 717–722. Copenhagen, Danemark.

MILLER, Georges A., BECKWITH, Richard, FELLBAUM, Christiane, GROSS, Derek et MILLER, Katherine J. (1990) : « Introduction to Wordnet : An On-line Lexical Database ». *International Journal of Lexicography*, 3(4), p. 335–391.

NEUVEL, Sylvain et FULOP, Sean A. (2002) : « Unsupervised Learning of Morphology Without Morphemes ». In *Proceedings of the Workshop on Morphological and Phonological Learning 2002*. ACL Publications, Philadelphia.

PIRRELLI, Vito et YVON, François (1999) : « The hidden dimension : a paradigmatic view of data-driven NLP ». *Journal of Experimental & Theoretical Artificial Intelligence*, 11(3), p. 391–408.

PLOUX, Sabine et VICTORRI, Bernard (1998) : « Constructions d'espaces sémantiques à l'aide de dictionnaires de synonymes ». *TAL*, 39(1), p. 161–182.

PLÉNAT, Marc (1988) : « Morphologie des adjectifs en *-able* ». *Cahiers de grammaire*, 13, p. 101–132.

RAJMAN, Martin, LECOMTE, Josette et PAROUBEK, Patrick (1997) : « Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique ».Rapp. Tech., EPFL & INaLF. GRACE GTR-3-2.1.

SCHONE, Patrick et JURAFSKY, Daniel S. (2000) : « Knowledge-Free Induction of Morphology Using Latent Semantic Analysis ». In *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, p. 67–72. Lisbon, Portugal.

SCHONE, Patrick et JURAFSKY, Daniel S. (2001) : « Knowledge-Free Induction of Inflectional Morphologies ». In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*. Pittsburgh, PA.

VOSSEN, Piek (éd.) (1998) : *EuroWordNet : A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

YARKOWSKY, David et WICENTOWSKI, Richard (2000) : « Minimally Supervised Morphological Analysis by Multimodal Alignment ». In *Proceedings of the Association of Computational Linguistics (ACL-2000)*, p. 207–216. Hong Kong.

FIGURE 1

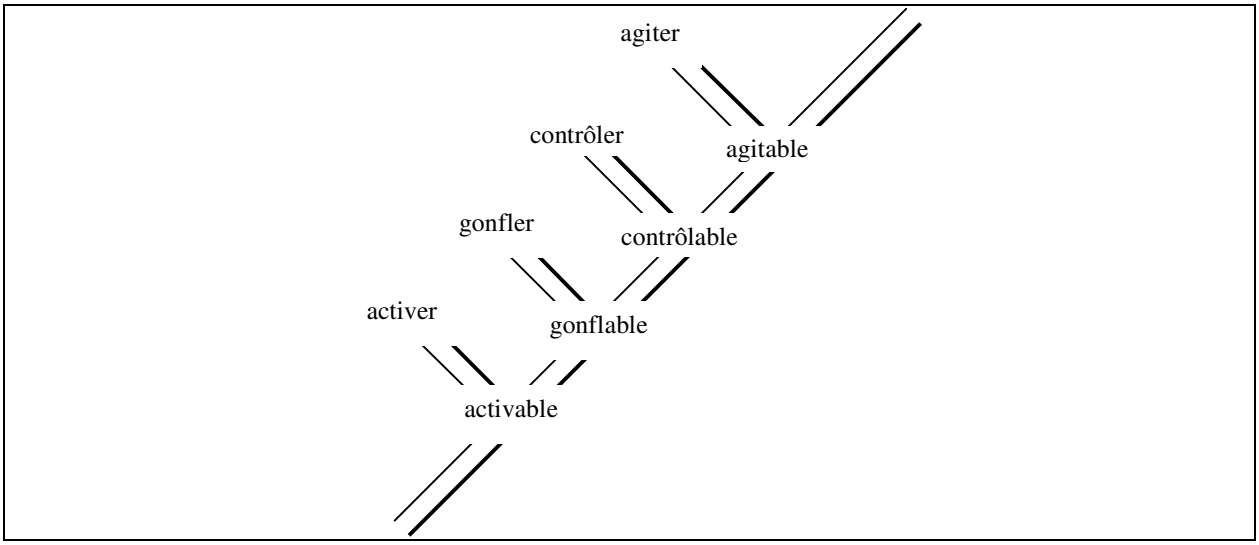


FIG. 1 : Portion du sous-graphe correspondant au suffixe *-able*.

FIGURE 2

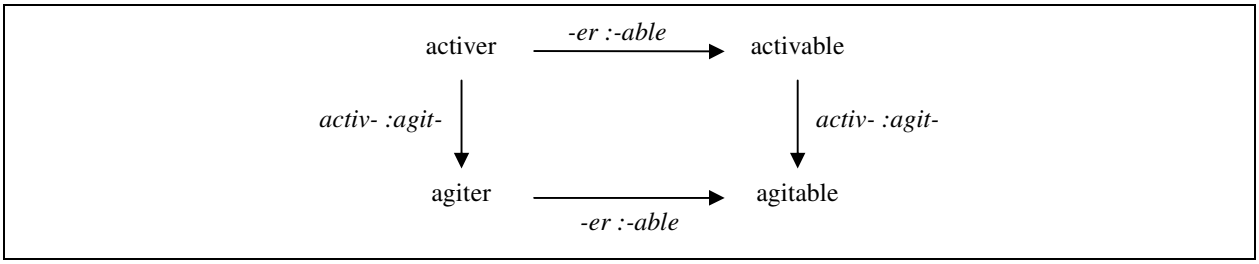
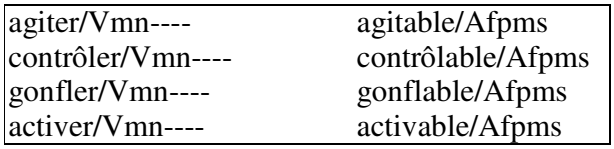


FIG. 2 : Exemple d'analogie graphémique.

