TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Création de clusters sémantiques dans des familles morphologiques à partir du TLFi

Nuria Gala[1] Nabil Hathout[2] Alexis Nasr[1] Véronique Rey[3] Selja Seppälä[1]
(1) LIF-TALEP, 163, Av. de Luminy case 901, 13288 Marseille Cedex 9
(2) CLLE-ERSS, 5, allées Antonio Machado, 31058 Toulouse Cedex 9
(3) EHESS, 2, rue de la Charité, 13002 Marseille
{nuria.gala, alexis.nasr, selja.seppala}@lif.univ-mrs.fr, nabil.hathout@univ-tlse2.fr,
veronique.rey-lafay@univmed.fr

Résumé. La constitution de ressources linguistiques est une tâche longue et coûteuse. C'est notamment le cas pour les ressources morphologiques. Ces ressources décrivent de façon approfondie et explicite l'organisation morphologique du lexique complétée d'informations sémantiques exploitables dans le domaine du TAL. Le travail que nous présentons dans cet article s'inscrit dans cette perspective et, plus particulièrement, dans l'optique d'affiner une ressource existante en s'appuyant sur des informations sémantiques obtenues automatiquement. Notre objectif est de caractériser sémantiquement des familles morpho-phonologiques (des mots partageant une même racine et une continuité de sens). Pour ce faire, nous avons utilisé des informations extraites du TLFi annoté morpho-syntaxiquement. Les premiers résultats de ce travail seront analysés et discutés.

Abstract. Building lexical resources is a time-consuming and expensive task, mainly when it comes to morphological lexicons. Such resources describe in depth and explicitly the morphological organization of the lexicon, completed with semantic information to be used in NLP applications. The work we present here goes on such direction, and especially, on refining an existing resource with automatically acquired semantic information. Our goal is to semantically characterize morpho-phonological families (words sharing a same base form and semantic continuity). To this end, we have used data from the TLFi which has been morpho-syntactically annotated. The first results of such a task will be analyzed and discussed.

Mots-clés : Ressources lexicales, familles morphologiques, clusters sémantiques, mesure de Lesk.

Keywords: Lexical resources, morphological families, semantic clusters, Lesk measure.

1 Introduction

Les ressources linguistiques sont indispensables aussi bien dans une perspective de traitement automatique de la langue que dans le cadre d'une utilisation humaine (apprentissage des langues, thérapie orthophoniste, etc.). Une des problématiques saillantes dans ce domaine concerne leur constitution : elle s'avère longue et coûteuse, spécialement lorsqu'on vise des informations fines comme la description de l'organisation dérivationnelle du lexique. La création de ce type de ressource peut être réalisée soit manuellement comme (Gala & Rey, 2008), soit à partir d'informations morphologiques dérivationnelles acquises automatiquement. Pour le français, citons le projet MorTAL (Hathout *et al.*, 2002; Dal *et al.*, 2004), les travaux en informatique médicale qui portent sur l'apprentissage de relations morphologiques en corpus spécialisés (Zweigenbaum *et al.*, 2003; Langlais *et al.*, 2009). Signalons également que plusieurs travaux portent sur l'acquisition automatique de familles morphologiques du français comme (Gaussier, 1999; Bernhard, 2007; Hathout, 2009; Lavallée & Langlais, 2010). Notre objectif à terme est de créer une ressource comparable à la base CELEX [1] (Baayen *et al.*, 1995), c'est-à-dire, un lexique avec des informations morpho-sémantiques fines pour au moins 50 000 mots du français.

Dans cet article, nous présentons une méthodologie de création de sous-familles ou clusters sémantiques dans des

1. CELEX décrit la phonologie et la morphologie de trois langues germaniques, l'anglais, l'allemand et le néerlandais. Elle contient notamment une analyse morphologique dérivationnelle fine pour l'ensemble des lexèmes : 52 447 entrées pour l'anglais, 51 728 pour l'allemand et 124 136 pour le néerlandais

familles morpho-phonologiques existantes. La section 2 décrit les principes théoriques et la ressource à la base de notre travail. Dans les sections 3 et 4, nous décrivons les critères utilisés et la méthode mise en œuvre pour le partitionnement automatique. Enfin, la dernière section analyse et discute les résultats obtenus.

