# Analyse morphologique non supervisée en domaine biomédical

## Application à la recherche d'information

**Vincent Claveau*** — **Ewa Kijak****

*IRISA – * CNRS – ** Univ. Rennes 1*
*Campus de Beaulieu, 35042 Rennes, France*
vincent.claveau@irisa.fr ewa.kijak@irisa.fr

RÉSUMÉ. *Dans le domaine biomédical, utiliser des termes spécialisés est essentiel pour accéder à l'information. Cependant, dans beaucoup de langues, ces termes sont des constructions morphologiques complexes qui compliquent cet accès à l'information. Dans cet article, nous nous intéressons à l'identification des composants morphologiques de ces termes et à leur utilisation pour une tâche de recherche d'information (RI). Nous proposons différentes approches reposant sur un alignement automatique avec une langue pivot particulière, le japonais, et sur un apprentissage par analogie permettant de produire des analyses morphologiques fines des termes d'une langue donnée. Ces analyses morphologiques sont ensuite utilisées pour améliorer l'indexation de documents biomédicaux. Les expériences rapportées montrent la validité de cette approche avec des gains en MAP de plus de 10 % par rapport à un système de RI standard.*

ABSTRACT. *In the biomedical field, using of specialized terms is key to access information. However, in most Indo-European languages, these terms are complex morphological structures. The presented work aims at identifying the various meaningful components of these terms and use them to improve biomedical Information Retrieval (IR). We present different approaches combining automatic alignments with a pivot language, Japanese, and analogical learning that allows an accurate morphological analysis of terms. These morphological analysis are used to improve the indexing of medical documents. The experiments reported in this paper show the validity of this approach with a 10% MAP improvement over a standard IR system.*

MOTS-CLÉS : *morphologie, terminologie biomédicale, alignement, apprentissage par analogie, indexation morphosémantique, recherche d'information biomédicale.*

KEYWORDS: *morphology, biomedical terminology, alignment, analogical learning, morphosemantic indexing, biomedical information retrieval.*

## 1. Introduction

Dans le domaine biomédical, les terminologies et leurs multiples usages sont au cœur de nombreuses applications. Elles sont utilisées pour structurer les connaissances mais aussi plus pragmatiquement pour indexer et formaliser l'information contenue dans des documents du domaine. L'exemple le plus connu est sans doute la terminologie MeSH® (Medical Subject Headings, `www.nlm.nih.gov/mesh`) qui a été développée pour indexer les articles de journaux spécialisés des bases PubMED®/MEDLINE® (`www.pubmed.gov`). Les termes recensés dans ces terminologies ont des caractéristiques bien particulières. Ainsi, dans la plupart des langues indo-européennes, les termes biomédicaux sont des compositions complexes de plusieurs constituants, souvent d'origine latine ou grecque. Cette complexité morphologique est un point important à prendre en compte pour les opérations basiques (manipulation, traduction, mise en relation sémantique de ces termes), et de ce fait, pour des opérations de haut niveau telles que la traduction artificielle, ou comme nous le montrons dans cet article, pour la recherche d'information (RI).

Dans cet article, nous nous intéressons au développement de techniques d'analyse morphologique pour les termes biomédicaux et nous montrons comment un système de RI du domaine peut en bénéficier. Plus précisément, nous présentons une technique permettant de segmenter un terme en ses constituants morphologiques, à savoir des morphes[1] tout en leur assignant des informations sémantiques. Au contraire de travaux existants (Deléger *et al.*, 2008 ; Markó *et al.*, 2005a, par exemple) résolument fondés sur une expertise humaine, nos techniques utilisent des approches non supervisées.

L'idée originale au cœur de notre approche est d'utiliser l'aspect multilingue des bases terminologiques existantes. Nous utilisons le japonais comme langue pivot, et plus précisément les termes en kanjis, pour aider à la décomposition des termes d'autres langues. De manière entièrement automatique, ceux-ci sont découpés en morphes et chaque morphe est associé aux kanjis correspondants. Les kanjis jouent donc un rôle de représentation sémantique pour les morphes. L'intérêt principal des kanjis à ce titre est que les termes japonais peuvent être vus comme une concaténation de mots élémentaires indépendants (cf. section 2.2 pour une discussion détaillée des intérêts de cette représentation pivot) ; ils peuvent même être trouvés dans un dictionnaire généraliste. Par exemple, le terme photochemotherapy peut être traduit en japonais par 光化学療法 ; la décomposition et l'alignement de ces deux termes donnent :

– photo ↔ 光 ('lumière') ;
– chemo ↔ 化学 ('chimie', 'médicament') ;
– therapy ↔ 療法 ('thérapie').

1. Dans cet article, à l'image de Mel'čuk (2006), nous distinguons le morphe, signe linguistique élémentaire dont le signifiant est un segment de la chaîne parlée, du morphème, unité abstraite vue dans ce travail comme une classe d'équivalence de morphes partageant un signifié identique et un signifiant proche, c'est-à-dire de réalisations sémantiquement équivalentes et formellement proches (Anderson, 2013, pour une vue exhaustive et historique des différentes écoles de définitions linguistiques du morphème).

Comme nous le voyons ici, chaque morphe est associé à des kanjis pouvant être utilisés comme des descripteurs plus adaptés à des problèmes d'indexation que le terme complet initial.

Dans cet article[2], nous présentons une technique automatique pour décomposer et assigner les kanjis correspondant à ces termes biomédicaux. Cette technique ne nécessite aucune intervention humaine mais repose sur la disponibilité de terminologies bilingues comme celles contenues dans l'UMLS. Cette analyse morphologique, et l'indexation des documents qu'elle permet, repose donc principalement sur l'étape d'alignement entre morphes et kanjis. Cet alignement est effectué avec une technique originale, adaptée aux données manipulées, et reposant sur un algorithme *Forward-Backward* et sur l'apprentissage par analogie. Nous montrons ensuite que ces correspondances entre morphes et kanjis, une fois obtenues automatiquement, peuvent être exploitées de différentes manières pour améliorer les résultats d'un système de RI.

L'article est structuré comme suit. Après un examen des travaux connexes en section 2, nous présentons successivement la technique d'alignement et ses résultats en sections 3 et 4. L'utilisation des décompositions morphologiques obtenues dans un cadre de RI est ensuite expliquée en section 5. Avant de conclure, nous détaillons en section 6 les évaluations conduites sur une collection de RI biomédicale.

## 2. Contexte scientifique

### 2.1. *Travaux connexes*

Différents travaux s'appuient sur la morphologie à des fins d'analyse terminologique. C'est plus particulièrement le cas dans le domaine biomédical. D'une part, les terminologies y jouent un rôle central pour beaucoup d'applications, et d'autre part, les termes sont le plus souvent construits par composition que l'on dit néoclassique. On a par exemple des termes comme magnétoencéphalographie que l'on peut décomposer en trois morphes : magnéto/encéphalo/graphie. La nature des constituants impliqués (Iacobini, 1999 ; Dal et Amiot, 2008) et des règles régissant ces compositions (Dal et Amiot, 2008 ; Fradin, 2005) en font en linguistique des objets d'étude particuliers. C'est aussi le caractère productif (de nombreux néologismes sont ainsi construits) et régulier (les mêmes règles et constituants sont utilisés dans ces néologismes) de ce type de composition qui en fait un phénomène important à traiter d'un point de vue applicatif.

Il existe quelques bases informatiques contenant des informations morphologiques (par exemple, pour le français le dictionnaire Biotop[3], et pour l'anglais le diction-

2. Une partie des travaux présentés dans cet article a fait l'objet de publications en conférences : décomposition par alignement (Claveau et Kijak, 2010 ; Claveau et Kijak, 2011), application à la recherche d'information (Claveau et Kijak, 2012 ; Claveau, 2012).

3. georges.dolisi.free.fr/

naire Dorland's[4]). Leur couverture est cependant loin d'être exhaustive, et le processus d'analyse morphologique d'un terme à partir de telles listes reste un problème, même en supposant l'existence de bases plus complètes. L'approche présentée dans cet article se veut une réponse originale à ces deux écueils. Pour autant, différents travaux sur l'analyse morphologique peuvent être mis en regard des nôtres.

On peut distinguer deux visions de l'utilisation de la morphologie comme outil d'analyse (lexicale ou terminologique). Dans la vision lexématique, la forme des termes est exploitée pour construire ou découvrir les relations qu'ils entretiennent entre eux mais sans avoir recours nécessairement à leur décomposition en morphes (Grabar et Zweigenbaum, 2002 ; Claveau et L'Homme, 2005 ; Hathout, 2009, par exemple). À l'opposé de cette utilisation implicite de la morphologie, la vision morphémique repose essentiellement sur une première étape d'analyse en morphes des termes. Beaucoup de travaux relèvent de ce cadre et peuvent se distinguer selon leur besoin d'expertise. Il y a d'une part les techniques automatiques, cherchant le plus souvent les séquences ou patrons récurrents de lettres, considérés comme des morphes, dans des listes de mots (Creutz et Lagus, 2005 ; Kurimo *et al.*, 2010, *inter alia*). Cependant, de telles techniques ne peuvent associer aucune connaissance sémantique aux morphes découverts. Il y a d'autre part les approches expertes, comme les travaux déjà mentionnés de Namer (2007), Deléger *et al.* (2008) ou Markó *et al.* (2005a) dans lesquels les morphes, leurs informations sémantiques et leurs règles de composition sont fournis soit par un expert, soit par des heuristiques, là aussi fournies manuellement (Baud *et al.*, 1999). Selon leur visée applicative, ces dernières techniques se distinguent par leur finesse d'analyse, allant de la segmentation et l'étiquetage avec des identifiants interlingues (Markó *et al.*, 2005a) à la décomposition hiérarchique et l'interprétation du composé (Namer, 2007). À notre connaissance, notre approche est la première à relever d'une approche automatique tout en fournissant des éléments d'analyse que l'on peut qualifier de morphosémantiques grâce aux informations associées à chaque morphe.

La construction de liens terminologiques à partir d'éléments d'analyse morphologique, comme nous le proposons en section 5, est également un domaine exploré par plusieurs auteurs. Ces méthodes se distinguent par l'emploi et le type d'expertise utilisée. Ainsi, les relations entre morphes peuvent être directement définies par l'expert (Namer et Zweigenbaum, 2004) et même typées ; Namer (2007) définit par exemple des relations de synonymie, hyponymie, méronymie et approximation en s'appuyant sur les relations lexicales définies dans les terminologies. Le coût de développement et la couverture de ces approches sont cependant des points critiques rarement discutés et évalués. D'autres approches, reposant sur une expertise censément plus légère, consistent à utiliser des règles heuristiques (Jacquemin, 1997 ; Daille, 2003). Les approches par apprentissage, incluant celles fondées sur les analogies formelles (cf. section 3.2), ont permis de déporter ce besoin d'expertise en une sélection d'exemples intéressants (Krovetz, 1993 ; Grabar, 2004 ; Claveau et L'Homme, 2005), voire à s'en passer (Gaussier, 1999 ; Moreau *et al.*, 2007 ; Hathout, 2009).

