Segmentation morphologique à partir de corpus

Delphine Bernhard

Laboratoire TIMC-IMAG
Institut de l'Ingénierie et de l'Information de Santé
Pavillon Le Taillefer – Faculté de Médecine
F-38706 LA TRONCHE cedex
Delphine.Bernhard@imag.fr

Mots-clefs – Keywords

Segmentation morphologique, alignement de segments de mots, corpus.

Morphological segmentation, word segments alignment, corpus.

Résumé – Abstract

Nous décrivons une méthode de segmentation morphologique automatique. L'algorithme utilise uniquement une liste des mots d'un corpus et tire parti des probabilités conditionnelles observées entre les sous-chaînes extraites de ce lexique. La méthode est également fondée sur l'utilisation de graphes d'alignement de segments de mots. Le résultat est un découpage de chaque mot sous la forme (préfixe*) + base + (suffixe*). Nous évaluons la pertinence des familles morphologiques découvertes par l'algorithme sur un corpus de textes médicaux français contenant des mots à la structure morphologique complexe.

We describe a method that automatically segments words into morphs. The algorithm only uses a list of words collected in a corpus. It is based on the conditional probabilities between the substrings extracted from this lexicon. The method also makes use of word segments alignment graphs. As a result, all words are segmented into a sequence of morphs which has the following pattern: (prefix*) + base + (suffix*). We evaluate the morphological families discovered by the algorithm using a corpus of French medical texts containing words whose morphological structure is complex.

1 Introduction

L'analyse des mots en morphèmes, qui sont les plus petites unités porteuses de sens, facilite l'exécution de diverses tâches telles que la recherche d'informations (Hahn et al., 2003), la construction de dictionnaires (Lovis et al., 1995) ou de terminologies (Zweigenbaum, Grabar, 2000). Les méthodes existantes permettent de découvrir des suffixes flexionnels ou dérivationnels (Gaussier, 1999), voire également des préfixes (Déjean, 1998; Schone, Jurafsky, 2001; Goldsmith, 2001; Creutz, Lagus, 2002). Cependant, l'analyse finale d'un mot se limite généralement à 3 unités morphologiques au plus ((préfixe?) + base + (suffixe?)).

Certaines langues, comme l'allemand, et langues de spécialité, comme le vocabulaire médical, présentent des caractéristiques nécessitant une segmentation plus fine, notamment en raison du procédé de composition à la base de la formation des mots. La procédure de segmentation que nous proposons permet d'obtenir un découpage de chaque mot sous la forme (préfixe*) + base + (suffixe*) sans imposer de limite au nombre d'affixes. Pour cela, nous avons combiné l'utilisation de trois propriétés caractérisant les morphèmes et leurs frontières :

- Il existe une frontière morphémique lorsqu'il est difficile de prédire le segment suivant en fonction des segments précédents. Par exemple, la méthode proposée par (Harris, 1955) pour les phonèmes et appliquée par (Hafer, Weiss, 1974) et (Déjean, 1998) à la langue écrite utilise le nombre de phonèmes différents qui peuvent suivre une suite de phonèmes. Un nombre élevé de phonèmes indique une frontière entre deux morphèmes. Ainsi, dans notre corpus, seule 1 lettre, le "o", peut suivre la séquence initiale "micr" en français tandis que la séquence "micro" peut être suivie par 10 lettres différentes, marquant ainsi une frontière morphémique. Nous avons expérimenté une méthode similaire utilisant les probabilités conditionnelles observées entre les sous-chaînes extraites d'une liste de mots pour prédire les frontières morphémiques d'un mot (section 2.1).

- La similitude graphique est un indice de lien morphologique. En effet, les mots morphologiquement liés partagent une base identique et diffèrent par leurs affixes. La découverte des bases et des affixes peut donc se faire en comparant la graphie des mots, par exemple en recherchant la plus longue chaîne initiale commune (Gaussier, 1999; Zweigenbaum, Grabar, 2000). Dans de nombreux cas cependant la base ne correspond pas à la chaîne initiale, notamment pour les formes préfixées comme "antihormone" ou "précancéreux " par rapport aux formes non préfixées "hormone" et "cancéreux". Nous avons donc choisi de dissocier les phases d'apprentissage des affixes (section 2.2) et des bases (section 2.3). Nous utilisons également une structure de données (graphe) permettant d'aligner les mots à partir d'une base qui peut apparaître à toute position dans les mots comparés.