2 Principes théoriques et justifications méthodologiques

2.1 Familles morpho-phonologiques

Ce travail s'inscrit dans le cadre d'une approche paradigmatique de la morphologie. Dans ce contexte, deux notions peuvent être utilisées pour rendre compte de la construction des unités lexicales : la notion de « famille » et celle de « série » (Roché *et al.*, 2011). En synchronie, la notion de « famille morphologique » ne réfère pas à une filiation historique mais à une relation de forme et de sens. La famille constitue alors un ensemble d'unités associées en raison d'un élément commun appelé « racine » et permettant le classement des mots du lexique en familles morphologiques (Gaussier, 1999; Hathout, 2005). Une famille morphologique est donc un ensemble de mots partageant une même racine (base) et une continuité de sens comme on peut le voir pour *boule* en figure 1 :

aboulie, dysboulie, abouler, boule, bouler, boulet, bouletage, bouleter, bouleté, boulette, boulier, boulisme, bouliste, boulocher, boulodrome, boulotter, débouler, rabouler, boulon, boulonnage, boulonner, boulonnerie, déboulonnage, déboulonnement, déboulonner, indéboulonnable, chamboulement, chambouler, éboulement, ébouler, éboulis

FIGURE 1 – Famille morphologique de *boule* dans la base Polymots.

En français, une base a généralement plusieurs thèmes (Roché, 2010) qui présentent des alternances vocaliques, par exemple, *œ*/*ou* (*cœur*/*courage*), *eu*/*o* (*fleur*/*floraison*), etc. Il existe aussi des alternances consonantiques, comme *d*/*t* (*verdure*/*verte*) ou *f*/*v* (*actif*/*activité*), et des phénomènes de consonne latente (*long*/*longuement* ; *tapis*/*tapissier* ; *petit*/*petitesse*). On observe par ailleurs que la taille des familles peut aussi varier de façon significative car certains mots sont morphologiquement isolés (*genou*, *intrigue*, *moment*, etc.), alors que d'autres appartiennent à des familles très nombreuses (*acte*/*ag-*, *fil*, *port*, etc.).

Polymots [2] (Gala & Rey, 2008) est une base lexicale pour le français qui décrit de telles familles morphologiques. Elle a été construite manuellement à partir du Petit Larousse 2000 (20 000 mots pleins ont été extraits et analysés morphologiquement : 2 004 familles ont été identifiées). L'objectif de Polymots étant la description du lexique en terme de dérivation, une partie des entrées du dictionnaire a été exclue en faisant une hypothèse *a priori* sur la rentabilité des bases : exclusion de mots qui construisent très peu de dérivés [3]. Sont exclus :

– les noms propres (et leurs dérivés) ;
– les noms composés qui ne sont pas graphiquement simples (qui contiennent, par exemple, un tiret) ;
– les termes techniques comme les noms de maladies, de champignons ou d'animaux, à l'exception des noms des animaux familiers comme *chien*, *chat*, *vache* qui eux figurent dans la base.

Le premier principe du regroupement en familles est le partage d'un radical morpho-phonologique. Une vérification dans deux dictionnaires étymologiques a permis de confirmer ou d'infirmer ces rapprochements. Par exemple, le mot *pause* a été rapproché de la forme *poser* (et donc inclu dans la même famille) et ce malgré une graphie différente. A l'issue d'une validation manuelle, nous avons exclu des mots comportant des marques de flexion (opposition de genre ou de nombre) sauf dans les cas où cette marque est significative (par exemple, *allumette*, *annales*, *fiançailles*, etc.). Par ailleurs, aucune information grammaticale (partie du discours) ou sociolinguistique (registre) n'était initialement fournie pour les mots. Des étiquettes grammaticales — extraites du Larousse — ont cependant été ajoutées dans la nouvelle version de la base, qui sera mise en ligne dans le courant de l'été 2011.

2. `http://polymots.lif.univ-mrs.fr`
3. Par exemple, dans le Petit Larousse, *France* n'a dans sa famille que *franciser* et *français*, *girafe* que *girafon* et *pot-au-feu* est seul dans sa famille dérivationnelle. Ce critère est en cours d'étude suite aux résultats obtenus lors des analyses : environ 15 % des bases de Polymots ont une rentabilité < 3 mots (*antenne*, *hangar*, *paupière*, etc.)