4. `www.dorlands.com/wsearch.jsp`

D'un point de vue plus technique, l'utilisation que nous faisons de terminologies bilingues évoque aussi les travaux sur la translitération, en particulier de l'arabe ou de l'alphabet katakana, que ce soit pour la traduction directe (Knight et Graehl, 1998, par exemple), ou la recherche de traduction (Chiao et Zweigenbaum, 2002 ; Tsuji *et al.*, 2002). Dans ce cadre, citons les travaux de Morin et Daille (2010). Ils proposent de faire correspondre les termes en kanjis aux termes en français à l'aide de règles morphologiques. Cependant, là encore, ces quelques règles sont proposées par un expert, et ne concernent qu'un cas particulier de dérivation. Une telle approche n'est par ailleurs pas adaptée aux composés néoclassiques. Dans d'autres travaux, des techniques de traduction de termes biomédicaux ont été proposées (Claveau, 2009 ; Harastani *et al.*, 2012, *inter alia*). Même si le but est différent ici, ces approches partagent des similitudes avec les nôtres. En effet, elles considèrent les termes comme des séquences de lettres et requièrent des alignements. Ces alignements sont le plus souvent faits avec des algorithmes 1-1, c'est-à-dire seulement capables d'aligner un caractère (éventuellement vide) de la langue source avec un caractère de la langue cible. Cependant, dans des travaux plus récents sur la phonétisation (Jiampojamarn *et al.*, 2007), les auteurs ont montré l'intérêt de considérer des alignements *many-to-many*.

Sur l'aspect applicatif de cet article, à savoir l'utilisation d'une analyse morphologique dans un contexte de recherche d'information, la littérature est plus fournie (Moreau et Sébillot, 2005, pour un panorama). Bien que les résultats dépendent de nombreux facteurs (langue, outil morphologique, taille de la collection, domaine...), il y a un consensus large sur l'intérêt de processus simples comme la racinisation (*stemming*). En effet, les raciniseurs sont disponibles dans de nombreuses langues, sont conceptuellement simples même si leur implémentation s'éloigne souvent de la réalité linguistique, et améliorent les performances des systèmes de RI. La lemmatisation, plus rarement utilisée en RI, améliore également les résultats dans la plupart des cas. Il faut noter que les seuls phénomènes morphologiques visés par ces outils, tels qu'ils sont implémentés, sont la flexion et la dérivation. Comme ils fonctionnent uniquement à partir d'opérations simples sur les préfixes et les suffixes (au sens informatique du terme), la composition morphologique reste hors de leur portée. Des auteurs ont noté l'importance de la tokénisation, c'est-à-dire la délimitation des termes d'indexation, dans le domaine biomédical (Jiang et Zhai, 2007 ; Trieschnigg *et al.*, 2007), faisant ressortir l'intérêt d'une prise en compte d'une tokénisation fondée sur les morphèmes, mais sans proposer de solutions effectives. Récemment, des outils d'analyse morphologique plus élaborés, développés dans le cadre de MorphoChallenge, ont été appliqués à des tâches de RI (Kurimo *et al.*, 2009). Ici aussi, les auteurs ont observé une amélioration sur certaines langues, comme le finnois, langue agglutinante. Pour l'anglais, l'emploi de tels outils a produit des résultats significativement plus faibles qu'avec un simple raciniseur. À ce titre, les bons résultats que nous présentons en section 6 confirme l'intérêt de notre approche.

### 2.2. *Motivations*

Comme cela a été souligné dans la section 2.1, une grande part des travaux en traitement automatique de la morphologie se fait désormais en recherchant des régularités de structure des mots-formes, en contexte ou non. Lorsque des approches automatiques, c'est-à-dire non fondées sur l'expertise humaine, sont utilisées, ces mots doivent alors être en quantité suffisante pour que la détection des régularités soit statistiquement possible. Ces approches pragmatiques peuvent sembler loin des théories de description morphologique, mais ont pour avantage de reposer résolument sur les données. Elles assurent notamment un compromis entre les critères de minimalité des morphes (la décomposition est la plus fine possible) et de minimalité de la description (la collection de morphes utilisée est la plus petite possible). Ce compromis étant une instanciation du rasoir d'Occam ou du *Minimum Description Length* (MDL), l'utilisation des techniques d'apprentissage (Mitchell, 1997) ou de compression est évidente.

C'est bien sûr cette même approche fondée sur les données qui guide notre travail, mais appliquée conjointement à deux langues. La décomposition de termes étant possible dans chacune des langues considérées, il est naturel de chercher à mettre en correspondance ces décompositions. Plus encore, chacune des langues offre un contexte à l'autre pour guider sa décomposition. L'alignement apparaît comme une solution simple pour mener cette décomposition conjointe en exploitant ce dernier point. Finalement, notre approche s'inscrit donc dans la même logique que les travaux de décomposition automatique à partir de liste de mots, mais en portant le raisonnement à des paires de mots, avec, en l'occurrence, l'un des mots servant de pivot. L'alignement décrit en section suivante cherche à produire des décompositions en morphes et en kanjis avec les mêmes visées de minimalité des décompositions et des descriptions.

Il est également important de noter que le choix des kanjis comme pivots n'est pas fortuit. Les kanjis sont l'un des trois jeux de caractères utilisés en japonais. Ils sont empruntés au chinois – on les appelle aussi sinogrammes – et sont particulièrement utilisés dans les domaines scientifiques. Contrairement à l'hiragana et au katakana, les deux autres jeux de caractères du japonais, les kanjis ne forment pas un syllabaire mais un système d'écriture dit sémantique, d'où leur nombre très important : environ 2 000 kanjis, dits officiels, sont appris dans le cursus scolaire japonais, mais beaucoup d'autres parmi les 50 000 à 60 000 sinogrammes existants sont utilisés, notamment dans les domaines spécialisés. Les kanjis peuvent être des pictogrammes, associant un concept à une image (schématique et parfois difficile à interpréter) par exemple 𥫱 (sang), représentant un calice sacrificiel recevant le sang[5]. Ils peuvent être aussi des idéogrammes, caractères plus complexes composés de plusieurs éléments, par exemple 炎 (flamme, inflammation), composé de deux fois 火 (feu). Ils peuvent aussi être des composés phono-sémantiques, avec un composant reprenant un kanji d'un mot phonologiquement proche (même de sens différent) et l'autre composant précisant son sens ici. Cependant dans notre travail, les kanjis, même étymologiquement composés, sont considérés comme une unité.

5. en.wiktionary.org/wiki/%E8%A1%80#Etymology

L'emploi des termes japonais comme pivots offre plusieurs avantages pour notre problème. Dans le domaine médical, ces termes sont souvent écrits en kanjis et consistent en la juxtaposition de plusieurs kanjis. Leur forme est invariable quelle que soit leur position dans le terme et quels que soient leurs kanjis voisins. Il faut noter cependant que l'unité minimale d'interprétation n'est pas forcément composée d'un seul kanji, certains mots sont composés de plusieurs kanjis qui ne peuvent pas s'interpréter indépendamment, comme dans 茉莉 (jasmin). Les termes en kanjis sont indépendants des racines grecques et latines utilisées dans la plupart des langues européennes. Dans notre processus, d'alignement présenté ci-après, cela empêche de détecter des régularités fortuites simplement fondées sur une étymologie commune. De plus, un segment d'un terme en kanjis est, la plupart du temps, un terme (ou un mot de la langue générale) en lui-même, à l'inverse des morphes (dont le caractère non autonome est justement un critère important pour définir le type de composition). Il est donc possible en théorie d'utiliser un dictionnaire pour accéder à leur sens. Enfin, les termes japonais comportent peu de kanjis (un terme est typiquement composé de un à six kanjis) par rapport aux lettres des termes en alphabet latin (beaucoup de termes dépassent les vingt caractères). Il est donc nécessaire de ne tester que quelques combinaisons lors de leur segmentation, ce qui réduit le coût calculatoire de l'alignement (cf. section suivante). Ces différentes raisons font des termes japonais écrits en kanjis de très bons pivots comparés aux autres langues possibles et disponibles dans les terminologies multilingues que nous utilisons.

D'autre part, la mise en œuvre de notre approche fait une hypothèse forte de parallélisme : le terme en kanjis doit être construit dans un ordre identique à celui des constituants des termes dans la langue étudiée. Cette hypothèse peut paraître peu réaliste, mais les résultats présentés dans les sections suivantes montrent qu'elle est raisonnable dans nos données. Quelques pistes d'explications sont fournies en discussion des résultats (section 4.2). Cette hypothèse posée, l'emploi d'une technique d'alignement sans distorsion (qui ne gère pas de changement d'ordre des éléments à aligner) comme celle présentée en section suivante est suffisante. Il faut noter que d'autres techniques d'alignement gérant la distorsion, par exemple issues des travaux en traduction statistique, permettraient tout de même d'employer la même approche d'analyse morphologique si cette hypothèse n'était pas validée, mais au prix d'un coût calculatoire plus important, et d'une qualité certainement moindre.

## 3. Alignement et analogie pour l'alignement

Comme nous l'avons expliqué précédemment, notre technique de décomposition morphologique repose sur l'alignement des termes avec leur traduction dans la langue pivot (japonais en kanjis). Tout ce qui est fourni est donc une liste de termes avec leur traduction en kanjis, sans autre prétraitement, et c'est le résultat de l'algorithme d'alignement qui va fournir à la fois le découpage en morphes et l'assignation des kanjis aux morphes correspondants.

Cet alignement repose sur un algorithme *Expectation-Maximization* (EM) que nous présentons brièvement dans la sous-section suivante (Jiampojamarn *et al.*, 2007, pour plus de détails et des exemples d'utilisation). La section 3.2 détaille la modification opérée sur cet algorithme standard pour qu'il manipule naturellement et automatiquement la variation morphologique, problème inhérent à la décomposition en morphes.

### 3.1. *Alignement EM*

L'algorithme d'alignement au cœur de notre approche est standard : il s'agit d'un algorithme de type *Baum-Welch* (Rabiner, 1989, pour une présentation didactique), étendu pour faire correspondre des séquences de symboles et non plus seulement produire des alignements 1-1. Dans notre cas, il prend en entrée des termes en français (ou en anglais ou autre) et leur traduction en kanjis, extraits, par exemple, d'une terminologie multilingue.

Le principe de cet algorithme itératif est d'alterner deux opérations : la première calcule une table de compte indiquant avec quel poids chaque alignement possible est rencontré, en s'appuyant sur la probabilité de cet alignement dans chaque paire considérée (cf. exemple ci-dessous), la seconde estime les probabilités d'alignement en s'appuyant à son tour sur la table de compte. Plus formellement, ces deux étapes sont détaillées dans l'algorithme 1 : pour chaque paire de termes $(x^T,y^V)$ à aligner ($T$ et $V$ sont les longueurs en kanjis et en lettres des termes), l'algorithme EM alterne les deux étapes précédentes. Dans l'étape d'*Expectation*, il compte tout d'abord les nombres d'occurrences de tous les alignements possibles entre sous-séquences de kanjis et de lettres (les longueurs maximales des sous-séquences de lettres et kanjis considérés pour l'alignement sont respectivement paramétrées par $maxX$ et $maxY$). Ces comptes sont collectés dans la table $\gamma$, et sont ensuite utilisés dans l'étape de *Maximization* pour estimer les probabilités d'alignement (table $\delta$, voir l'extrait en figure 3).