- Si l'on remplace les mots par la liste des morphèmes qu'ils contiennent, il est possible de comprimer les données du lexique. La meilleure segmentation d'un ensemble de mots est alors celle qui donne la représentation la plus compacte des données et qui réutilise un maximum de morphèmes. Ce principe est notamment utilisé par (Goldsmith, 2001) et (Creutz, Lagus, 2002). Ainsi, notre algorithme privilégie la réutilisation d'unités morphologiques apprises lors de la première phase de l'apprentissage (section 2.2).

Nous présentons dans un premier temps notre méthode de segmentation morphologique (section 2). Nous décrivons ensuite l'évaluation effectuée à partir d'une liste de mots extraits d'un corpus médical français (section 3). Enfin, nous discutons les résultats et proposons des possibilités d'évolution de l'algorithme (section 4).

2 Présentation de la méthode

Nous détaillons tout d'abord notre méthode de segmentation morphologique basée sur les probabilités conditionnelles (section 2.1) permettant de découvrir une liste d'affixes initiaux (section 2.2). Cette liste d'affixes est ensuite utilisée pour la découverte des bases et la segmentation finale (section 2.3).

2.1 Segmentation basée sur les probabilités conditionnelles

Nous pouvons à partir de chaque mot former un graphe directionnel dont les nœuds sont étiquetés par une sous-chaîne du mot. Il existe un arc entre deux nœuds si les sous-chaînes correspondantes se suivent immédiatement dans le mot. Pour tout chemin du graphe tel que le premier nœud n'a pas prédécesseur, la concaténation des sous-chaînes associées aux nœuds du chemin est équivalente au mot. Afin de faire la distinction entre les sous-chaînes apparaissant en début, milieu et fin de mot, les frontières de mots sont marquées par le caractère "#". Tout mot de longueur N est ainsi décomposable en un ensemble de sous-chaînes de longueur 1 (caractères) à N+2 (mot complet auquel s'ajoutent les marques de début et de fin "#"). La figure 1 représente une sous-partie du graphe des sous-chaînes du mot "microcalcifications".

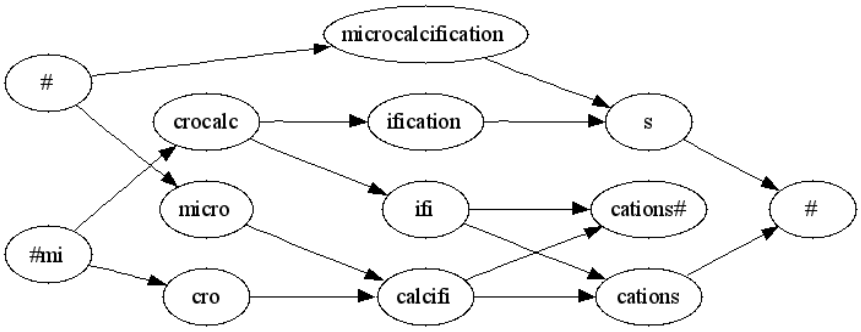


Figure 1 : Sous-partie du graphe des sous-chaînes du mot "microcalcifications".

Pour un arc du graphe reliant deux nœuds étiquetés par les sous-chaînes s_1 et s_2, nous pouvons estimer les probabilités conditionnelles $p(s_1|s_2)$ et $p(s_2|s_1)$ à partir du nombre d'occurrences dans le lexique de s_1, s_2 et de leur concaténation notée $s_1.s_2$:

$$p(s_1 | s_2) = \frac{f(s_1.s_2)}{f(s_2)} \quad \text{et} \quad p(s_2 | s_1) = \frac{f(s_1.s_2)}{f(s_1)}$$

Si les sous-chaînes s_1 et s_2 appartiennent à des morphèmes différents, $p(s_1|s_2)$ et $p(s_2|s_1)$ sont faibles car il est difficile de prédire la succession de s_1 et s_2 . A l'inverse, si elles appartiennent au même morphème, les probabilités conditionnelles sont plus fortes. Nous conservons le maximum de $p(s_1|s_2)$ et $p(s_2|s_1)$.