2.2 Caractérisation sémantique

Le regroupement des mots sur la base de critères morphophonologiques et étymologiques produit souvent des familles hétérogène du point de vue sémantique, et rend leur exploitation pour le TAL difficile. Pour réduire cette difficulté, plusieurs directions ont été explorées. La première par (Gala *et al.*, 2009) qui présentent un premier enrichissement de Polymots par un ensemble d'unités de sens associé à chaque mot. Ces unités ont été acquises dans des corpus structurés librement accessibles comme Wikipédia ou Wiktionnaire et ont permis une première caractérisation en termes de cohésion et de dispersion sémantique [4].(Gala *et al.*, 2009) ont ainsi mis en évidence que pour certaines familles la cohésion sémantique est très forte (par exemple, le mot *terre* est présent dans les définitions lexicographiques de *aterrir*, *enterrer*, *extraterrestre*, *terrain*, *terrasse*, etc.), dans d'autres, il existe une dispersion sémantique plus grande : une ou plusieurs composantes du mot base sont présentes dans les mots dérivés. Par ailleurs, il est difficile dans certains cas de percevoir en synchronie les liens sémantiques entre des mots d'une même famille (par exemple, *dispenser* et *dispendieux*, dans la famille de *penser*, le premier dérivé est lié à l'idée d'*autoriser*, le deuxième à la sphère du *pécunier*).

Étant donné que les informations sémantiques peuvent être utiles dans les applications en traitement automatique des langues, nous avons jugé nécessaire de mieux caractériser sémantiquement les familles morphologiques de Polymots. Ansi, il nous a paru important de délimiter, dans chaque famille, des ensembles de mots partageant une même notion sémantique. Nous avons alors partitionné des familles sous des clusters sémantiques, que nous définissons comme des sous-familles morphologiques où les mots présentent une forte cohésion sémantique autour d'une notion commune. Par exemple, la famille de *boule* qui contient initialement 31 mots dérivés peut être divisés en six clusters homogènes comme en figure 2 :

Composante commune	**Cluster**
trouble	aboulie, dysboulie
boule	abouler, boule, bouler, boulet, bouletage, bouleter, bouleté, boulette, boulier, boulisme, bouliste, boulocher, boulodrome, débouler, rabouler
boulon	boulon, boulonnage, boulonner, boulonnerie, déboulonnage, déboulonnement, déboulonner, indéboulonnable
chambouler	chamboulement, chambouler
ébouler	éboulement, ébouler, éboulis

TABLE 1 – Clusters sémantiques dans la famille morphologique de *boule*.

3 Création automatique de clusters sémantiques

Nous avons retenu trois critères pour la création des clusters sémantiques : le premier est morphologique (partage d'affixes), le deuxième lexico-sémantique(partage de mots communs dans les définitions du TLFi) et le troisième lexico-morphologique (présence de mots de la même famille dans les définitions du Larousse). La découpage des familles en clusters a été réalisé à partir d'un corpus de 658 mots appartenant à 20 familles morpho-phonologiques de Polymots. Chacune de ces familles a été initialement partitionné à la main, en fonction du partage d'une unité de sens à tous les membres du cluster (voir colonne 1 dans la table 1).

3.1 Critère morphologique

Le critère morphlogique retenu est le nombre d'affixes communs que chaque mot partage avec les autres mots de sa famille morpho-phonologique. Pour chaque paire de mots dérivés (e_i, e_j) dans une famille, notre système calcule le score $s_1(e_i, e_j)$, qui n'est autre que le nombre d'affixes communs, par exemple : s_1(*déboulonner*, *in-déboulonnable*) $= |\{$*dé-*, *-on*$\}| = 2$; s_1(*boulonner*, *déboulonner*) $= 1$; s_1(*boulonner*, *déboulonnement*) $= 1$; s_1(*chamboulement*, *boulonner*) $= 0$, etc.