FIGURE 3



agiter/Vmn----	agitable/Afpms
contrôler/Vmn----	contrôlable/Afpms
gonfler/Vmn----	gonflable/Afpms
activer/Vmn----	activable/Afpms

FIG. 3 : Série proportionnelle dont les lemmes sont catégorisés

FIGURE 4

–	fraternaliste/Afpms ; fraternaliste/Ncms ; fraternisation/Ncfs ; fraternellement/Rgp ; fraternel/Afpms ; fraterniser/Vmn---- ; fraternité/Ncfs
–	freinateur/Afpms ; freinateur/Ncms ; freination/Ncfs ; freinage/Ncms ; freinette/Ncfs ; freineur/Ncms ; freiner/Vmn---- ; freinée/Ncfs ; frein/Ncms

FIG. 4 : Deux familles constructionnelles constituées par DéClique.

FIGURE 5

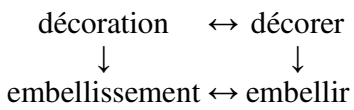


FIG. 5 : Exemple de quadruplets morpho-synonymiques.

FIGURE 6

alchimie/Ncfs	: alchimiste/Ncms	:: archimagie/Ncfs	: archimage/Ncms
alternance/Ncfs	: alternant/Afpms	:: récurrence/Ncfs	: récurrent/Afpms
facétie/Ncfs	: facétieux/Afpms	:: drôlerie/Ncfs	: drôle/Afpms
fouiller/Vmn----	: fouilleur/Afpms	:: fureter/Vmn----	: fureteur/Afpms
introniser/Vmn----	: intronisation/Ncfs	:: couronner/Vmn----	: couronnement/Ncms
métaphore/Ncfs	: métaphorique/Afpms	:: symbole/Ncms	: symbolique/Afpms
révéler/Vmn----	: révélateur/Ncms	:: divulguer/Vmn----	: divulgateur/Ncms
rigoler/Vmn----	: rigolade/Ncfs	:: blaguer/Vmn----	: blague/Ncfs
sobriété/Ncfs	: sobre/Afpms	:: Simplicité/Ncfs	: simple/Afpms
toucher/Vmn----	: touchant/Afpms	:: troubler/Vmn----	: troublant/Afpms

FIG. 6 : Exemples de quadruplets analogiques.

TABLEAU 1

filtres supplémentaires	quadruplets	couples	précision %	rappel %	Verbaction I_V	rappel %	*fr*-I_{fr}
aucun	47 426	26 870	85,5	93,4	2 956	63,4	177
(3)	43 377	24 499	**95,0**	93,3	2 935	65,1	166
(5)	39 474	20 861	89,0	97,4	2 878	65,5	131
(7)	40 535	21 929	94,0	96,6	2 864	61,0	139
(3) et (7)	36 649	19 581	92,0	96,5	2 843	62,0	129
(3), (5) et (7)	35 208	18 286	89,5	97,4	2 821	69,0	120

TAB. 1 : Rappel et précision des couples extraits pour différentes combinaisons des filtres.

FIGURE 7

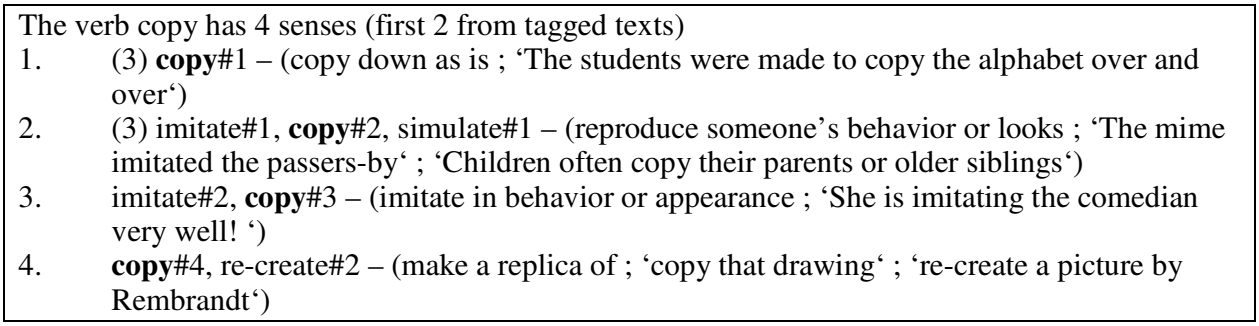


FIG. 7 : Synsets du verbe *copy*.

TABLEAU 2

	entrées	couples	ratio
S-dict	43 055	127 274	3,0
M-dict	62 477	283 422	4,5
L-dict	64 168	2 213 331	34,5

TAB. 2 : Taille des dictionnaires.

TABLEAU 3

	quadruplets	couples	précision
S-dict	15 876	10 211	93,5%
M-dict	37 104	24 205	85,5%
L-dict	175 883	53 755	63,0%

TAB. 3 : Résultats de l'acquisition des quadruplets morpho-synonymiques.

TABLEAU 4

	rappel	I_c	$C_w / F(C_C)$
S-dict	87,0 %	2 860	6 065
M-dict	82,8 %	5 077	13 203
L-dict	73,7 %	6 813	38 789

TAB. 4 : Rappel relativement à la base CELEX.