Pour chaque position dans le mot, nous obtenons ainsi un ensemble de valeurs correspondant aux maximums des probabilités conditionnelles entre les sous-chaînes se terminant à cette position et celles commençant à cette position. Le nombre de valeurs n'est pas le même pour toutes les positions. Afin d'analyser ces données et de les comparer, nous calculons à chaque position dans le mot le maximum, le quartile supérieur, la médiane, la moyenne, le quartile inférieur et le minimum[1] de ces valeurs. La figure 2 représente graphiquement ces données pour le mot "microcalcifications".

Nous pouvons observer des minimums locaux sur ces courbes à diverses positions dans le mot. Ces minimums indiquent des frontières morphémiques potentielles. Un minimum local constitue un point de segmentation potentiel si la différence avec le maximum précédent et le maximum suivant est au moins égale à un écart-type des valeurs de la courbe. Nous appliquons un deuxième critère afin de garantir la précision de la segmentation : une frontière morphémique est validée si elle correspond à des points de segmentation potentiels sur au moins la moitié des courbes. Les résultats ont montré qu'une courbe ne pouvait être

[1] La médiane et les quartiles inférieur et supérieur permettent de partager l'ensemble des valeurs en 4 sous-groupes de valeurs de même taille : 25 % des valeurs sont inférieures au quartile inférieur, 50 % sont inférieures à la médiane et 75 % sont inférieures au quartile supérieur.

privilégiée par rapport à une autre : nous avons donc appliqué la même pondération à chacune d'elles. Sur la figure 2, les frontières morphémiques validées sont indiquées par des flèches. Pour le mot "microcalcifications", la segmentation proposée est donc "micro + calcification + s".

