Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique

Delphine Bernhard
TIMC-IMAG, Institut d'Ingénierie de l'Information de Santé
Faculté de Médecine, 38706 La Tronche cedex
Delphine.Bernhard@imag.fr

Résumé. Cet article présente un système d'acquisition de familles morphologiques qui procède par apprentissage non supervisé à partir de listes de mots extraites de corpus de textes. L'approche consiste à former des familles par groupements successifs, similairement aux méthodes de classification ascendante hiérarchique. Les critères de regroupement reposent sur la similarité graphique des mots ainsi que sur des listes de préfixes et de paires de suffixes acquises automatiquement à partir des corpus traités. Les résultats obtenus pour des corpus de textes de spécialité en français et en anglais sont évalués à l'aide de la base CELEX et de listes de référence construites manuellement. L'évaluation démontre les bonnes performances du système, indépendamment de la langue, et ce malgré la technicité et la complexité morphologique du vocabulaire traité.

Abstract. This article describes a method for the unsupervised acquisition of morphological families using lists of words extracted from text corpora. It proceeds by incrementally grouping words in families, similarly to agglomerative hierarchical clustering methods. Clustering criteria rely on graphical similarity as well as lists of prefixes and suffix pairs which are automatically acquired from the target corpus. Results obtained for specialised text corpora in French and English are evaluated using the CELEX database and manually built reference lists. The evaluation shows that the system perfoms well for both languages, despite the morphological complexity of the technical vocabulary used for the evaluation.

Mots-clés : familles morphologiques, classification, apprentissage non supervisé.

Keywords: morphological families, clustering, unsupervised learning.

1 Introduction

L'analyse morphologique est une tâche importante dans divers domaines du traitement automatique des langues comme la reconnaissance de la parole, la communication alternative et augmentée, la traduction automatique ou la recherche d'informations. Dans ce dernier cas, l'utilité des connaissances morphologiques se justifie par la proximité sémantique des variantes flexionnelles ou dérivationnelles. Il est également possible d'exploiter

la structure morphologique des mots pour l'acquisition de relations sémantiques telles que l'hyponymie (Buitelaar & Sacaleanu, 2002) ou l'antonymie (Schwab *et al.*, 2005). Les ressources décrivant les liens morphologiques n'étant pas disponibles à l'heure actuelle pour toutes les langues et tous les domaines, ces applications sont fréquemment associées à l'acquisition automatique de connaissances morphologiques à partir de textes. Les méthodes d'analyse morphologique non supervisée sont variées : comparaison de graphies (Zweigenbaum & Grabar, 2000), recherche d'analogies (Lepage, 1998), modèles probabilistes (Creutz & Lagus, 2005) ou segmentation par optimisation (Goldsmith, 2001; Creutz & Lagus, 2002). Elles se distinguent également par le type de résultats obtenus : mots découpés en segments morphémiques ou liens morphologiques.

Le travail présenté dans cet article relève du second type de méthode car il consiste en l'acquisition de familles morphologiques, c'est-à-dire des groupes de mots liés deux à deux par un lien morphologique d'affixation (préfixation ou suffixation) ou de composition. Nous formulons la question de l'acquisition de familles morphologiques comme un problème de classification. En effet, l'objectif de la classification est d'organiser un ensemble de données en groupes homogènes et contrastés : dans notre cas, les groupes souhaités sont des familles de mots morphologiquement reliés. La méthode que nous proposons prend pour point de départ une liste de mots et les groupe en familles d'une manière similaire aux méthodes de classification ascendante hiérarchique utilisées en analyse de données. Elle a de plus la particularité d'être non supervisée et n'est donc pas liée à une langue ou à un domaine précis.

Nous allons dans un premier temps décrire les diverses étapes de la méthode avant de présenter et d'analyser les résultats obtenus pour des corpus de textes techniques (médecine et volcanologie) en français et en anglais. Nous nous intéressons plus particulièrement au vocabulaire technique car il se caractérise par l'utilisation fréquente des procédés de composition et de dérivation, notamment par préfixation.