L'étape d'*Expectation* repose sur une approche *forward-backward* (algorithme 2) : elle calcule les probabilités *forward* $\alpha$ et *backward* $\beta$. Son principe est illustré en figure 1 : il consiste à estimer le poids de chaque alignement possible (ici 学癆/himiot) en fonction de sa probabilité ($\delta$) mais aussi de la probabilité dans la paire d'arriver à cet alignement (par tous les découpages possibles) – c'est $\alpha$ –, et de la probabilité de partir de cet alignement – c'est $\beta$. Ce poids est alors utilisé pour mettre à jour les comptes $\gamma$ pour cet alignement particulier et sera utilisé à la prochaine itération. Plus précisément, à chaque position $t$,$v$ des termes, $\alpha_{t,v}$ est la somme des probabilités de tous les alignements possibles de $(x_1^t,y_1^v)$, c'est-à-dire du début des termes jusqu'à la position courante, selon les probabilités d'alignement de $\delta$ (cf. algorithme 4). De manière similaire, $\beta_{t,v}$ est calculé en considérant $(x_t^T,y_v^V)$. Ces probabilités sont alors utilisées pour réestimer les comptes dans $\gamma$. Dans cette version de l'algorithme EM, l'étape de *Maximization* (algorithme 3) consiste simplement à estimer les probabilités d'alignement $\delta$ en normalisant les comptes de $\gamma$.

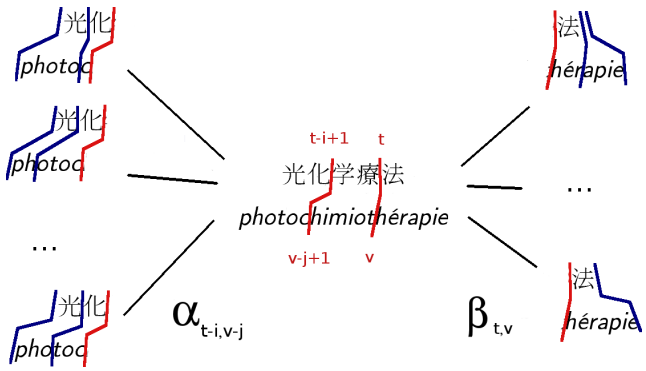


**Figure 1.** *Illustration de la phase* forward-backward *pour l'alignement* 学療/himiot

**Algorithme 1** *Algorithme EM*

Entrée : liste des paires $(x^T, y^V)$, $maxX$, $maxY$
**while** changements dans $\delta$ **do**
  initialisation de $\gamma$ à 0
  **for all** paire $(x^T, y^V)$ **do**
    $\gamma$ = Expectation($x^T$, $y^V$, $maxX$, $maxY$, $\gamma$, $\delta$)
  $\delta$ = Maximization($\gamma$)
**return** $\delta$

**Algorithme 2** *Expectation*

Entrée : $(x^T, y^V)$, $maxX$, $maxY$, $\gamma$
$\alpha$ := Forward-many2many( $x^T$, $y^V$, $maxX$, $maxY$, $\delta$ )
$\beta$ := Backward-many2many( $x^T$, $y^V$, $maxX$, $maxY$, $\delta$ )
**if** $\alpha_{T,V} > 0$ **then**
  **for** $t = 1...T$ **do**
    **for** $v = 1...V$ **do**
      **for** $i = 1...maxX$ t.q. $t - i \geq 0$ **do**
        **for** $j = 1...maxY$ t.q. $v - j \geq 0$ **do**
          $\gamma(x^t_{t-i+1}, y^v_{v-j+1})$ += $\frac{\alpha_{t-i,v-j}\delta(x^t_{t-i+1}, y^v_{v-j+1})\beta_{t,v}}{\alpha_{T,V}}$
**return** $\gamma$

**Algorithme 3** *Maximization*

Entrée : $\gamma$
**for all** sous-séquence $a$ t.q. $\gamma(a,\cdot) > 0$ **do**
  **for all** sous-séquence $b$ t.q. $\gamma(a,b) > 0$ **do**
    $\delta(a,b) = \frac{\gamma(a,b)}{\Sigma_x \gamma(a,x)}$
**return** $\delta$

**Algorithme 4** *Forward-many2many*

Entrée : $(x^T, y^V)$, $maxX$, $maxY$, $\delta$
$\alpha_{0,0} := 1$
**for** $t = 0...T$ **do**
  **for** $v = 0...V$ **do**
    **if** $(t > 0 \vee v > 0)$ **then**
      $\alpha_{t,v} = 0$
    **if** $(v > 0 \wedge t > 0)$ **then**
      **for** $i = 1...maxX$ s.t. $t - i \geq 0$ **do**
        **for** $j = 1...maxY$ s.t. $v - j \geq 0$ **do**
          $\alpha_{t,v} \mathrel{+}= \delta(x^t_{t-i+1}, y^v_{v-j+1})\alpha_{t-i,v-j}$
**return** $\alpha$

Le processus EM est répété jusqu'à ce que les probabilités $\delta$ soient stables. Quand la convergence est atteinte (Gupta et Chen, 2010, pour une discussion sur les propriétés de convergence des algorithmes de ce type), l'alignement consiste simplement à trouver, pour chaque paire, les correspondances maximisant $\alpha_{T,V}$ (cf. figure 2). En plus de ce résultat, nous conservons également les probabilités d'alignement contenues dans $\delta$, que nous utilisons en section 4.5 pour décomposer les termes inconnus (absents des données d'alignement).

Cette technique n'est pas fondamentalement différente de celles largement utilisées en traduction statistique. Mais certaines particularités sont importantes à noter. D'une part, cette approche permet de gérer le phénomène de *fertilité*, c'est-à-dire la capacité d'aligner depuis ou vers une chaîne vide (par souci de place, ces cas ne sont pas présentés dans les algorithmes simplifiés présentés ci-dessus). En revanche, comme nous l'avons expliqué, la *distortion*, c'est-à-dire le réordonnancement de morphes, ne peut pas être facilement géré par ce type d'algorithme.

### 3.2. *Normalisation morphémique automatique*

Comme nous l'avons vu, l'étape de *Maximization* sert à calculer la probabilité de traduction d'une séquence de kanjis par une séquence de lettres. Par exemple, pour la séquence de kanjis 菌 ('*bactérie*'), il peut y avoir plus d'une entrée dans la table $\delta$ l'associant par exemple à bactérie, à bactério (comme dans bactério/lyse) et à bactéri (dans myco/bactéri/ose), chacune avec une certaine probabilité. Cette dispersion de probabilité, qui est bien sûr néfaste pour l'algorithme, est causée par la variation morphémique : bactério, bactérie, et bactéri sont trois morphes du même morphème, et nous souhaiterions que leurs probabilités se renforcent les unes les autres. L'adaptation que nous proposons tente de répondre à ce problème en permettant à l'étape de *Maximization* de grouper automatiquement les différents morphes appartenant à un même morphème. Pour ce faire, nous utilisons une technique simple mais adaptée s'appuyant sur le calcul d'analogies formelles.

3.2.1. *Analogie entre morphes*

Une analogie est une relation entre quatre éléments que nous notons $a : b :: c : d$, et qui peut se lire $a$ *est à* $b$ *ce que* $c$ *est à* $d$ (Lepage, 2000, pour une présentation plus détaillée). Les analogies ont été utilisées dans de nombreux travaux de TAL, notamment pour la traduction de phrases (Lepage, 2000) ou de termes (Langlais et Patry, 2007 ; Langlais *et al.*, 2008). Elles sont aussi au cœur de nos travaux déjà mentionnés sur la structuration terminologique (Claveau et L'Homme, 2005) ou de (Hathout, 2009), desquels nous nous inspirons pour formaliser notre problème.

L'idée originale que nous souhaitons mettre en œuvre est d'appliquer cet outil standard de la morphologie non pas à l'échelle du mot-forme, comme en morphologie lexématique, mais à l'échelle du morphe. Ainsi, dans notre cadre, une analogie possible entre quatre morphes est : dermato : dermo :: hémato : hémo. Sachant que dermato et dermo appartiennent au même morphème, on peut inférer que c'est aussi le cas pour hémato et hémo. Une telle analogie, construite sur la représentation graphémique des mots, est dite formelle.

Selon Stroppa et Yvon (2005), les analogies formelles peuvent être définies en termes de *factorisations*. Notons $\overleftarrow{\oplus}$ l'opérateur (non commutatif) de concaténation à droite ($abc\overleftarrow{\oplus}d = abcd$), et $\overleftarrow{\ominus}$ l'opérateur associé de soustraction ($abc\overleftarrow{\oplus}d\overleftarrow{\ominus}d = abc\overleftarrow{\ominus}c\overleftarrow{\oplus}c = abc$) ; similairement, nous notons $\overrightarrow{\oplus}$ et $\overrightarrow{\ominus}$ les opérateurs concaténant ou soustrayant à la gauche du premier argument. Soit $a$ une séquence de caractères (un terme dans notre cas) sur un alphabet $\Sigma$, une factorisation de $a$, notée $f_a$ est une séquence de $n$ facteurs $f_a = (f_a^1,...,f_a^n)$, telle que $a = f_a^1\overleftarrow{\oplus}f_a^2\overleftarrow{\oplus}...\overleftarrow{\oplus}f_a^n$. Une analogie formelle est définie comme suit.

**Définition 1** $\forall(a,b,c,d) \in \Sigma,[a : b :: c : d]$ *ssi il existe des factorisations* $(f_a,f_b,f_c,f_d) \in (\Sigma^{*n})^4$ *de* $(a,b,c,d)$ *telles que,* $\forall i \in [1,n],(f_b^i,f_c^i) \in \{(f_a^i,f_d^i),(f_d^i,f_a^i)\}$. *Le plus petit* $n$ *pour lequel cette définition est vérifiée est le degré de l'analogie.*

Pour la plupart des langues indo-européennes, comme le français et l'anglais, les interactions entre les morphes composant un terme sont principalement susceptibles d'en modifier les préfixes et suffixes[6]. Nous nous intéressons donc aux analogies de degré au plus trois, c'est-à-dire avec trois facteurs : préfixe $\overleftarrow{\oplus}$ base $\overleftarrow{\oplus}$ suffixe.

En pratique, pour vérifier si quatre morphes sont en analogie, nous construisons une règle de réécriture entre une des deux paires de morphes (par exemple dermato-dermo) et vérifions si elle s'applique à l'autre paire (hémato-hémo). Cette règle fait intervenir les trois facteurs ; la base est définie comme étant la plus longue sous-chaîne commune (lcss) entre les deux mots (par exemple, lcss(dermato-dermo) = derm). Ainsi, pour passer de dermato à dermo, il faut ôter à cette base ato et ajouter o. La règle est donc : $r = \text{lcss}(\text{morphe}_1,\text{morphe}_2) \overleftarrow{\ominus} \text{ato} \overleftarrow{\oplus} \text{o}$.

6. Préfixe et suffixe sont ici utilisés au sens informatique du terme : des chaînes de caractères apparaissant en début ou en fin de séquences, indépendamment de leur pertinence linguistique.

Cette règle permet bien de réécrire hémato en hémo, donc, hémato,hémo est en analogie avec dermato,dermo.