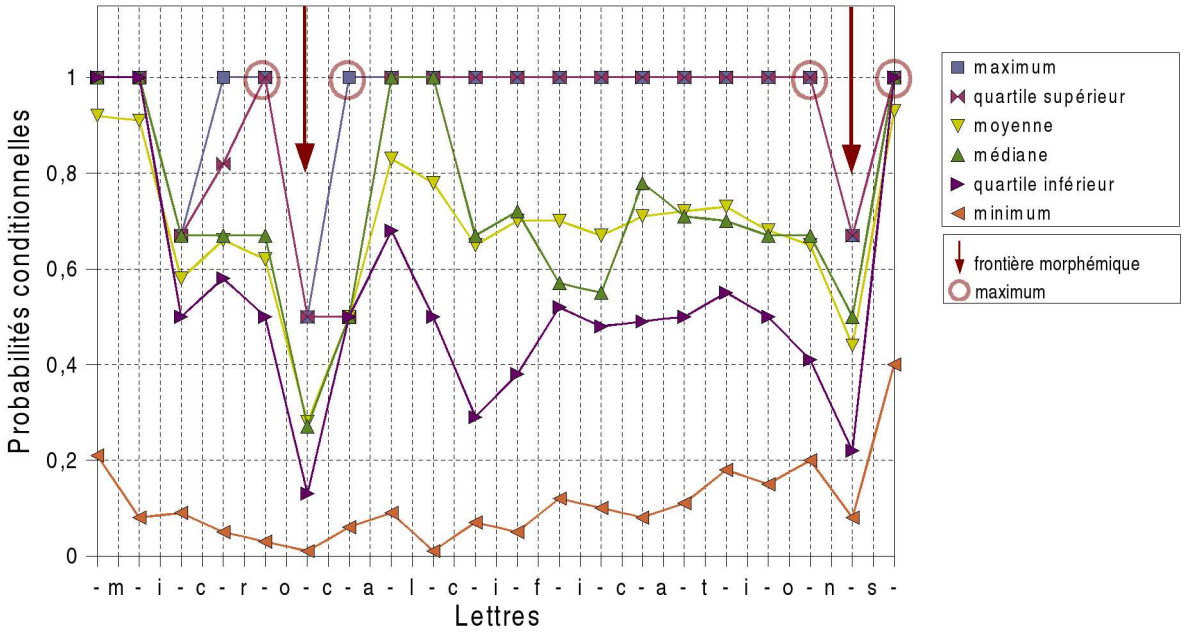


Figure 2 : Variation des probabilités conditionnelles pour le mot "microcalcifications".

Cette méthode permet d'identifier des frontières morphémiques mais n'est pas suffisante pour reconstituer des familles morphologiques. Elles peut donner lieu à des segmentations erronées notamment pour les mots courts ("foyer" donne "fo + yer", "mêlé" : "mê + lé") ou lorsque l'on trouve peu de mots de la même famille dans le corpus ("incubées" : "incu + bées"). Le tableau 1 donne les segmentations obtenues pour un ensemble de mots appartenant à la famille de "microcalcifications". On constate l'absence de segmentation des mots les plus courts et du segment "calcification", ainsi qu'une segmentation erronée du préfixe "macro" en "ma + cro". Cette dernière peut s'expliquer par une prépondérance du préfixe "micro" (25 occurrences) par rapport au préfixe "macro" (13 occurrences). De plus, en l'absence de base commune, il n'est pas possible de reconstituer la famille morphologique. Cette méthode de segmentation est donc utilisée uniquement pour obtenir une liste d'affixes initiaux qui sera réutilisée pour la découverte des bases et la segmentation finale des mots.

Mots	Segmentation
calcifier	calcifier
calcifiée	calcifiée
calcifiés	calcifi + és
calcifiées	calcifiées
calcifications	calc + ification + s
microcalcification	micro + calcification
micro-calcifications	micro- + calcification + s
macrocalcifications	ma + cro + calcification + s

Tableau 1 : Exemples de segmentations obtenues en utilisant les probabilités conditionnelles.

2.2 Découverte des affixes initiaux

La sélection des affixes valides peut se faire en utilisant un critère de fréquence (Déjean, 1998) : dans ce cas, seuls sont conservés les affixes qui dépassent un certain nombre d'occurrences après la phase d'apprentissage. Nous proposons un autre critère de sélection des affixes initiaux. En effet, les mots morphologiquement complexes sont généralement plus longs que les mots morphologiquement simples. L'apprentissage des affixes initiaux n'est donc effectué que sur les mots les plus longs du lexique et non pas sur l'ensemble des mots. Le nombre de mots du lexique d'apprentissage est paramétrable (seules quelques centaines de mots sont nécessaires).

Nous segmentons chacun des mots du lexique d'apprentissage en utilisant la méthode de segmentation basée sur les probabilités conditionnelles décrite dans la section 2.1. Dans la mesure où la segmentation met à jour aussi bien des préfixes que des suffixes, il n'est pas possible de déterminer le type (préfixe, suffixe ou base) d'un segment uniquement en utilisant des critères positionnels. (Vergne, 2003) utilise les différences de fréquence et de longueur pour distinguer mots vides (fréquents et courts) et mots pleins (rares et longs) dans un énoncé. Nous pouvons établir un parallèle entre mots vides et affixes d'une part et mots pleins et bases d'autre part. En effet, une base est généralement moins fréquente et plus longue qu'un affixe : la combinaison de ces deux critères nous permet de repérer une pseudo-base[2] parmi les segments proposés. Les pseudo-bases identifiées par cette méthode correspondent aux segments soulignés dans le tableau 1. Le tableau 2 donne quelques exemples de mise en oeuvre de cette méthode. Dans le cas du mot "chimio-hormonothérapie", effectif minimal (2 : "chimio") et longueur maximale (9 : "othérapie") ne correspondent pas, la segmentation est alors considérée comme non valide.