4. Pour des raisons d'espace, il est impossible de décrire ici la méthode utlisée. Les auteurs renvoyent le lecteur à l'article (Gala *et al.*, 2009)

Les affixes n'ont été comparés que pour les membres d'une même famille. Dans certains cas le score s_1 peut induire en erreur comme pour s_1(*chamboulement*, *déboulonnement*) $= 1$ alors qu'on cherche à séparer ces deux mots de la famille *boule* dans deux clusters sémantiques différents (liés aux composantes 'chambouler' et 'boulon' respectivement).

3.2 Critère lexico-sémantique

Le deuxième critère mesure la similarité sémantique sur la base des mots partagés par les définitions lexicographiques. Cette technique initialement proposée par (Lesk, 1986) a été reprise par de nombreux auteurs, notamment (Banerjee & Pedersen, 2002). Nous avons utilisé les unités lexicales présentes dans les définitions du *Trésor de la Langue Française informatisé* (TLFi) annotées avec l'analyseur syntaxique MACAON (Nasr *et al.*, 2010). Dans le TLFi, chaque mot correspond à un vocable (ou plusieurs, en cas d'homonymie) et chaque vocable peut avoir différents sens (lexies). Lorsqu'une lexie est accompagnée d'une définition, celle-ci est extraite et envoyée dans la chaîne de traitement de MACAON. Le texte est tout d'abord segmenté en phrases, puis en unités pré-lexicales et en mots. Les mots se voient ensuite associer une catégorie morpho-syntaxique, avant d'être lemmatisés. Chaque mot de la définition ainsi annoté est présenté sous la forme suivante : *forme* | *catégorie* | *lemme* (voir figure 2). Pour la création des clusters, seules les définitions du vocable sont retenues (pas celles des expressions figées liées au vocable) et seuls les lemmes des mots pleins (en gras dans l'exemple) sont gardés.

```
<VOCABLE id="28590" text="ÉBOULEMENT, " cat="subst. masc."> <S n="1">
<DEFI id="28590-1">Chute|nc|chute de|prep|de ce|pro|ce qui|prorel|qui s'|clr|s'
éboule|v|ébouler .|poncts|. </DEFI> <DEFI id="28590-2" ind="P. méton.">Amas/nc/amas
de/prep/de matériaux/nc/matériau éboulés/vppart/ébouler ./poncts/. </DEFI>
```

FIGURE 2 – Définition d'*éboulement* dans le TLFi, étiquetée et lemmatisée.

Pour chaque paire de mots dérivés appartenant à une même famille, nous attribuons un score lexical $s_2(e_i, e_j)$ correspondant à l'indice de Jaccard (rapport entre le nombre de lemmes communs aux définitions des deux mots dans le TLFi et le nombre total de lemmes qui apparaissent dans ces définitions). Par exemple : s_2(*éboulement, ébouler*) $= 0.000$; s_2(*éboulement, éboulis*) $= 0.222$; s_2(*ébouler, éboulis*) $= 0.036$, etc. Pour ces exemples, s_2 a été calculé à partir des ensembles de mots suivants extraits des définitions du TLFi présentées en figure3 :

ÉBOULEMENT	chute, ébouler, amas, matériau
ÉBOULER	faire, tomber, désagrégation, affaisser, formation, naturel, artificiel, écrouler, effondrer, laisser, roulant
ÉBOULIS	chute, ébouler, amas, matériau, généralement, naturel, ensemble, débris, rocheux, détacher, abrupt, former, talus, incliné, plan, conique, fort, pente

FIGURE 3 – Mots pleins des définitions du TLFi utilisés pour le calcul de s_2 entre mots de la même famille morphologique.

3.3 Critère lexico-morphologique

Le troisième critère est l'apparition de mots de la même famille morphologique dans les définitions du dictionnaire Larousse. On trouve par exemple *boule* dans les définitions de *boulet*, *boulier*, *boulisme*, *bouliste*, *boulocher*, etc. Nous avons ainsi défini un trait binaire $s_3(e_i, e_j)$ qui vaut 1 si un mot de la famille apparait dans l'une des définitions de l'entrée et 0 sinon (e_i est le mot traité, e_j un mot de sa famille présent ou non dans la défintion) : s_3(*boulet, boule*) $= 1$; s_3(*bouletage, boulette*) $= 1$; s_3(*bouleté, boulet*) $= 1$; s_3(*boulette, boule*) $= 1$; etc.