2 Description de la méthode

Le système prend pour entrée les données suivantes :

- Une liste des mots d'un corpus L
- Une liste de préfixes P
- Une liste de signatures (ou paires de suffixes) S

Les deux dernières listes sont obtenues à partir de la première à l'aide du module d'apprentissage d'affixes décrit dans (Bernhard, 2006). Celui-ci utilise les probabilités transitionnelles entre sous-chaînes pour repérer les zones de faible probabilité et ainsi découper les mots en radical et affixes. Nous avons adapté ce module pour qu'il produise non seulement une liste de préfixes et de suffixes mais également une liste de paires de suffixes qui apparaissent avec la même base et qui sont donc mutuellement substituables sur l'axe paradigmatique[1]. Par exemple, les suffixes de la paire (*s*,*ique*) peuvent se combiner à la base *climat* pour former les mots *climat**s*** et *climat**ique***. La même signature se retrouve dans les paires de mots *volcan**s*** – *volcan**ique*** et *océan**s*** – *océan**ique***. La notion de signa-

[1]Il faut noter que les préfixes et les suffixes sont acquis automatiquement, de manière non supervisée. Par conséquent, aucune distinction n'est faite entre les affixes flexionnels et dérivationnels.

ture est présente dans de nombreux travaux en acquisition automatique de connaissances morphologiques, parfois sous des dénominations différentes : *paires de suffixes* (Gaussier, 1999), *règles morphologiques* (Grabar & Zweigenbaum, 1999) ou *schémas de suffixation* (Hathout, 2005).

Nous allons maintenant détailler l'ensemble des étapes menant à l'acquisition des familles morphologiques.

2.1 Familles initiales

Avant apprentissage, il y a autant de familles que de mots dans la liste donnée en entrée : chaque mot constitue sa propre famille. Les familles formées au cours du processus d'apprentissage sont représentées par un radical R. De plus, chaque famille comprend deux sous-familles, sauf si elle correspond à une feuille dans la hiérarchie : dans ce cas, elle contient un mot unique et n'a pas de sous-famille.

2.2 Étape 1 : regroupement de familles à partir de l'inclusion de mots

Le premier critère de regroupement des familles est l'inclusion de mots : il s'agit de repérer les mots formés par préfixation à partir d'un autre mot de la liste, selon une procédure détaillée ci-dessous :

Soient :
– $m_1, m_2, \ldots, m_i$ et m_j des mots de longueur minimale égale à 4 ;
– $F_1, F_2, \ldots, F_i$ des familles telles que $F_1 = [m_1], F_2 = [m_2], \ldots, F_i = [m_i]$;
– F_j une famille telle que $F_j = [m_j]$.
Les familles $F_1, F_2, \ldots, F_i$ et F_j sont regroupées pour former une nouvelle famille F_k si $m_1 = E_1 + m_j, m_2 = E_2 + m_j, \ldots, m_i = E_i + m_j$
où E_1, E_2, …, E_i représentent une suite maximale d'un ou plusieurs préfixes de la liste P, éventuellement séparés par des tirets, tels que chaque préfixe ait une longueur minimale de 3.
Le radical de la nouvelle famille F_k est m_j.

Par exemple, si F_1 = [sub-océaniques], F_2 = [océaniques] et F_3 = [intra-océaniques] alors il est possible de former une nouvelle famille F_4 telle que $F_4 = F_1 \cup F_2 \cup F_3$ = [sub-océaniques, océaniques, intra-océaniques]. En effet, les mots *sub-océaniques* et *intra-océaniques* contiennent tous le mot *océaniques*. De plus, ils débutent par les préfixes *sub+* et *intra+*. Le radical de la nouvelle famille est *océaniques*.

2.3 Étape 2 : regroupement de familles à partir des préfixes

Après avoir procédé à un premier regroupement des mots en fonction des mots inclus, nous utilisons d'autres critères de regroupement, basés sur la comparaison des graphies des radicaux des familles existantes et des préfixes auxquels ils peuvent être associés. En

effet, lorsque deux mots partagent un même préfixe et que leurs bases sont graphiquement similaires, alors il y a de fortes chances pour qu'ils soient également morphologiquement liés. Prenons l'exemple des mots suivants : *neuro-oncologist* et *neuro-oncology*. Ces deux mots débutent tous deux par le préfixe *neuro-* suivi d'une même chaîne de caractères de longueur 7 : *oncolog*. La combinaison de deux indices, à savoir le partage d'un préfixe, suivi d'une chaîne commune, est un indice suffisant dans la plupart des cas pour conclure que les mots sont morphologiquement liés.

Nous appliquons ces remarques de la manière suivante :

Soient :
- F_1 et F_2 deux familles ;
- R_1 le radical représentant F_1 ;
- R_2 le radical représentant F_2.