Segments	micro	calcification	s
Effectifs	37	6	8289
Longueurs	5	13	1
Segments	multi	dimension	nelle
Effectifs	32	5	41
Longueurs	5	9	5
Segments	chimio-	hormon	othérapie
Effectifs	2	29	25
Longueurs	7	6	9

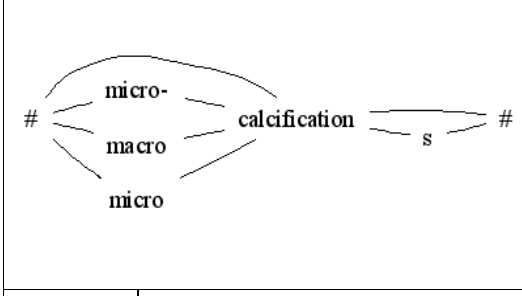
Tableau 2 : Repérage de la base parmi des segments. Les cases grisées correspondent respectivement à l'effectif minimal et à la longueur maximale.

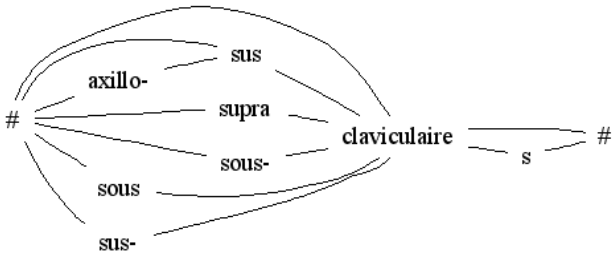
Afin d'augmenter le nombre d'affixes extraits nous recherchons la pseudo-base dans l'ensemble des mots du lexique et nous procédons à l'alignement de ces mots en les insérant dans un graphe. La pseudo-base sert de point d'ancrage. Nous appliquons des critères supplémentaires pour vérifier la validité de la pseudo-base en fonction des mots dans lesquels elle apparaît :

[2] Nous réutilisons ici le terme "pseudo-base" employé par (Schone, Jurafsky, 2001). Il peut en effet s'agir d'une base non valide ou contenant des affixes.

- Le nombre de mots de cette liste doit être supérieur ou égal à 2.
- La pseudo-base doit débuter un de ces mots. Par exemple, la segmentation du mot "radiopharmaceutiques" est "radio + pharmac + eutiques". Le segment "eutiques" est reconnu comme étant la pseudo-base car il est à la fois le plus long et le moins fréquent. Les 5 mots contenant "eutiques" sont : "chimiothérapeutiques", "postthérapeutiques", "pharmaceutiques", "radiopharmaceutiques", et "thérapeutiques". Aucun des mots du corpus ne commence par "eutiques", la pseudo-base est donc considérée comme non valide.

La figure 3 donne deux exemples d'alignement à partir des pseudo-bases "calcification" et "claviculaire" ainsi que les préfixes et les suffixes découverts par l'alignement ("#" indique le préfixe ou suffixe vide). Seuls sont conservés les préfixes de longueur supérieure ou égale à deux. L'ensemble des préfixes et suffixes obtenus à la fin de cette étape constitue la liste des affixes initiaux.





préfixes	micro-, micro, macro, #	préfixes	axillo-, sus, sus-, sous, sous-, supra, axillo-sus, #
suffixes	s, #	suffixes	s, #

Figure 3 : Alignement des mots contenant les pseudo-bases "calcification" et "claviculaire".

2.3 Découverte des bases et segmentation des mots

Cette dernière phase permet de segmenter tous les mots du corpus en réutilisant les affixes découverts lors de la phase précédente pour obtenir une liste des bases présentes dans le corpus :

1. Pour chaque mot du lexique, recherche des affixes initiaux qu'ils contient. Si l'on retranche d'un mot les diverses combinaisons possibles de ces affixes et de la chaîne vide, il est possible d'obtenir des pseudo-bases. Le mot lui-même constitue une pseudo-base, ce qui permet notamment de conserver dans la liste des pseudo-bases les mots non fléchis. Ces pseudo-bases doivent avoir une longueur minimale de 3. Par exemple, le mot "calcifiées" contient les suffixes initiaux "iées", "ées", "es" et "s". Les pseudo-bases obtenues en retranchant ces suffixes initiaux sont : "calcifiée", "calcifié", "calcifi" et"calcif", auxquelles s'ajoute "calcifiées" (mot complet).