4 Clustering

Le regroupement d'éléments d'une même famille au sein de sous-familles (ou clusters) a été réalisé à l'aide d'un algorithme de clustering hiérarchique. Son principe est simple : étant donné une famille $\mathcal{F} = \{l_1, \ldots l_n\}$, et une

Famille	Taille famille	Rand index	Famille	Taille famille	Rand index
mode	42	0.90	paille	26	0.54
pot	15	0.68	ten	55	0.91
val	58	0.79	meuble	28	0.68
onde	24	0.74	pens	27	0.79
presse	58	0.81	terre	37	0.62

TABLE 2 – Evaluation du partitionnement automatique sur un ensemble de 10 familles

mesure de similarité $s(\cdot,\cdot)$, on commence par réaliser une partition initiale de $\mathcal{F}$, qui consiste à créer une partie pour tout élément de $\mathcal{F}$: $\{\{l_1\},\dots\{l_n\}\}$. Les deux parties $\mathcal{P}_x$ et $\mathcal{P}_y$ les plus proches au sens de la mesure de similarité s sont alors fusionnées au sein d'une même partie. Le processus est réitéré sur la nouvelle partition. Le processus prend fin lorsque tous les éléments de $\mathcal{F}$ ont été regroupés au sein d'une seule partie ou lorsque la similarité entre les deux parties les plus proches est inférieure à un seuil τ.

La mesure de similarité que nous avons utilisée est une combinaison linéaire des trois critères s_1, s_2 et s_3. Ainsi la similarité entre les deux éléments e_i et e_j est $s(e_i,e_j) = \alpha s_1(e_i,e_j)+\beta s_2(e_i,e_j)+\gamma s_3(e_i,e_j)$ avec $\alpha+\beta+\gamma = 1$. Cette mesure est étendue à la mesure entre deux parties $\mathcal{P}_x$ et $\mathcal{P}_y$ en réalisant simplement la moyenne sur tous les couples $(e_i,e_j) \in \mathcal{P}_x \times \mathcal{P}_y$:

$$s(\mathcal{P}_x,\mathcal{P}_y) = \frac{1}{|\mathcal{P}_x \times \mathcal{P}_y|} \sum_{(e_i,e_j)\in\mathcal{P}_x\times\mathcal{P}_y} s(e_i,e_j)$$

5 Évaluation

L'évaluation a été réalisée sur un ensemble de 20 familles pour lesquelles la partition de référence a été réalisée manuellement. Les dix premières familles ont été utilisées comme ensemble de développement et ont permis d'optimiser les coefficients α, β et γ ainsi que le seuil τ.

Les paramètres optimaux ainsi obtenus ont ensuite été utilisés pour évaluer le partitionnement sur l'ensemble de test, composé des dix dernières familles. La mesure d'évaluation utilisée est le *rand index*, qui est une mesure de comparaison de partitions d'un même ensemble. Etant donné deux partitions X et Y de l'ensemble E, on définit :

- a comme le nombre de paires d'éléments de E qui appartiennent à la même partition dans X et dans Y
- b comme le nombre de paires d'éléments de E qui n'appartiennent pas à la même partition dans X ni dans Y
- c comme le nombre de paires d'éléments de E qui appartiennent à la même partition dans X mais n'appartiennent pas à la même partition dans Y
- d comme le nombre de paires d'éléments de E qui n'appartiennent pas à la même partition dans X mais appartiennent à la même partition dans Y.

Le rand index de X et Y est alors :

$$\frac{a+b}{a+b+c+d}$$

La valeur maximale que peut prendre le rand index est 1, lorsque les deux partitions sont identiques, et 0 lorsqu'il n'existe aucun couple dont les deux éléments appartiennent à la même partie dans X mais n'appartiennent pas à une même partie dans Y ou inversement.