Les deux familles F_1 et F_2 sont regroupées dans une nouvelle famille F_3 ssi :

1. $R_1 = \alpha + s_1$ et $R_2 = \alpha + s_2$, où α est une chaîne de caractères de longueur minimale égale à 4 et s_1 et s_2 sont des chaînes de caractères différant au moins par leur premier caractère.
2. Il existe au moins un mot $m_1 \in F_1$ et un mot $m_2 \in F_2$ tels que m_1 et m_2 incluent le même préfixe.

Le radical R_3 de la nouvelle famille F_3 est le mot le plus court parmi R_1 et R_2.

Par exemple, si :

- F_1 = [océanique, intra-océanique] avec R_1 = *océanique* ;
- F_2 = [océaniques, sub-océaniques, intra-océaniques] avec R_2 = *océaniques*

alors il est possible de former une nouvelle famille :

$F_3 = F_1 \cup F_2$ = [océanique, intra-océanique, océaniques, sub-océaniques, intra-océaniques].

En effet, R_1 et R_2 partagent une chaîne initiale commune de longueur 9, *océanique*, et les mots *intra-océanique* de F_1 et *intra-océaniques* de F_2 ont en commun le préfixe *intra*. Le radical de F_3 est le radical le plus court, à savoir *océanique*.

2.4 Étape 3 : regroupement de familles à partir des signatures

La dernière étape de la classification consiste à utiliser la liste de signatures S donnée en entrée et à découvrir de nouvelles signatures à partir des regroupements opérés lors des étapes précédentes. Ces signatures vont permettre à leur tour d'effectuer de nouveaux regroupements, selon le principe du bootstrapping. Le processus se termine lorsqu'il n'est plus possible de découvrir de nouvelles signatures.

2.4.1 Découverte de nouvelles signatures

La découverte de nouvelles signatures se fait à partir des familles déjà constituées au cours des étapes précédentes. Les mots non préfixés de chaque famille sont comparés deux à deux afin d'obtenir une liste de signatures, selon la méthode suivante :

Soient m_1 et m_2 deux mots non préfixés appartenant à la famille F
tels que $m_1 = \alpha + s_1$ et $m_2 = \alpha + s_2$
avec $|\alpha| \geq 4$ et s_1 et s_2 des chaînes de caractères différant au moins par leur premier caractère.
Nous appellerons signature la paire de suffixes (s_1,s_2) et $sig(F, F)$ l'ensemble des signatures formées à partir d'une famille F, c'est-à-dire par comparaison bijective des mots non préfixés de F. Toutes ces signatures sont ajoutées à la liste des signatures S.

Prenons l'exemple de la famille suivante, formée lors des étapes 1 et 2 :

[trachyandésite, andésite, trachy-andésite, andésites, trachy-andésites, trachyandésites, andésitique, trachy-andésitique, trachyandésitique, trachy-andésitiques, trachyandésitiques, andésitiques].

La comparaison des graphies des mots non préfixés de cette famille conduit à l'identification des paires de suffixes suivantes : (ϵ, s), $(e, ique)$, $(e, iques)$, $(es, ique)$ et $(es, iques)$ (voir Figure 1).

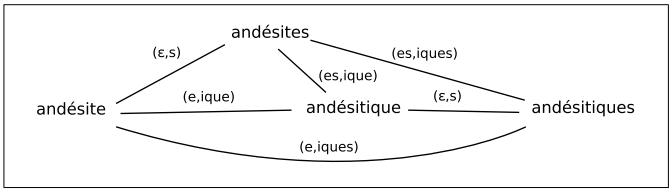

FIG. 1: Identification de signatures

2.4.2 Fusion de familles à l'aide des signatures

Les signatures ainsi acquises sont utilisées pour fusionner des familles. Le critère d'agglomération repose sur un indice p qui mesure la proportion de signatures valides partagées entre deux familles que l'on cherche à fusionner :

Soient :
- F_1 et F_2 deux familles ;
- l_1 le nombre de mots non préfixés de F_1 ;
- l_2 le nombre de mots non préfixés de F_2 ;
- S la liste de signatures fournies en entrée et découvertes à partir des familles déjà constituées.

$$p = \frac{|sig(F_1, F_2) \cap S|}{l_1 \cdot l_2}$$

Dans les expériences relatées dans la suite de cet article, nous avons fusionné deux familles lorsque $p \geq 0.5$.