2. Pour chaque pseudo-base, recherche des mots contenant la pseudo-base dans le lexique. Par exemple, les mots contenant la pseudo-base "calcifié" sont : "calcifié", "calcifiée", "calcifiés", "calcifiées".

3. Construction du graphe d'alignement des mots contenant la pseudo-base. Nous appliquons deux contraintes sur le graphe d'alignement : il doit contenir un nombre suffisant d'affixes initiaux et le nombre de mots contenant la pseudo-base ne doit pas excéder un certain seuil. Nous avons fixé expérimentalement la proportion d'affixes initiaux à 75 % et le nombre maximal de mots à 50. Afin d'éviter l'absence d'analyse lorsque ces critères ne sont pas vérifiés, deux tentatives supplémentaires sont effectuées : nous supprimons dans un premier temps les préfixes inconnus du graphe d'alignement, puis, si l'alignement n'est toujours pas considéré comme valide, nous supprimons les suffixes inconnus. La meilleure pseudo-base est celle dont le graphe d'alignement contient le plus grand nombre d'affixes initiaux. Les mots contenant la meilleure pseudo-base sont alors segmentés en fonction de leur alignement. Le tableau 3 représente les pseudo-bases obtenues à partir du mot "calcifiées" ainsi que les segmentations correspondantes. La meilleure pseudo-base est "calcifi".

calcifiées	calcifi	calcifiée	calcif	calcifié
calcifiées	calcifi**+er**	calcifiée	calcif+ier	calcifié
	calcifi**+é**	calcifiée+**s**	calcif+ié	calcifié+**e**
	calcifi**+é+e**		calcif+ié+**e**	calcifié+**s**
	calcifi**+é +s**		calcif+ié+**s**	calcifié+**e+ s**
	calcifi**+é+e+s**		calcif+ié+**e**+**s**	
	calcifi**+cation**		calcif+ication	
	calcifi**+cation+s**		calcif+ication+**s**	
	micro+calcifi**+cation**		**micro+**calcif+ication	
	micro+calcifi**+cation+s**		**micro+**calcif+ication+**s**	
	micro-+calcifi**+cation+s**		**micro-+**calcif+ication+**s**	
	macro+calcifi**+cation+s**		**macro+**calcif+ication+**s**	

Tableau 3 : Pseudo-bases et segmentations obtenues à partir du mot "calcifiées". Les affixes initiaux sont marqués en gras. Les segmentations validées se trouvent dans la colonne grisée.

3 Évaluation

Nous avons évalué la méthode sur un corpus composé de 80 documents traitant du cancer du sein. Ce corpus se compose de 33 articles scientifiques et de 47 pages web. Il comprend environ 280 000 mots pour 12 587 formes différentes. Nous évaluons la validité des bases associées à chaque mot et non pas la position des points de segmentation. L'ensemble des mots qui contiennent la même base forme une famille morphologique. Nous vérifions si deux mots associés à la même base par l'algorithme (comme c'est le cas pour "microcalcification" et "calcifier") sont effectivement morphologiquement liés. Nous avons automatisé l'évaluation des variantes flexionnelles en utilisant le fichier DLF (formes simples) produit par INTEX (Silberztein, 1993) par l'application du dictionnaire DELAF. Dans ce fichier, chaque forme est associée à un ou plusieurs lemmes. Si deux formes sont associées au même lemme dans le fichier DLF et à la même base par l'algorithme, la relation entre les deux formes est considérée comme valide (relation flexionnelle). L'analyse du lexique par INTEX produit

12 116 relations binaires entre mots du corpus. En ce qui concerne les mots reliés par dérivation ou composition ainsi que les mots ne figurant pas dans le fichier DLF, l'évaluation a été faite manuellement.