Les résultats de l'évaluation sur notre ensemble de test sont présentés dans la table 2. Cette table montre une grande disparité dans les résultats obtenus pour les différentes familles : la moyenne du rand index sur l'ensemble de 10 familles est de $74,6$ pour un écart type de $9,4$. Cela s'explique en partie par le manque de stabilité des résultats obtenus lors de l'optimisation des paramètres sur l'ensemble d'évaluation. En effet, lors de cette étape, la recherche des paramètres optimaux est effectuée par famille. La moyenne de chacun des paramètres est ensuite effectuée pour obtenir les paramètres qui seront utilisés dans la phase de test. Or les paramètres optimaux lors de la première étape présentent une grande disparité. Les paramètres optimaux pour une famille sont par conséquent éloignés des paramètres définitifs. Pour illuster cela, nous avons calculé la moyenne du rand index sur l'ensemble

de développement en utilisant les paramètres définitifs, le résultat est de $64,4$, qu'il convient de comparer avec une moyenne de $81,5$, obtenue en utilisant les paramètres optimaux pour chaque famille.

6 Conclusions et perspectives

Nous avons présenté une méthode de partitionnement automatique de familles morpho-phonologiques en clusters sémantiques. À l'aide de trois critères (morphologique, lexico-sémantique et lexico-morphologique) nous avons calculé une mesure qui a permis le regroupement des mots d'une même famille en clusters autour d'une même notion sémantique. Malgré la disparité de résultats entre les familles et l'échantillon réduit choisi pour ce premier travail, les résultats obtenus sont encourageants. Les perspectives de ce travail concernent la généralisation du clustering pour l'ensemble de familles existantes dans Polymots (2 004 familles à ce jour). Nous comptons, également, rajouter des critères plus fins, comme la prise en compte d'informations syntaxiques, afin d'obtenir de meilleurs résultats pour l'ensemble des familles.

Références

BAAYEN R. H., PIEPENBROCK R. & VAN RIJN H. (1995). The Celex lexical database (release 1) [cd-rom].

BANERJEE S. & PEDERSEN T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, p. 136–145, London, UK : Springer-Verlag.

BERNHARD D. (2007). Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique. In F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds., *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, p. 367–376, Toulouse : ATALA IRIT.

DAL G., HATHOUT N. & NAMER F. (2004). Morphologie constructionnelle et traitement automatique des langues : le projet mortal. *Lexique*, **16**, 199–229.

GALA N. & REY V. (2008). Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In *Actes de TALN 2008 (Traitement automatique des langues naturelles)*, Avignon : ATALA LIA.

GALA N., REY V. & TICHIT L. (2009). Dispersion sémantique dans des familles morpho-phonologiques : éléments théoriques et empiriques. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

GAUSSIER E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *ACL Workshop on Unsupervised Learning in Natural Language Processing*, College Park, MD.

HATHOUT N. (2005). Exploiter la structure analogique du lexique construit : une approche computationnelle. *Cahiers de Lexicologie*, **87**(2).

HATHOUT N. (2009). Acquisition of morphological families and derivational series from a machine readable dictionary. In F. MONTERMINI, G. BOYÉ & J. TSENG, Eds., *Selected Proceedings of the 6th Décembrettes : Morphology in Bordeaux*, Somerville, Mass. : Cascadilla Proceedings Project.

HATHOUT N., NAMER F. & DAL G. (2002). An experimental constructional database : The mortal project. *Many Morphologies*, p. 179–209.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2009). Improvements in analogical learning : application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, p. 487–495, Stroudsburg, PA, USA : Association for Computational Linguistics.

LAVALLÉE J.-F. & LANGLAIS P. (2010). Analyse morphologique non supervisée par analogie formelle. In *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal : ATALA RALI.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : how to tell a pine code from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, p. 24–26, Toronto, Canada.

NASR A., BÉCHET F. & REY J. (2010). Macaon une chaîne linguistique pour le traitement de graphes de mots. In *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal : ATALA RALI.

ROCHÉ M. (2010). Base, thème, radical. *Revue linguistique de Vincennes*, **39**, 95–134.

ROCHÉ M., BOYÉ G., HATHOUT N., LIGNON S. & PLÉNAT M. (2011). *Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en* anti-, In *Des unités morphologiques au lexique*, p. 251–318. Hermes Science-Lavoisier : Paris. Chapitre rédigé par N. Hathout.

ZWEIGENBAUM P., HADOUCHE F. & GRABAR N. (2003). Apprentissage de relations morphologiques en corpus. In B. DAILLE, Ed., *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, Batz-sur-mer : ATALA IRIN.