Prenons l'exemple des familles représentées sur la Figure 2. Les signatures connues sont représentées par un arc plein tandis que les signatures inconnues sont représentées en pointillés. Ces deux familles sont fusionnées car le rapport du nombre de signatures connues sur le nombre total de signatures possibles est égal à 0.5.

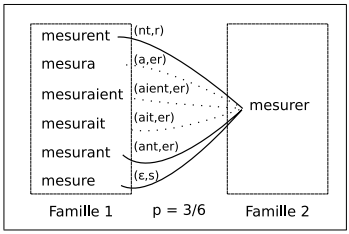

Fig. 2: Fusion de familles

Le dendrogramme de la Figure 3 illustre l'intérêt des regroupements effectués aux diverses étapes de la méthode. La seule famille formée à l'issue des deux premières étapes est [satellites, microsatellites, microsatellite, sous-satellite, satellite, mini-satellite]. L'étape de fusion de familles à partir des signatures partagées permet le regroupement de mots comme [satellitaire, satellitaires] ou [satellisation, satellisait].

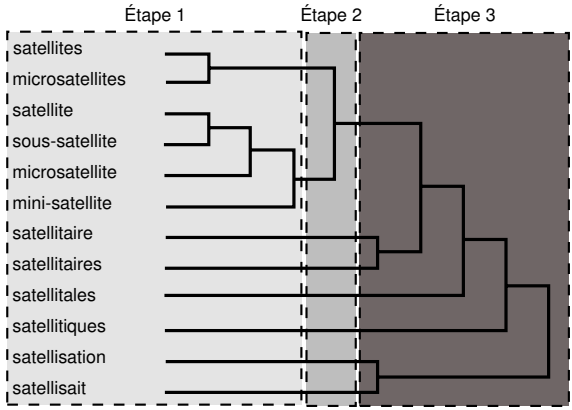

Fig. 3: Familles obtenues par classification à l'issue des trois étapes.

L'étape 3 d'agglomération à partir des signatures partagées est répétée tant que de nouvelles signatures sont acquises à partir des regroupements effectués et tant que ces signatures permettent de regrouper des familles. Le nombre de signatures différentes augmente fortement au cours des premières itérations, puis se stabilise. Le processus d'acquisition de nouvelles signatures, et par conséquent d'apprentissage de familles, s'achève au bout de 10 à 15 itérations.

3 Évaluation

Afin d'évaluer les résultats de la méthode, nous avons utilisé 4 corpus différents, en anglais et en français, couvrant deux domaines spécialisés distincts, la volcanologie et le cancer du sein. Dans la suite de cet article, ils seront désignés respectivement par `volcano-en`, `volcano-fr`, `cancer-en` et `cancer-fr`. Ces corpus ont été construits automatiquement à partir du Web en utilisant la méthode décrite dans (Baroni & Bernardini, 2004). Les listes de mots extraites de ces corpus comprennent entre 47 000 et 86 000 formes différentes.

3.1 Méthode d'évaluation

L'évaluation des résultats nécessite de disposer de familles morphologiques de référence auxquelles sont comparées les familles obtenues automatiquement par classification. Nous avons utilisé deux sources pour les familles de référence : nous avons d'une part élaboré manuellement des listes de référence et, pour l'anglais, nous avons extrait des familles de référence à partir des segmentations contenues dans la base CELEX (Baayen *et al.*, 1995). Les listes de référence construites manuellement contiennent des familles de mots pour le domaine du cancer du sein en français et en anglais. Elles contiennent 3 250 familles en anglais et 1 964 familles en français. Les familles morphologiques maximales de CELEX sont déterminées à partir des relations morphologiques de dérivation, de composition et de conversion, ce qui permet d'obtenir 14 880 familles de référence.

Nous avons évalué les familles induites par rapport aux familles de référence en utilisant les mesures proposées par (Schone & Jurafsky, 2000; Schone & Jurafsky, 2001). La méthode d'évaluation consiste à faire la somme des proportions de mots corrects (C), insérés (I) et supprimés (D) dans les familles morphologiques de tous les mots w de la liste d'évaluation. Si X_w est l'ensemble des mots appartenant à la famille morphologique d'un mot w selon le système à évaluer et Y_w est l'ensemble des mots appartenant à la famille morphologique de w selon CELEX ou toute autre base de référence, alors :

$$C = \sum_{\forall w} \frac{|X_w \cap Y_w|}{|Y_w|} \quad ; \quad D = \sum_{\forall w} \frac{|Y_w - (X_w \cap Y_w)|}{|Y_w|} \quad et \quad I = \sum_{\forall w} \frac{|X_w - (X_w \cap Y_w)|}{|Y_w|}$$

À partir de ces valeurs, il est également possible de calculer la précision, le rappel (et par conséquent la F-mesure) du système. La précision est égale à $C/(C+I)$ et le rappel à $C/(C+D)$.