3.2.2. *Normalisation par analogie*

En pratique, pour notre problème, il est nécessaire d'avoir des exemples (comme dermato et dermo précédemment) pour utiliser les analogies pour regrouper les morphes en morphèmes. Pour contourner ce problème, nous utilisons une technique simple d'amorçage : si deux morphes sont recensés dans $\gamma$ comme des traductions d'une même séquence de kanjis, et si ces deux morphes partagent une sous-chaîne commune de longueur supérieure à un certain seuil, alors nous supposons qu'ils appartiennent à un même morphème. Par exemple, on observe que parmi les morphes alignés à 咽頭, pharyngo et pharynx partagent une longue sous-chaîne commune ; on suppose que ce sont deux morphes d'un même morphème. À partir de ces paires d'amorçage, nous construisons les règles $r$ de préfixation et suffixation nous permettant de grouper d'autres paires de morphes. Plus une règle est trouvée fréquemment, plus elle est considérée comme sûre. Nous collectons donc toutes les règles générées à chaque itération avec leur nombre d'occurrences, et nous n'appliquons que les plus fréquentes à l'itération suivante (contrairement aux amorces, ces règles peuvent alors être appliquées à des morphes très courts). L'ensemble du processus est donc entièrement automatique.

La nouvelle étape de *Maximization*, incluant cette normalisation morphémique, est résumée dans l'algorithme 5. Elle permet que tous les morphes considérés comme appartenant à un même morphème aient une probabilité égale et renforcée.

**Algorithme 5** *Maximization* avec normalisation morphémique

Entrée : $\gamma$
**for all** séquence de kanjis $k$ t.q. $\gamma(k,\cdot) > 0$ **do**
  **for all** morphes $m_1$,$m_2$ t.q. $\gamma(k,m_1) > 0 \wedge \gamma(k,m_2) > 0 \wedge$ lcss($m_1$,$m_2$) > seuil **do**
    construire la règle de préfixation et suffixation $r$ pour $m_1$,$m_2$
    incrémenter le score de $r$
  **for all** sous-séquence $b$ t.q. $\gamma(a,b) > 0$ **do**
    construire l'ensemble $\mathcal{M}$ de tous les morphes associés à $b$ à l'aide des $n$ règles de réécriture les plus fréquentes de l'itération précédente
    $\delta(a,b) = \frac{\sum_{c \in \mathcal{M}} \gamma(a,c)}{\sum_{x} \gamma(a,x)}$
**return** $\delta$

## 4. Validation expérimentale

### 4.1. *Données d'évaluation*

Les données utilisées dans les expériences rapportées ci-après sont issues du MetaThesaurus de l'UMLS (Tuttle *et al.*, 1990). Le MetaThesaurus regroupe plusieurs ter-

minologies dans plusieurs langues et associe à chaque terme un identifiant conceptuel (CUI, *Concept Unique Identifier*). Les CUI sont indépendants de la langue ; ils permettent donc de construire très facilement des listes de termes dans la langue étudiée associés à leur équivalent en japonais. Dans cet article, nous présentons les résultats pour deux langues, le français et l'anglais. Dans les deux cas, nous ne considérons que les termes japonais entièrement écrits en kanjis, et uniquement les termes simples (composés d'un seul mot-forme) français ou anglais. Environ 14 000 paires anglais-kanjis et 8 000 français-kanjis sont collectées de cette façon. Un symbole de fin de chaîne (' ;') est ajouté aux termes français et anglais. Nous avons sélectionné aléatoirement environ 1 000 paires pour le français et 500 pour l'anglais pour évaluer les performances de notre technique d'alignement. Ces paires ont été alignées manuellement (par l'auteur avec l'aide de locuteurs japonais) et servent de vérité terrain.

Les différentes versions de l'algorithme d'alignement ont donc été lancées sur les paires d'entraînement. À convergence, l'algorithme produit l'alignement le plus probable entre chaque terme et sa traduction en kanjis, accompagnés de la probabilité (extrait sur l'anglais en figure 2). Nous récupérons également la table de probabilités $\delta$ (extrait en figure 3) et les règles de réécriture analogiques de la dernière itération.

```
卵子形成/ovogenesis          卵子:ovo 形成:genesis;          P=0.000761417
高炭酸症/hypercapnia         高:hyper 炭酸症:capnia;         P=2.84108e-06
状胞性血/drepanocytosis      状:drepano 胞:cyt 性血:osis;    P=7.89005e-09
角膜全移植/keratoplasty      角膜全:kerato 移植:plasty;      P=0.000652548
下垂体切除/hypophysectomy    下垂体:hypophys 切除:ectomy;    P=1.88705e-06
灼痛/causalgia               灼:caus 痛:algia;               P=8.7077e-05
                             ...
```

**Figure 2.** *Extrait des alignements produits pour l'anglais par l'algorithme EM avec normalisation morphémique*

```
上/ia;              0.00099
上嫌/euphor         4.950495e-05
上嫌/euphoria;      4.950495e-05
上炎/itis;          4.470132e-52
上狭窄/ostenosis;   5.59957e-23
上狭窄/stenosis;    7.716783e-17
上皮/carcino        2.568568e-311
...
```

**Figure 3.** *Extrait de la table de probabilités $\delta$ produite pour l'anglais à la dernière itération de l'algorithme EM avec normalisation morphémique*

### 4.2. *Résultats d'alignement*

Dans les différentes expériences rapportées ci-dessous, les performances sont évaluées en termes de précision : un alignement est considéré comme correct si tous les composants de la paire sont correctement alignés (c'est l'équivalent du *sentence error rate* en traduction artificielle).

Pour chaque paire, l'algorithme EM indique la probabilité de l'alignement proposé. Il est donc possible de ne considérer que les alignements de probabilités supérieures à un certain seuil. En faisant varier ce seuil, nous pouvons calculer une précision selon le nombre de termes alignés. Les figures 4 et 5 présentent respectivement les résultats obtenus pour le français et l'anglais. Nous y indiquons les performances de l'algorithme EM avec et sans normalisation morphémique. À des fins de comparaison, nous indiquons aussi les résultats de GIZA++ (Och et Ney, 2003), un outil d'alignement de référence dans le domaine de la traduction artificielle. Les différents modèles IBM et jeux de paramètres de GIZA++ ont été testés ; les résultats rapportés sont les meilleurs obtenus (avec un modèle IBM 4 sans distorsion).

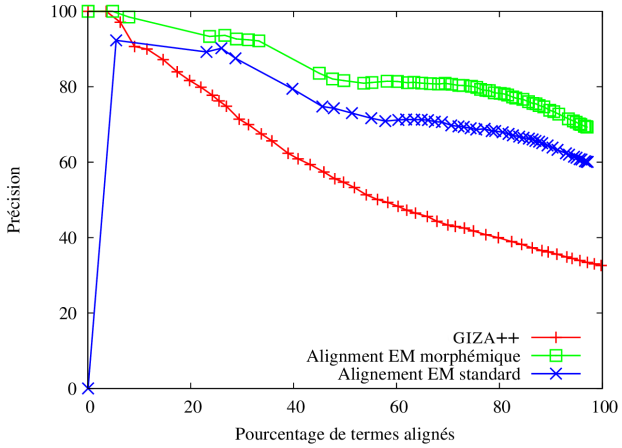


**Figure 4.** *Précision de l'alignement français-kanjis selon le nombre de paires de tests alignées*

Comme espéré, l'intérêt de la normalisation morphémique apparaît clairement dans ces deux expériences. Dans le pire cas (c'est-à-dire quand on tente d'aligner toutes les paires), on obtient 70 % de précision pour le français et 80 % pour l'anglais. Cela correspond à une amélioration de 10 % quel que soit le nombre de paires considérées. Il faut noter que la normalisation a aussi un autre intérêt puisqu'elle permet une réduction de complexité : moins d'itérations sont nécessaires pour atteindre la convergence de l'algorithme EM.

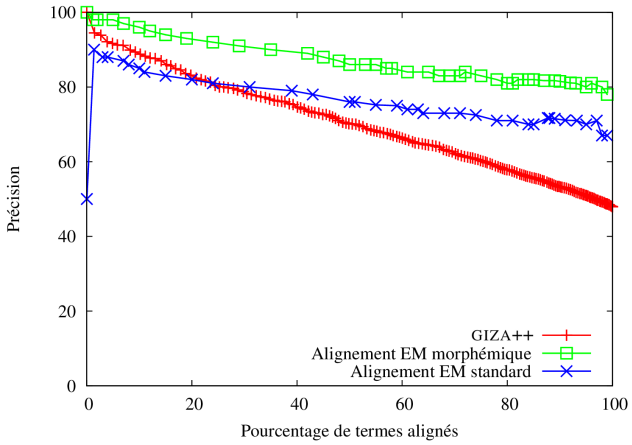


**Figure 5.** *Précision de l'alignement anglais-kanjis selon le nombre de paires de tests alignées*

Pour étudier l'impact de la taille de l'ensemble d'apprentissage de notre approche, nous présentons en figure 6 la précision en fonction du nombre de paires utilisées pour le français. L'augmentation rapide des performances illustre l'importance des

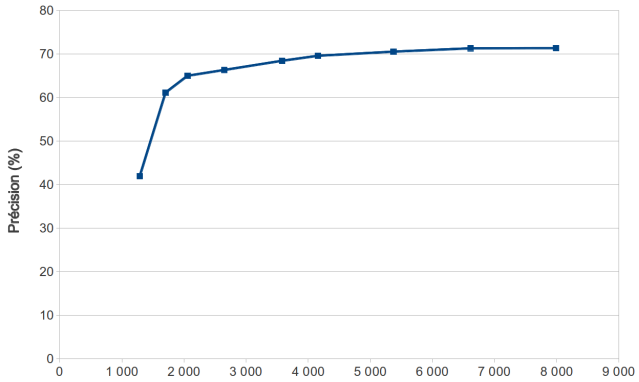


**Figure 6.** *Précision de l'alignement sur le français selon le nombre d'exemples*

exemples pour couvrir la plupart des morphes utilisés. L'augmentation ensuite plus lente mais continue s'explique également par l'apport de nouveaux constituants mais aussi par l'importance des redondances des morphes et des kanjis, même déjà rencontrés, pour que l'algorithme estime au mieux leur probabilité d'alignement.

### 4.3. *Mise en regard d'autres approches*

En complément des expériences précédentes, nous nous intéressons dans cette sous-section à la comparaison de notre approche avec deux outils emblématiques du domaine : Morfessor[7] (Creutz et Lagus, 2005) et DeriF[8] (Namer, 2007). Cette comparaison porte uniquement sur les capacités de segmentation de ces outils, c'est-à-dire le découpage en morphes. En effet, cette tâche est la seule opérée par Morfessor, et la comparaison des informations sémantiques de DeriF avec celles apportées par nos kanjis est difficile à évaluer. Nous mesurons donc la précision de ces outils sur cette tâche de segmentation en utilisant la vérité terrain de l'expérience précédente.

Morfessor est un représentant de la classe des outils fondé sur le principe MDL pour détecter au sein d'une liste de mots les séquences de lettres pouvant former des composants morphologiques. La taille de la liste fournie peut être un paramètre important pour la qualité des résultats ; nous la faisons donc varier pour observer l'évolution des performances de l'outil (à partir d'une liste de 13 000 termes simples du français collectés dans l'UMLS). DeriF reposant, quant à lui, sur une expertise humaine, les segmentations évaluées sont celles directement produites par l'outil dans sa version actuelle. Ces segmentations, qui sont hiérarchques, sont ici aplanies pour la comparaison. Nous opérons également quelques opérations de normalisation visant à ne pas le pénaliser du fait de certains choix concernant les suffixes (DeriF propose un suffixe ie dans acr/odyn/ie) différents de ceux induits par notre approche (odynie est considéré comme une réalisation du morphème odyn- s'il n'induit pas de sens différent attesté par une différence de kanjis).