Le tableau 4 présente les résultats de l'évaluation en fonction de la taille du lexique d'apprentissage et le tableau 5 donne quelques exemples de familles morphologiques obtenues pour un lexique d'apprentissage de 100 mots. La précision décroît avec l'augmentation du nombre de mots du lexique d'apprentissage tandis que le rappel des relations flexionnelles augmente pour culminer à environ 75 %. Pour 600 mots dans le lexique d'apprentissage, le rappel n'est pas meilleur, ce qui semble indiquer un seuil limite dans l'apprentissage.

Nombre de mots du lexique d'apprentissage	100	200	300	400	500
Préfixes initiaux	35	69	95	113	147
Suffixes initiaux	18	77	112	138	177
Nombre de bases	9 389	6 989	6 740	6 224	5 336
Nombre total de relations	5 284	15 794	17 610	21 889	33 837
Relations flexionnelles (rappel)	3 436 (28,4%)	7 319 (60,4%)	7 291 (60,2%)	8 116 (67,0%)	9 261 (76,4%)
Relations dérivationnelles et compositionnelles valides	1 740	7 145	8 343	10 524	14 807
Relations non valides	108	1 330	1 976	3 249	9 769
Précision	98,0%	91,6%	88,8%	85,2%	71,1%

Tableau 4 : Résultats de l'évaluation.

Bases	Variantes
pathologi	anatomo-pathologie, anatomo-pathologique, anatomo-pathologiques, anatomo-pathologiste, anatomopathologie, anatomopathologique, anatomopathologiques, anatomopathologiste, anatomopathologistes, cytopathologique, cytopathologiste, histopathologie, histopathologique, histopathologiques, pathologie, pathologies, pathologique, pathologiques, pathologiste, pathologistes, physio-pathologique, physiopathologiques, radio-pathologiques, radiologique-anatomopathologique, radio-pathologique, radiopathologiques.
thérapi	chimiothérapie, chimiothérapies, chimiothérapique, chimiothérapiques, hormono-chimiothérapique, physiothérapie, polychimiothérapie, radio-chimiothérapie, radiothérapie, radiothérapique, thérapie, thérapies.

Tableau 5 : Exemples de familles morphologiques obtenues à partir d'un lexique d'apprentissage de 100 mots.

4 Discussion et conclusion

L'algorithme que nous avons mis au point permet de traiter à la fois les procédés de flexion ("pathologique", "pathologiques"), dérivation ("pathologie", "pathologique") et composition ("anatomopathologique"). De plus, il est possible de repérer les préfixes et les suffixes au cours de la même procédure, grâce à l'alignement des segments de mots à partir d'une base qui peut se situer à n'importe quelle position dans le mot.

La taille du corpus a été volontairement limitée afin de permettre une validation manuelle des mots reliés par dérivation ou composition. Les méthodes d'apprentissage sur corpus sont généralement dépendantes de la taille du corpus d'apprentissage : plus le corpus est important, meilleurs sont les résultats. Nous obtenons néanmoins un très bon pourcentage de relations valides (entre 98 et 71 %) avec uniquement 12 587 formes différentes (le français compte plusieurs centaines de milliers de formes différentes).

Nous n'avons pas directement mesuré le rappel. Nous en avons une indication indirecte à partir du nombre de relations découvertes et du rappel des relations flexionnelles du fichier DLF. Ce rappel est assez bon (de l'ordre des 60 à 70 %) et ce d'autant plus qu'un certain nombre de relations flexionnelles ne peuvent être retrouvées par un algorithme de segmentation morphologique (c'est le cas notamment des verbes irréguliers dont des formes très différentes peuvent être ramenées au même lemme; par exemple les formes "veut", "veux", "voudrez", "voudront", "voulant", "vouloir", "voulu" correspondent toutes au même lemme "vouloir"). D'une manière générale, les variantes des bases et les cas de doublement des consonnes constituent des pierres d'achoppement qui peuvent diminuer le rappel de l'algorithme. Ainsi, les mots "estrogènes" et "estrogénique" ont deux bases différentes : "estrogèn" et "estrogéniqu". De même "fractions" et "fractionnement" correspondent aux bases "fraction" et "fractionne". La méthode devrait être améliorée pour prendre ces cas en compte.