3.2 Résultats obtenus

Les résultats sont détaillés dans la Table 1[2]. Ils démontrent la grande précision du système, qui est d'environ 80-90% suivant le corpus. De plus, malgré les contraintes imposées sur la longueur minimale des préfixes et des bases, le rappel est assez élevé et se situe autour

[2]Ces résultats ont été obtenus pour une valeur du paramètre N du module d'apprentissage des affixes égale à 10. N est un paramètre permettant de contrôler le processus d'apprentissage des affixes. Plus N est grand, plus le nombre de préfixes, de suffixes et, par conséquent, de signatures, est important.

de la barre des 70%. Ce résultat est d'autant plus remarquable que la méthode ne traite pas le cas des mots composés. La F-mesure varie peu sur l'ensemble des corpus, ce qui montre que les principes de regroupement utilisés sont valables aussi bien pour l'anglais que pour le français, et pour des domaines différents.

Référence	CELEX		Listes construites manuellement			
Corpus	cancer-en	volcano-en	cancer-en	volcano-en	cancer-fr	volcano-fr
Précision	79.3	81.4	89.8	91.2	91.5	93.1
Rappel	72.9	73.2	71.4	75.0	69.3	73.3
F-mesure	75.9	77.1	79.5	82.3	78.9	82.0

Tab. 1: Résultats obtenus pour les différents corpus et familles de référence.

3.3 Analyse des résultats

L'examen plus appronfondi des résultats montre que différents types de variantes sont groupés par l'algorithme :
- variantes orthographiques comme *tumor* (variante américaine) et *tumour* (variante britannique).
- variantes flexionnelles comme *traitement* et *traitements*.
- variantes dérivationnelles suffixées comme *traiter* et *traitement* et préfixées comme *auto-examen* et *examen*.
- composés savants comme *hormonothérapie* et *immunothérapie*. Il faut toutefois noter que dans ce cas, les chaînes *hormono* et *immuno* sont considérées comme des préfixes, car la méthode ne traite pas explicitement de la composition.

Malgré sa bonne précision, le système commet deux types d'erreur : sur-regroupement, c'est-à-dire le groupement de mots qui n'appartiennent pas tous à la même famille morphologique et sous-regroupement, c'est-à-dire l'absence de regroupement pour des mots appartenant à la même famille. La première erreur a pour conséquence de faire baisser la précision du système, tandis que la seconde conduit à une baisse du rappel. Ces erreurs peuvent survenir à toutes les étapes de la classification :

- À l'étape 1, malgré les contraintes imposées sur la longueur des préfixes et des mots, des mots peuvent être injustement considérés comme étant formés par combinaison d'un préfixe et d'un autre mot. Ainsi, le mot anglais *missing* est analysé comme étant la forme préfixée par *mis* du mot *sing*.
- À l'étape 2, il arrive que des familles soient fusionnées alors même qu'elles sont morphologiquement disjointes. Par exemple, la famille [médiane, paramédiane] est fusionnée avec la famille [socio-médical, paramédical, médical] car les mots *médiane* et *médical* commencent par la même chaîne de caractères de longueur 4 et apparaissent tous deux sous forme préfixée avec *para*.
- À l'étape 3 enfin, les contraintes imposées sur la longueur minimale de la base, égale à 4, peuvent empêcher le regroupement de certaines familles comme [île] et [îles], et donc induire une baisse du rappel.

4 Conclusion et perspectives

Nous avons présenté une méthode non supervisée d'acquisition de familles morphologiques. Malgré la simplicité de la méthode, les résultats obtenus sont très bons, notamment en terme de précision. L'analyse des résultats montre que les liens morphologiques découverts sont variés : flexion, dérivation et composition. De plus, l'apprentissage est effectué uniquement à partir d'une liste de mots et n'utilise aucune ressource externe. L'approche peut donc être directement appliquée à des langues et des domaines différents, à condition que les mots soient formés par concaténation linéaire de morphèmes.