La figure 7 représente la précision des trois systèmes, selon le nombre de mots (pour Morfessor) et le nombre de paires de mots (pour notre système). On y retrouve sans surprise la même forme de courbe d'apprentissage pour notre approche que précédemment, avec un précision assez élevée dès le départ, et une précision maximale de 80,5 %. Bien entendu, ces résultats sont directement corrélés à ceux de l'alignement. Les résultats de Morfessor sont, quant à eux, relativement constants selon le nombre d'exemples. Un examen des erreurs explique ce résultat *a priori* surprenant et fait apparaître une propension à sursegmenter sur la base de quelques ressemblances fortuites entre mots. À nombre d'exemples égal, notre approche offre de bien meilleurs résultats, mais elle dispose de plus d'informations à travers les traductions en kanjis. Le système DeriF donne des résultats proches des nôtres sur cette tâche de segmentation. Ses erreurs relèvent principalement d'une absence de segmentation s'expliquant peut-être par la présence de morphes inconnus (*e.g.* améloblastome, athétose, micrognathisme, sparganose, sérosite ne sont pas segmentés), ou des configurations de morphes particulières (argyrose, cholangite, angiocholite). Dans de plus rares cas, il s'agit d'erreurs de segmentation (morphe incorrectement reconnu ; *e.g.* ré/tinoblastome). L'absence de segmentation ou la segmentation partielle sont aussi les erreurs principales de notre approche, causées par des morphes et des kanjis peu rencontrés. L'importance de

7. www.cis.hut.fi/projects/morpho
8. www.cnrtl.fr/outils/DeriF/requete.php

ces erreurs diminue cependant à mesure que les données d'entraînement augmentent. Dans l'ensemble, DeriF et notre approche s'accordent sur 70,5 % des segmentations, et plus de 80 % des segmentations correctes de l'un sont trouvées par l'autre.

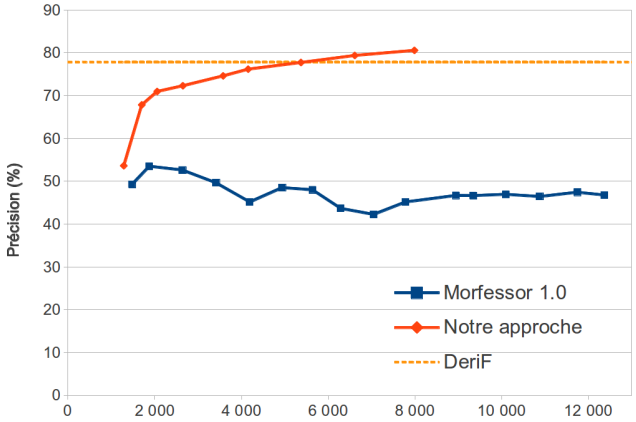


**Figure 7.** *Précision de la segmentation sur le français selon le nombre d'exemples*

La comparaison des représentations sémantiques entre DeriF et notre approche est difficile à mesurer quantitativement. D'une part DeriF offre une segmentation hiérarchique que notre système n'est pas capable de produire. D'autre part, les informations sémantiques de DeriF sont assez générales et parfois manquantes, alors que celles apportées par les kanjis sont propres au terme. Cela montre la potentielle complémentarité des systèmes : notre approche pourrait fournir à DeriF des informations sémantiques dans certains cas plus adaptées et renforcer sa couverture, tout en bénéficiant de l'avantage du découpage hiérarchique de DeriF, comme l'illustre le tableau 1.

### 4.4. *Discussion des résultats et retour sur les hypothèses*

Un examen manuel des résultats d'alignement montre que la plupart des erreurs d'alignement sont des mises en défaut de notre hypothèse de parallélisme : certaines paires français-kanjis ou anglais-kanjis ne peuvent pas être décomposées de manière similaire. Par exemple, le terme anxiolytiques est traduit en kanjis par un terme se lisant littéralement 'médicament pour la dépression'. Parmi ces erreurs, certaines impliquent un terme qui n'est pas un composé soit du côté français, soit du côté japonais, soit des deux ; par exemple, méninges est traduit par le terme 脳膜 qui se lit littéralement 'cerveau membrane'. Ces problèmes mènent alors à des décompositions qui ne sont pas minimales (deux morphes ne sont pas séparés) voire complètement erronées. Les autres erreurs sont causées par un manque de données d'entraînement : quelques morphes ou séquences apparaissent seulement une fois, ou toujours combinés avec un autre morphe, ce qui trompe le processus de segmentation.

| Terme | DeriF | Décomposition en kanjis |
|---|---|---|
| asbestose | (Partie de – Type particulier de) affection chronique en rapport avec le(s) asbeste | 石綿 'amiante' 症 'maladie' |
| anurie | Absence de appareil urinaire | 無 'manque' 尿 'urine' 症 'maladie' |
| stéréopsie | Affection du/des vue liée au(x) stère | 立体 'tridimensionnel' 視 'regard, vision' |
| calcinose | (Partie de – Type particulier de) affection chronique en rapport avec le(s) calcin | 石灰 'calcaire' 症 'maladie' |
| pneumopéricarde | (Partie de – Type particulier de) péricarde en rapport avec le(s) poumon | 気 'air' 心膜 'péricarde' 症 'maladie' |
| sarcolemme | (Partie de – Type particulier de) lemme en rapport avec le(s) chair | 筋 'muscle' 鞘 'enveloppe' |

**Tableau 1.** *Exemples d'analyses produites par DeriF pouvant bénéficier de notre approche*

Mais au-delà de ces quelques erreurs, il est intéressant de constater que l'hypothèse de parallélisme des constructions des termes est majoritairement vérifiée. Si le principe même de détecter des régularités se correspondant dans les deux langues ne faisait pas de doute (cf. section 2.2), le fait que l'ordre des morphes et des kanjis soit identique peut paraître plus surprenant. Cela peut néanmoins s'expliquer par deux éléments. D'une part, bien que l'ordre des mots en grammaire japonaise soit souvent qualifié comme en partie libre (les rôles des mots peuvent être marqués au moyen de particules qui leur sont accolées), un principe de base est d'indiquer les attributs d'un objet avant l'objet, les actants avant le prédicat marquant l'action, et, d'une manière générale, l'élément régi avant son recteur (Nakamura-Delloye, 2007, pour une présentation didactique et étayée). Un tel ordre se reflétant dans la structure des termes biomédicaux japonais correspond effectivement à la structure de beaucoup des termes anglais ou français manipulés (Dal et Amiot, 2008 ; Namer, 2007) pour lesquels cet ordre est justement une marque de la composition néoclassique[9]. D'autre part, une explication plus pragmatique tient peut-être à la façon dont les terminologies japonaises utilisées dans l'UMLS ont été construites. Par exemple, le JAMAS (Japanese Medical Abstracts Society , `www.jamas.gr.jp`) précise avoir construit le *Thesaurus for Medical and Health Related Terms*, utilisé dans l'UMLS, en se basant sur le MeSH anglais. Ses développeurs indiquent n'avoir utilisé que des termes apparaissant dans des articles publiés en japonais (Onogi *et al.*, 2004), mais il n'est pas à exclure que la

9. Dal et Amiot (2008) notent ainsi : « *Dans les langues romanes, l'ordre des constituants est [...] différent, selon qu'il s'agit de composition néoclassique ou de composition ordinaire : dans le premier cas, le constituant sémantiquement recteur se trouve à droite, dans le second, il est à gauche.* »

recherche d'équivalent ait induit un biais dans le choix des termes japonais, favorisant ceux aux structures proches des termes anglais.

Il est également intéressant d'observer le comportement de la méthode sur des termes dont la composition ne fait pas intervenir que des constituants non autonomes (au sens syntaxique) munis d'un sens référentiel d'origine gréco-latine. Cette origine gréco-latine des composants, associée au fait qu'ils existent de manière autonome ou non, est une condition controversée pour définir la composition néoclassique (Dal et Amiot, 2008, pour une discussion complète). Bien entendu, ces considérations n'ont pas d'impact sur notre approche, l'origine et le statut d'autonomie syntaxique n'étant pas des critères intervenant dans la décomposition. Ainsi, diétothérapie, exemple de composés dont le statut est discuté, est régulièrement aligné avec 食事 'repas, diète' et 療法 'thérapie'.

Par ailleurs, certains termes reposent aussi sur d'autres opérations morphologiques comme la dérivation. Par exemple, la préfixation par re- dans le terme anglais reintubation n'est pas une composition néoclassique. Là encore, cette distinction n'a pas de conséquence sur l'approche. En effet, ces constituants seront considérés comme les autres et éventuellement alignés aux kanjis correspondants (re- est ainsi aligné le plus souvent avec 再 'encore, répété'). De plus, ces alignements permettent de mettre en avant le sens contextuel de ces constituants (Corbin, 2001). Ainsi, le terme anglais dehydration est traduit par 脱水 'enlever' 'eau', alors que de- dans dehumanisation est traduit par 非 'non' (humanisé), dans deceleration il est traduit par 減 'moins' (de vitesse), dans decerebration il est traduit par 除 'éliminer' (le cerveau)... Même si ces considérations sont hors de portée de cet article, le parallèle dressé avec le japonais pourrait, là encore, guider une analyse linguistique de ces phénomènes en s'appuyant sur des données réelles. Notons également que certains suffixes propres au vocabulaire médical (*e.g.* en français -ite, -ose), sont naturellement considérés de la même manière que les autres constituants et régulièrement alignés avec les kanjis correspondant au sens porté (inflammation, maladie...). Cela rejoint l'analyse de Namer (2007), amenée à les intégrer dans ce qu'elle nomme éléments de formation sous un statut d'exception. En revanche, dans l'UMLS, les variations flexionnelles et certaines variations dérivationnelles sont classées sous un même identifiant de concept. Il n'est donc pas possible de faire apparaître de différences d'analyse sur des cas de conversion, comme par exemple, entre reintubation et reintubate, tous deux traduits par 再挿管. D'autres cas de dérivation, entraînant des différences conceptuelles suffisamment importantes pour être référencés dans l'UMLS selon des CUI différents, sont néanmoins observables. Là encore, pour notre approche, ces cas sont traités comme pour la composition néoclassique et les kanjis alignés apportent les informations nécessaires à la segmentation et à l'interprétation. On a ainsi dent/歯, dentition/歯列 ('dent' 'rangée'), dentiste/歯科医師 ('dent' 'spécialiste'), et aussi odontologie/歯学 ('dent' 'science').

### 4.5. *Traduction et segmentation de termes inconnus*

Notre technique d'alignement peut être utilisée comme première étape pour traduire un terme inconnu (absent des données ayant servi à l'algorithme d'alignement), en s'appuyant sur ses composés morphologiques potentiellement connus. La traduction de termes a déjà fait l'objet de plusieurs travaux, principalement pour pallier les erreurs dites *out-of-vocabulary* lors des tâches de traduction artificielle de textes. Beaucoup de ces travaux cherchent des traductions dans des ressources textuelles : corpus parallèles ou comparables (Chiao et Zweigenbaum, 2002 ; Fung et Yee, 1998), Web (Lu *et al.*, 2005). Quelques auteurs se sont intéressés à ce même problème mais sans autres données que les paires de termes, en s'appuyant sur les similarités (cognats) d'une langue à l'autre (Schulz *et al.*, 2004, par exemple), ou sur les similarités des opérations permettant de passer d'une langue à l'autre (Langlais et Patry, 2008 ; Claveau, 2009 ; Harastani *et al.*, 2012). C'est bien sûr dans le cadre de ces derniers travaux que nous nous situons.