Il faut également noter que cette méthode n'est pas a priori limitée à l'analyse du français car elle ne repose sur aucune ressource externe au corpus. Nous avons d'ailleurs obtenu quelques exemples de segmentations de mots anglais, dans la mesure où ces mots contiennent des segments communs au français et à l'anglais (par exemple "radio + therapy"). Il reste à tester la méthode sur des corpus étrangers, notamment des corpus de langues agglutinantes telles que l'allemand, afin d'éprouver son efficacité. Des expériences supplémentaires devraient également être conduites sur des corpus français généralistes et de taille plus importante.

La méthode de segmentation finale des mots décrite utilise une structure de mot simplifiée : (préfixe*) + base + (suffixe*). Nous obtenons par exemple les segmentations suivantes : "lymphangi + osarcome", "lymphangi + te" et "ostéo + sarcom + e". Il est intéressant de noter que l'interfixe "o" reliant "lymphangi" et "sarcom" a bien été repéré. Cependant, seule une base, "lymphangi", a été identifiée, "osarcome" étant considéré comme un suffixe. Il serait souhaitable d'affiner l'algorithme sur ce point afin de repérer plusieurs bases par mot, comme c'est souvent le cas pour les termes spécialisés. Le repérage de plusieurs bases par mot permettrait de mettre le mot "lymphangiosarcome" en relation à la fois avec "lymphangite", "lymphe" ou "ostéosarcome" par le biais des bases "lymph" et "sarcom".

De plus, à l'heure actuelle, l'algorithme utilise uniquement des critères formels pour procéder à la segmentation morphologique des mots. Or les relations morphologiques impliquent à la fois forme et sens. Il paraît donc nécessaire d'intégrer des informations d'ordre sémantique à une méthode de segmentation morphologique. La prise en compte d'informations sémantiques

peut se faire a priori en effectuant l'apprentissage à partir de mots sémantiquement reliés telles que les listes de termes synonymes issus d'une terminologie (Zweigenbaum, Grabar, 2000) ou a posteriori en utilisant des mesures de distance sémantique basées sur la distribution des mots dans le corpus pour valider les segmentations (Schone, Jurafsky, 2001; Zweigenbaum et al., 2003). Nous envisageons de compléter l'algorithme avec des informations de ce type afin d'en augmenter la précision.

Références

Creutz M., Lagus K. (2002), Unsupervised Discovery of Morphemes, *Proceedings of the 6th Meeting of the ACL Special Interest Group of Computaional Phonology (SIGPHON)*, 21-30.

Déjean H. (1998), Morphemes as Necessay Concept for Structures Discovery from Untagged Corpora, *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning*, 295-298.

Gaussier E. (1999), Unsupervised Learning of Derivational Morphology from Inflectional Lexicons, *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, 24-30.

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

Hafer M.A, Weiss S.F. (1974), Word Segmentation by Letter Successor Varieties, *Information Storage and Retrieval*, Vol. 10, pp. 371-385.

Hahn U., Honeck M., Shulz S. (2003), Subword-Based Text Retrieval, *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, 108.1.

Harris Z. (1955), From Phoneme to Morpheme, *Language*, Vol. 31, pp. 190-222.

Lovis C., Michel PA., Baud, R., Scherrer JR.. (1995), Word Segmentation Processing: A Way to Exponentially Extend Medical Dictionaries, *Proceedings of the 8th World Congress on Medical Informatics*, 28-32.

Schone P., Jurafsky D. (2001), Knowledge-Free Induction of Inflectional Morphologies, *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, 183-191.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Paris, Masson.

Vergne J. (2003), Un outil d'extraction terminologique endogène et multilingue, Actes de *TALN 2003*, 139-148.

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

Zweigenbaum P., Hadouche F., Grabar N. (2003), Apprentissage de relations morphologiques en corpus, Actes de *TALN 2003*, 285-294.