Des évaluations complémentaires sont toutefois nécessaires. Nous n'avons pour l'heure testé le système que pour du vocabulaire issu de corpus de spécialité en français et en anglais. Il serait intéressant d'évaluer les performances pour d'autres langues plus complexes et pour du vocabulaire non technique. L'utilisation des préfixes aux deux premières étapes de la classification suppose que la langue traitée utilise ce procédé de formation. Or ce procédé n'est pas présent dans toutes les langues : le turc par exemple n'emploie que très peu de préfixes, ce qui rend les deux premières étapes du traitement inutiles. Reste alors à déterminer si le système est capable de produire des regroupements pertinents en utilisant uniquement les suffixes. On peut se poser une question similaire pour le vocabulaire moins technique, où le procédé de préfixation est utilisé moins fréquemment. Des expérimentations complémentaires pourront nous permettre de répondre à ces questions.

Les perspectives d'améliorations du système sont diverses. En effet, le système procède à une classification hiérarchique ascendante stricte, sans parenté multiple et donc sans possibilité pour un mot d'appartenir à deux voire à plusieurs familles différentes. Ceci est souhaitable pour les mots composés, qui font partie de plusieurs familles morphologiques. Il faudrait donc recourir à une forme de classification « floue ». Le système ne permet pas non plus la découverte de nouveaux préfixes, en complément de ceux injectés dans le système lors de la phase d'initialisation. On pourrait envisager d'appliquer une phase de bootstrapping similaire à celle qui permet la découverte de nouvelles signatures. De plus, il serait pertinent d'utiliser la fréquence des signatures au cours de la classification, afin de procéder aux regroupements correspondant aux signatures les plus fréquentes. Les informations contextuelles, disponibles dans les corpus, pourraient également permettre d'améliorer encore les résultats, notamment en terme de précision en validant le fusionnement de deux familles en fonction de la similarité de leurs contextes d'occurrence.

Les perspectives applicatives directement envisageables concernent la recherche d'information et la classification de documents. En recherche d'information, les familles morphologiques peuvent être utilisées pour l'extension de requêtes (Moreau & Claveau, 2006). Pour la catégorisation de documents, les familles peuvent servir de descripteurs des documents à classer (Witschel & Biemann, 2006).

Références

Baayen R. H., Piepenbrock R. & Gulikers L. (1995). *The Celex Lexical Database (Release 2) [CD-ROM].* Philadelphia, PA : Linguistic Data Consortium.

Baroni M. & Bernardini S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa & R. Silva,

Eds., *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, p. 1313–1316, Lisbon, Portugal.

Bernhard D. (2006). Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. In M. Kurimo, M. Creutz & K. Lagus, Eds., *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, p. 19–23, Venice, Italy.

Buitelaar P. & Sacaleanu B. (2002). Extending Synsets with Medical Terms. In *Proceedings of the First International WordNet Conference*, Mysore, India.

Creutz M. & Lagus K. (2002). Unsupervised Discovery of Morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, p. 21–30.

Creutz M. & Lagus K. (2005). Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland.

Gaussier E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Methods in Natural Language Processing* : University of Maryland.

Goldsmith J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, **27**(2), 153–198.

Grabar N. & Zweigenbaum P. (1999). Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In P. Amsili, Ed., *Actes de TALN 1999*, p. 175–184, Cargèse.

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

Lepage Y. (1998). Solving analogies on words : an algorithm. In *Proceedings of the 17th international conference on Computational Linguistics*, volume 1, p. 728–734, Morristown, NJ, USA : Association for Computational Linguistics.

Moreau F. & Claveau V. (2006). Extension de requêtes par relations morphologiques acquises automatiquement. In *Actes de la Troisième Conférence en Recherche d'Informations et Applications CORIA 2006*, p. 181–192.

Schone P. & Jurafsky D. (2000). Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. In *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop*, Lisbon, Portugal.

Schone P. & Jurafsky D. (2001). Knowledge-Free Induction of Inflectional Morphologies. In *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics*, p. 1–9.

Schwab D., Lafourcade M. & Prince V. (2005). Extraction semi-supervisée de couples d'antonymes grâce à leur morphologie. In *Actes de TALN 2005*, p. 73–82.

Witschel H. F. & Biemann C. (2006). Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization. In S. Werner, Ed., *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, volume 1, p. 197–204, Joensuu, Finland.

Zweigenbaum P. & Grabar N. (2000). Liens morphologiques et structuration de terminologie. In *Actes de IC 2000 : Ingénierie des Connaissances*, p. 325–334.