Dans l'expérience reportée ici, nous traduisons des termes français vers le japonais. En pratique, nous utilisons les probabilités $\delta$ pour générer une traduction. La mise en œuvre que nous utilisons ici est très simple : les probabilités de traduction des morphes dans $\delta$ sont exploitées dans un algorithme de type Viterbi ; nous n'utilisons donc pas de modèle de langue. Ce processus de traduction a un autre avantage très intéressant : il produit l'alignement du terme de la langue source avec sa traduction. Le terme est donc segmenté en morphes, et les morphes sont étiquetés par les kanjis correspondants. Cela produit bien une décomposition morphosémantique du terme inconnu.

Pour les besoins de cette expérience, 128 termes et leurs traductions en kanjis ont été sélectionnés aléatoirement pour constituer notre jeu de test (ils ont bien sûr été retirés des données alimentant notre algorithme d'alignement). Ces termes français sont ensuite traduits grâce aux probabilités d'alignement $\delta$ et comparés à la référence.

| Référence | UMLS | Web |
|---|---|---|
| Correctement traduit (et segmenté) | 58 | 82 |
| Incorrectement traduit (ou segmenté) | 34 | 10 |
| Non traduit | 36 | 36 |

**Tableau 2.** *Résultats (précision en pourcentage) de la traduction de termes inconnus selon une référence stricte (*UMLS*) et une référence large (Web)*

Les résultats de cette petite expérience sont présentés dans le tableau 2. Sur 128 termes, 58 ont été correctement traduits (et segmentés), soit une précision de 45 %. Il y a deux types d'erreurs : 34 termes ont reçu une mauvaise traduction, et pour les 36 restants, aucune traduction-décomposition n'a été trouvée. En examinant ces derniers, on trouve sans surprise des termes qui ne sont pas des composés savants, ou des composés dont un ou plusieurs constituants n'apparaissent pas dans les données d'entraînement. La précision mesurée sur les seuls termes traduits est donc de 63 %, ce

qui semble très encourageant étant donnée la simplicité de notre mise en œuvre. Autre point intéressant, un examen manuel des résultats montre que pour certains cas notés comme des erreurs, les premières traductions proposées sont souvent des paraphrases correctes, non attestées dans l'UMLS mais attestées sur le Web dans des sites institutionnels japonais (gouvernement, universités...) ou des documents hébergés (PDF) du domaine biomédical. Avec cette référence plus large, contrôlée par un japanophone, la précision de traduction (et décomposition) atteint 89 %.

## 5. Analyse morphosémantique pour la recherche d'information

Comme cela a été souligné précédemment, la recherche d'information dans le domaine biomédical a quelques spécificités du fait de l'utilisation de termes spécialisés. À ce titre, l'utilité de prendre en compte des informations morphologiques riches a déjà été démontrée, mais seulement avec des ressources développées manuellement (Markó *et al.*, 2005b). En complément des évaluations intrinsèques de notre approche présentées dans la section précédente, nous évaluons ci-après son intérêt dans une tâche de recherche d'information sur une grande collection de documents biomédicaux en anglais. Plus précisément, nous montrons comment tirer parti de toutes les informations produites par l'alignement et sous quelle forme elles peuvent être ensuite exploitées pour indexer des documents.

### 5.1. *Graphes morphosémantiques*

À l'issue de la phase d'alignement, il est possible d'étudier les correspondances récurrentes entre les morphes des termes anglais et kanjis. Plus un morphe est aligné fréquemment avec une même séquence de kanjis, plus on les suppose liés sémantiquement. Tous ces liens forment un graphe : les nœuds sont les kanjis et les morphèmes (morphes groupés par les règles de réécriture collectées par analogie), et les arcs entre kanjis et morphèmes sont pondérés en fonction de la fréquence d'alignement dans les 14 000 paires de l'UMLS. La figure 8 illustre un tel graphe sur un exemple jouet ; l'épaisseur des arcs est proportionnelle aux poids associés.

Cette représentation nous permet de mettre au jour différents types de relations sémantiques entre les morphèmes. Pour ce faire, nous explorons le voisinage de chaque morphème en suivant les arcs et en tenant compte des poids associés à ces liens (de manière similaire à l'algorithme de Folk-Fulkerson (Heineman *et al.*, 2008) ; cela revient à donner à chaque nœud une certaine quantité d'énergie qui la distribue à ses voisins proportionnellement à la force du lien qui les unit, ceux-ci la distribuant à leur tour, etc.).

La figure 9 présente les kanjis (traduits manuellement), sous la forme d'un nuage de mots (la taille et la couleur représentent donc l'énergie atteignant ce nœud) pour le morphème représenté par ome; (suffixe pour 'tumeur').

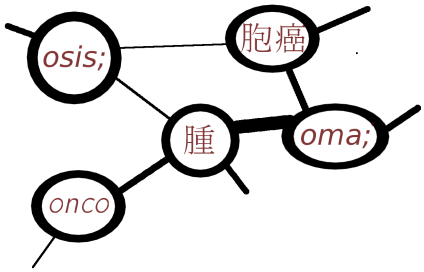


**Figure 8.** *Graphe morphosémantique morphème-kanji*


**Figure 9.** *Nuage de kanjis pour* ome;

Sous le même format, la figure 10 présente les morphèmes trouvés dans le voisinage du même morphème ome;. Les nœuds atteints sont considérés comme conceptuellement proches et sont supposés mettre en lumière des liens de synonymie ou de quasi-synonymie. Il est intéressant de noter que d'autres suffixes proches sont trouvés, mais aussi des morphèmes plus couramment utilisés en préfixe comme onco.

L'alignement et la segmentation produits par notre algorithme permettent aussi d'étudier les cooccurrences des morphèmes. On peut ainsi s'intéresser aux affinités de premier ordre, c'est-à-dire observer quels sont les morphèmes qui cooccurrent fréquemment ensemble. On peut aussi rechercher les affinités de second ordre, c'est-à-dire rechercher les morphèmes partageant les mêmes morphèmes cooccurrants. Ces affinités de second ordre doivent nous permettre de grouper les morphèmes partageant une certaine proximité sémantique. On espère ainsi capturer des relations

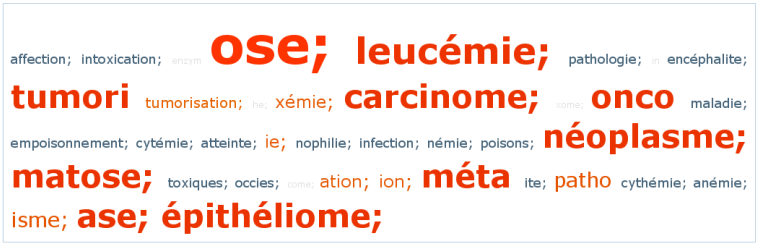


**Figure 10.** *Nuage des affinités de premier ordre du morphème français ome;*

comme celles définies manuellement par Namer et Baud (2007) : synonymie ou quasi-synonymie, cohyponymie, et proximité (physique, fonctionnelle ou procédurale). Par exemple, le nuage de la figure 11 regroupe les morphèmes proches, selon cette affinités de second ordre, du morphème gastro (signifiant estomac). La plupart de ces voisins désignent des organes, et il est même intéressant de noter que ce sont les organes anatomiquement les plus proches qui sont les plus liés.


**Figure 11.** *Nuage des affinités de second ordre du morphème français gastro*

Ces informations de différentes natures nous permettent d'identifier des relations entre termes grâce à leurs composants, à construire des synonymes, ou, plus largement, à explorer la base terminologique grâce à ces éléments morphologiques. Cependant, à notre connaissance, aucune ressource morphosémantique comparable et suffisamment complète n'est disponible pour permettre une évaluation directe de nos résultats. Dans la suite de cet article, nous proposons une évaluation indirecte à travers une tâche de recherche d'information dans laquelle nous tentons d'exploiter au mieux ces informations morphologiques.

### 5.2. *Représentation morphémique pour la recherche d'information*

Pour intégrer les informations morphologiques dans un système de RI, nous adoptons une représentation simple des documents pour l'indexation : ils seront vus comme des sacs de mots et/ou morphèmes, selon les différents systèmes proposés ci-dessous. Les morphèmes sont ceux obtenus en décomposant les termes biomédicaux, et pour certains systèmes ceux qui leur sont liés (au sens des affinités discutées dans la sous-section précédente). Le but de ces expériences est bien sûr d'utiliser les informations morphologiques pour être capable de retourner un document contenant gastrodynia à une requête contenant stomachalgia.

Lors de la phase d'indexation, les termes sont décomposés. Deux cas peuvent se produire : soit le terme est connu car il apparaît dans les paires servant à l'alignement, soit il est inconnu. Dans le premier cas, nous utilisons simplement sa décomposition telle qu'elle est produite par l'algorithme d'alignement. Dans le second cas, nous tentons de le décomposer en suivant le principe expliqué en section 4.5. Dans ces deux cas, un autre produit de l'algorithme d'alignement est utilisé : il s'agit des règles de réécriture produites par les analogies et collectées à la dernière itération EM. Elles nous permettent de détecter les morphes appartenant aux mêmes morphèmes. Ainsi une requête contenant hemo peut être mise en correspondance avec des documents contenant haemo, hemato ou même emia;.

Nous proposons un système de référence et quatre systèmes utilisant les informations morphologiques. Tous reposent sur une approche RI standard connue pour ses bonnes performances, à savoir un modèle vectoriel avec une tokénisation similaire à celle opérée dans l'outil de RI TERRIER (Ounis *et al.*, 2006) et un schéma de pondération Okapi BM-25 qui peut être vu comme un TF-IDF moderne (Robertson *et al.*, 1998, pour plus de détails) ; les valeurs par défaut des paramètres $b$, $k_1$, $k_3$ de cette pondération ont été gardées. Ce système a été notamment utilisé dans le cadre de la tâche de filtrage dont nous utilisons les données (cf. *infra* où il a donné les meilleurs résultats (Robertson et Hull, 2000).

Le système de référence effectue une indexation classique des documents fondée sur les mots avec une racinisation de Porter (Porter, 1980). Les autres systèmes sont les suivants.

1) Le premier système utilisant des informations morphologiques que nous proposons repose sur les morphèmes. Il considère simplement les morphèmes produits par la décomposition des mots des documents (et des requêtes) comme les éléments d'indexation (les termes sont aussi conservés pour l'indexation). Le poids assigné à chaque morphème prend en compte la probabilité de décomposition : c'est le produit de cette probabilité et du poids BM-25.

2) Le deuxième système repose sur les kanjis. Là encore, les termes des documents sont décomposés, et les kanjis les plus proches sont utilisés comme éléments d'indexation. Ces kanjis sont ceux identifiés dans le voisinage des morphes produits par la décomposition des termes (voir la section 5.1).

3) Le troisième système reprend la même représentation que le premier, mais étend les requêtes avec les affinités de premier ordre de ses morphèmes. Les morphèmes utilisés en extension sont pondérés selon leur proximité dans le graphe et le poids des morphèmes qu'ils étendent.

4) Le dernier système est similaire au troisième mais utilise cette fois les affinités de second ordre pour étendre les requêtes.

## 6. Expériences de RI biomédicale

Cette section présente le contexte expérimental et les résultats obtenus par les systèmes décrits dans la section précédente. Le but de ces expériences est de caractériser le gain que l'on peut attendre d'un tel composant morphologique pour un système de RI dédié au domaine biomédical.

### 6.1. *Contexte expérimental*

Dans les expériences rapportées ci-après, nous utilisons le jeu de données issu de la tâche de filtrage (*filtering track*) de la conférence TREC-9. Ce jeu s'appuie lui-même sur la collection OHSUMED, qui est composée de 350 000 résumés de MEDLINE. À ces documents, 4 000 requêtes et leurs jugements de pertinence ont été développés pour TREC-9. Les requêtes comportent plusieurs champs : le sujet, qui est un terme de la terminologie MeSH, et la définition de ce terme. Nous utilisons cette collection comme une collection d'évaluation de RI classique, et ne considérons que le champ sujet pour les requêtes.

### 6.2. *Résultats*

La figure 12 synthétise les résultats de tous les systèmes en termes de précision moyenne (*mean average precision*, MAP). À des fins de comparaison, nous indiquons également les résultats obtenus par notre système de référence (cf. section 5.2) et par trois systèmes de RI biomédicale proposés par Dinh et Tamine (2010) : leur propre système de référence également fondé sur Okapi (mais sans racinisation), noté Dinh 0, et deux systèmes implémentant des stratégies de désambiguïsation de concepts médicaux (Dinh 1 et Dinh 2). Ces résultats sont obtenus dans les mêmes conditions expérimentales que les nôtres mais sur un sous-ensemble de 48 requêtes.

Le tableau 3 détaille les résultats du système de référence et des deux systèmes fondés sur la décomposition morphologique (systèmes 1 et 2, cf. section 5.2). Outre la MAP, les performances sont évaluées à l'aide des mesures classiques en RI : la précision sur les 5, 10, ... , 1 000 premiers documents (P@x), la précision moyenne

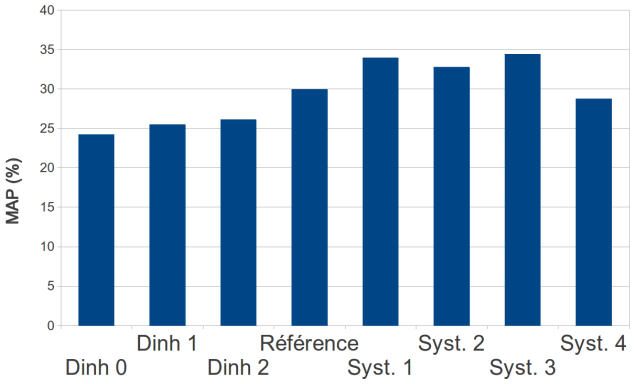


**Figure 12.** *Précision moyenne des différents systèmes de RI sur la collection* OHSUMED*/TREC-9*

| | Référence (BM-25 + racinisation) | Système 1 fondé morphèmes | Système 2 fondé kanjis |
|---|---|---|---|
| MAP | 29,93 | 33,94 (+13,4 %) | 32,76 (+9,5 %) |
| IAP | 31,74 | 35,55 (+12 %) | 34,49 (+8,6 %) |
| R-prec | 35,28 | 39,64 (+12,3 %) | 38,59 (+9,4 %) |
| P@5 | 69,87 | 73,45 (+5,1 %) | 71,70 (+2,6 %) |
| P@10 | 67,99 | 71,31 (+4,9 %) | *69,65 (+2,4 %)* |
| P@50 | 52,98 | 56,90 (+7,4 %) | 55,24 (+4,3 %) |
| P@100 | 40,86 | 44,56 (+9,1 %) | 43,39 (+6,2 %) |
| P@500 | 15,11 | 17,21 (+13,9 %) | 16,92 (+12 %) |
| P@1000 | 8,72 | 10,10 (+15,86 %) | 9,95 (+14,2 %) |

**Tableau 3.** *Performances des systèmes fondés morphèmes et kanjis sur la collection* OHSUMED*, avec les requêtes TREC*

interpolée (IAP) et la R-précision (R-prec). Pour s'assurer que les différences constatées entre deux systèmes sont statistiquement significatives, nous utilisons un test de Wilcoxon ($p = 0{,}05$) (Hull, 1993) ; les différences avec le système de référence qui ne sont pas statistiquement significatives apparaissent en italique.

Le système fondé morphèmes, reposant uniquement sur la décomposition et le groupement des morphes en morphèmes obtient d'excellents résultats, avec un gain de MAP de 13 %. Comme attendu, cette décomposition améliore plus spécifiquement les performances en fin de listes (P@100 et supérieurs), puisqu'elle permet de ramener des documents pertinents même s'ils ne contiennent pas exactement les termes de la requête. Le système fondé kanjis obtient des résultats très similaires, ce qui confirme

l'équivalence en termes de représentation sémantique des morphèmes et des kanjis. Ce résultat, un peu décevant, s'explique par deux points. D'une part, en pratique, dans certaines requêtes, il apparaît que les kanjis sont de fait trop génériques pour capturer le sens exact attendu, ou n'apportent aucune information additionnelle par rapport aux morphèmes. D'autre part, aucune sélection n'est faite sur les morphèmes à remplacer par leurs kanjis associés, et certains kanjis ont des propriétés (notamment leur fréquence documentaire) qui diffèrent fortement des morphèmes sources puisqu'ils peuvent aussi être associés à d'autres morphèmes. Une étude plus approfondie de ce phénomène qui mènerait à une pondération adaptée à ces cas est une piste importante à explorer pour utiliser cette indexation reposant sur les kanjis.

Le tableau 4 présente de la même manière les résultats des deux derniers systèmes proposés (systèmes 3 et 4, cf. section 5.2), fondés sur l'extension de requête. Ces deux

| | référence (BM-25 + racinisation) | Système 3 affinités 1er ordre | Système 4 affinités 2nd ordre |
|---|---|---|---|
| MAP | 29,93 | 34,40 (+14,9 %) | 28,74 (-3,9 %) |
| IAP | 31,74 | 36,63 (+15,4 %) | 30,80 (-2,9 %) |
| R-prec | 35,28 | 39,92 (+13,2 %) | 34,38 (-2,6 %) |
| P@5 | 69,87 | 71,76 (+2,7 %) | *68,65 (-1,7 %)* |
| P@10 | 67,99 | 70,46 (+3,6 %) | *66,20 (-2,6 %)* |
| P@50 | 52,98 | 56,30 (+6,7 %) | 50,50 (-4,68 %) |
| P@100 | 40,86 | 44,69 (+9,4 %) | 39,07 (-4,38 %) |
| P@500 | 15,11 | 17,98 (+18,9 %) | *15,01 (-0,64 %)* |
| P@1000 | 8,72 | 10,56 (+21,1 %) | *8,77 +0,66 %)* |

**Tableau 4.** *Performances des systèmes avec extension de requêtes sur la collection* OHSUMED*, avec les requêtes TREC*

systèmes ont des résultats très contrastés. En effet, l'extension avec des affinités de premier ordre donne de bons résultats, bien supérieurs à la référence. La précision est légèrement inférieure au système 1, mais son rappel est sensiblement meilleur grâce aux documents ramenés par les morphèmes étendant la requête. En revanche, les affinités de second ordre dégradent clairement les performances. Les affinités ajoutées à la requête, la plupart du temps, cassent la spécificité de l'information demandée initialement, ce qui amène le système à ramener des documents portant sur des sujets proches mais pas suffisament pertinents pour la requête. Ce résultat négatif d'extension avec des liens plus lointains est malgré tout important. Leur importance en RI biomédicale a en effet fait l'objet de plusieurs suppositions non prouvées : Namer (2007) indique que les liens de cohyponymie et coméronymie sont utiles en RI, mais Namer et Baud (2007) supposent le contraire. Il faut également noter que ce résultat est également similaire aux constatations faites sur la langue générale (Voorhees, 1998).

## 7. Conclusion et perspectives

Dans ce travail, l'utilisation originale de l'aspect multilingue des terminologies biomédicales, et notamment du japonais comme langue pivot, permet de mener une analyse morphologique fine des termes. Cette analyse ouvre de nombreuses opportunités de manipulation intelligente des termes reposant sur leurs composants morphologiques. Pour ce faire, nous avons montré qu'une simple technique d'alignement de séquences était suffisante, mais qu'elle pouvait bénéficier de plusieurs adaptations plus fines à la tâche visée. Nous avons montré d'une part l'intérêt de prendre en compte la variation morphémique durant le processus d'alignement, grâce à l'apprentissage par analogie.

Les bons résultats obtenus nous ont ainsi permis de nous attaquer aux problèmes causés par la complexité morphologique en RI biomédicale qui ne peuvent être capturés par les outils usuels tels que les raciniseurs. À ce titre, nos préoccupations sur le rôle de la morphologie pour l'accès à l'information dans le domaine biomédical ne sont pas nouvelles (Markó *et al.*, 2005a ; Deléger *et al.*, 2008), mais, à notre connaissance, nous sommes les premiers à proposer un système automatique, directement disponible pour de nombreuses langues. Bien sûr, notre approche repose essentiellement sur l'existence de l'UMLS, groupement de terminologies multilingues contenant des termes en kanjis.

De nombreuses perspectives sont ouvertes par ce travail, que ce soit des améliorations directes ou des travaux à plus long terme. Tout d'abord, d'un point de vue technique, il serait intéressant de considérer des segmentations plus complexes que les simples segmentations linéaires que nous avons implémentées. Pour ce faire, les propriétés syntaxiques des kanjis (certains attendent des arguments de type agent ou objet, par exemple), pourraient nous aider à mieux structurer les différents morphèmes. L'analyse morphosémantique, telle que nous l'effectuons dans la section 5.1, pourrait aussi s'appuyer sur les relations sémantiques entre kanjis qui peuvent être trouvées dans des dictionnaires généraux du japonais.

Concernant les aspects abordés dans les dernières sections, beaucoup de perspectives sont également ouvertes. Par exemple, notre approche actuelle ne nous permet pas de typer les liens entre morphes ; mais des heuristiques simples (telles que l'inclusion de chaîne Grabar et Zweigenbaum (2002)) ou des techniques utilisées en analyse distributionnelle pourraient permettre de mieux caractériser ces relations. Bien sûr, se pose alors le problème de l'évaluation de ce type de travaux, et particulièrement de la vérité terrain, puisque de telles ressources de référence n'existent pas. L'évaluation indirecte par une tâche comme celle de la RI que nous utilisons dans cet article pourrait dans ce cas être utilisée de nouveau, avec éventuellement d'autres collections biomédicales telles que celles de la tâche TREC Genomics (Hersch et Voorhees, 2009).

Enfin, une adaptation de ces principes aux termes complexes est envisageable. La principale difficulté est que l'hypothèse de parallélisme n'est plus réaliste pour certaines langues comme le français puisque l'ordre des mots est généralement différent de celui néoclassique ; il faut donc être capable de gérer le réordonnancement

des mots-formes composant ces termes complexes, ce qui correspond au phénomène de distorsion dans l'alignement. Cette perspective est essentielle, notamment en RI, puisque les termes complexes sont extrêmement courants mais apparaissent sous de nombreuses variantes, empêchant la mise en correspondance entre une requête et des documents contenant des variantes d'un même terme (Nenadic *et al.*, 2005). Ce problème de distorsion se poserait aussi pour l'analyse morphologique de mots de la langue générale (pour lesquels des dictionnaires bilingues pourraient facilement fournir les données) qui ne relèverait pas de l'ordre de composition néoclassique. Cependant des solutions d'alignement plus complexes gérant ces phénomènes existent et leur adaptation semble prometteuse.

Remerciements

Ce travail a été en partie financé par OSEO, l'agence française de l'innovation, dans le cadre du programme de recherche Quæro (`www.quaero.org`).

## 8. Bibliographie

Anderson S. R., « The Morpheme : Its Nature and Use », *in* M. Baerman (ed.), *The Oxford Handbook of Inflection*, Oxford University Press, 2013.

Baud R., Rassinoux A.-M., Ruch P., Lovis C., Scherrer J.-R., « The power and limits of a rule-based morpho-syntactic parser », *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association. Transforming Health Care through Informatics. AMIA*, Washington, DC, USA, p. 22-26, 1999.

Chiao Y.-C., Zweigenbaum P., « Looking for French-English translations in comparable medical corpora », *Journal of the American Medical Informatics Association*, 2002.

Claveau V., « Translation of Biomedical Terms by Inferring Rewriting Rules », *in* V. Prince, M. Roche (eds), *Information Retrieval in Biomedicine : Natural Language Processing for Knowledge Integration*, IGI - Global, 2009.

Claveau V., « Unsupervised and semi-supervised morphological analysis for Information Retrieval in the biomedical domain », *Proceedings of the Computational Linguistics (COLING) Conference*, Mumbai, Inde, 2012.

Claveau V., Kijak E., « Analyse morphologique en terminologie biomédicale par alignement et apprentissage non supervisé », *Actes de la conférence Traitement automatique des langues naturelles, TALN'10*, ATALA, Montréal, Québec, Canada, July, 2010.

Claveau V., Kijak E., « Morphological Analysis of Biomedical Terminology with Analogy-Based Alignment », *Proceedings of RANLP conference*, Hissar, Bulgaria, 2011.

Claveau V., Kijak E., « Analyse morphologique fine pour la recherche d'information biomédicale », *Actes de la conférence CORIA*, Bordeaux, France, 2012.

Claveau V., L'Homme M.-C., « Structuring Terminology by Analogy-Based Machine Learning », *Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhaguen, Denmark, 2005.

Corbin D., « Préfixe et suffixe, du sens aux catégories », *Journal of French Language Studies*, vol. 11, n° 1, p. 41-69, 2001.

Creutz M., Lagus K., Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0, Technical report, Publications in Computer and Information Science, Report A81, Helsinki University of Technology, 2005.

Daille B., « Conceptual Structuring through term Variation », *Proceedings of the ACL'03 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, Sapporo, Japon, p. 9-16, 2003.

Dal G., Amiot D., « La composition néoclassique en français et ordre des constituants », *in* D. Amiot (ed.), *La composition dans une perspective typologique*, Arras : Artois Presse Université, p. 89-113, 2008.

Deléger L., Namer F., Zweigenbaum P., « Morphosemantic parsing of medical compound words : Transferring a French analyzer to English », *International Journal of Medical Informatics*, vol. 78, p. 48-55, 2008. Supplement 1.

Dinh D., Tamine L., « Recherche d'information sémantique dans les documents biomédicaux : approche basée sur le sens précis des concepts », *Actes du XXVIII° congrès INFORSID*, 2010.

Fradin B., « On a semantically grounded difference between derivation and compounding », *in* W. U. Dressler, D. Katovsky, F. Rainer (eds), *Morphology and its Demarcations*, Amsterdam / Philadelphia : John Benjamins, 2005.

Fung P., Yee L. Y., « An IR Approach for Translating New Words from Non-parallel, Comparable Texts », *Proc. of 36th Annual Meeting of the Association for Computational Linguistics ACL*, Montréal, Canada, 1998.

Gaussier E., « Unsupervised Learning of Derivational Morphology from Inflectional Corpora », *Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics, ACL 99*, Maryland, États-Unis, p. 24-30, 1999.

Grabar N., Terminologie médicale et morphologie. Acquisition de ressources morphologiques et leur utilisation pour le traitement de la variation terminologique, Thèse de doctorat, Université Paris 6, 2004.

Grabar N., Zweigenbaum P., « Lexically-based terminology structuring : Some inherent limits », *Proc. of International Workshop on Computational Terminology, COMPUTERM*, Taipei, Taiwan, 2002.

Gupta M. R., Chen Y., « Theory and Use of the EM Algorithm », *Foundations and Trends in Signal Processing*, vol. 4, n° 3, p. 223-296, 2010.

Harastani R., Daille B., Morin E., « Neoclassical Compound Alignments from Comparable Corpora », *Proceedings of the CICLing conference*, New Dehli, Inde, 2012.

Hathout N., « Acquisition morphologique à partir d'un dictionnaire informatisé », *Actes de la conférence Traitement automatique des langues naturelles, TALN'09*, Senlis, France, 2009.

Heineman G. T., Pollice G., Selkow S., *Algorithms in a Nutshell*, Oreilly Media, chapter Chapter 8 :Network Flow Algorithms, 2008.

Hersch W., Voorhees E., « TREC genomics special issue overview », *Information Retrieval*, vol. 12, n° 1, p. 1-15, 2009.

Hull D., « Using Statistical Testing in the Evaluation of Retrieval Experiments », *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis, 1993.

Iacobini C., « Distinguishing derivational prefixes from initial combining forms », *Proceedings of the First Mediterranean Morphology Meeting*, Mytilene, Greece, 1999.

Jacquemin C., « Guessing morphology from terms and corpora », *Proceedings of the 20th International Conference on Research and Development in Information Retrieval, SIGIR 97*, Philadelphie, États-Unis, p. 156-165, 1997.

Jiampojamarn S., Kondrak G., , Sherif T., « Applying many-to-many alignments and hidden Markov models to letter-to-phoneme conversion », *Proc. of the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA, 2007.

Jiang J., Zhai C., « An empirical study of tokenization strategies for biomedical information retrieval », *Information Retrieval*, vol. 10, nº 4-5, p. 341-363, 2007.

Knight K., Graehl J., « Machine Transliteration », *Computational Linguistics*, vol. 24, nº 4, p. 599-612, 1998.

Krovetz R., « Viewing Morphology as an Inference Process », *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, États-Unis, 1993.

Kurimo M., Creutz M., Turunen V., « Morpho challenge evaluation by information retrieval experiments », *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, Springer-Verlag, Berlin, Heidelberg, p. 991-998, 2009.

Kurimo M., Virpioja S., Turunen V. T., *(Eds), Proceedings of the MorphoChallenge 2010*, Espoo, Finlande, 2010.

Langlais P., Patry A., « Translating Unknown Words by Analogical Learning », *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, p. 877-886, June, 2007.

Langlais P., Patry A., « Enrichissement d'un lexique bilingue par apprentissage analogique », *Traitement automatique des langues*, vol. 49, nº 1, p. 13-40, 2008.

Langlais P., Yvon F., Zweigenbaum P., « Translating Medical Words by Analogy », *Proc. of the workshop on Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP) 2008*, Washington, DC, 2008.

Lepage Y., « Languages of analogical strings », *Proc. of the 18th conference on Computational linguistics, COLING'00*, Universität des Saarlandes, Saarbrücken, Germany, 2000.

Lu W.-H., Lin S.-J., Chan Y.-C., Chen K.-H., « Semi-Automatic Construction of the Chinese-English MeSH Using Web-Based Term Translation Method », *Proc. of AMIA annual symposium*, 2005.

Markó K., Schulz S., Han U., « Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain », *Methods of Information in Medicine*, 2005a.

Markó K., Schulz S., Medelyan O., Hahn U., « Bootstrapping Dictionaries for Cross-Language Information Retrieval », *Proceedings of the 28th Annual International ACM SIGIR Confe-*

*rence on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brésil, 2005b.

Mel'čuk I., *Aspects of the Theory of Morphology*, Trends in Linguistics. Studies and Monographs, Mouton de Gruyter, Berlin, March, 2006.

Mitchell T. M., *Machine Learning*, McGraw-Hill, 1997.

Moreau F., Claveau V., Sébillot P., « Automatic morphological query expansion using analogy-based machine learning », *Proceedings of the European Conference on Information Retrieval, ECIR'07*, Rome, Italie, avril, 2007.

Moreau F., Sébillot P., Contributions des techniques du traitement automatique des langues à la recherche d'information, Technical Report n° 1690, IRISA, 2005.

Morin E., Daille B., « Compositionality and lexical alignment of multi-word terms », *Language Resources and Evaluation (LRE)*, 2010.

Nakamura-Delloye Y., Alignement automatique de textes parallèles français-japonais, Thèse de doctorat en linguistique, Université Paris 7, 2007.

Namer F., « Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue », *Traitement Automatique des Langues – TAL*, vol. 46, n° 2, p. 157-181, 2007.

Namer F., Baud R., « Defining and relating biomedical terms : towards a cross-language morphosemantics-based system », *International Journal of Medical Informatics*, vol. 76, p. 226-233, 2007.

Namer F., Zweigenbaum P., « Acquiring meaning for French Medical Terminology : contribution of Morphosemantics », *Proceedings of the 11th MEDINFO International Conference*, San Francisco, USA, p. 535-539, 2004.

Nenadic G., Spasic I., , Ananiadou S., « Mining biomedical abstracts : What's in a term ? », *Proceedings of the IJCNLP 2004*, vol. 3248 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, p. 797-806, 2005.

Och F. J., Ney H., « A Systematic Comparison of Various Statistical Alignment Models », *Computational Linguistics*, vol. 29, n° 1, p. 19-51, 2003.

Onogi Y., Ohe K., Tanaka M., Nozoe A., Sasaki T., Sato M., Kikuchi Y., Shinohara T., Suzuki H., Kaihara S., Seyama Y., « Mapping Japanese Medical Terms to UMLS Metathesaurus », *Medinfo 2004 : Proceedings of the 11th World Conference on Medical Informatics*, San Francisco, USA, 2004.

Ounis I., Amati G., Plachouras V., He B., Macdonald C., Lioma C., « Terrier : A High Performance and Scalable Information Retrieval Platform », *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

Porter M. F., « An Algorithm for Suffix Stripping », *Program*, vol. 14, p. 130-137, 1980.

Rabiner L. R., « A tutorial on hidden Markov models and selected applications in speech recognition », *Proceedings of the IEEE*, p. 257-286, 1989.

Robertson S. E., Walker S., Hancock-Beaulieu M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proceedings of the 7th Text Retrieval Conference, TREC-7*, p. 199-210, 1998.

Robertson S., Hull D., « The TREC-9 filtering track final report », *Proceedings of the TREC 2000 conference*, 2000.

Schulz S., Markó K., Sbrissia E., Nohama P., Hahn U., « Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon », *Proc. of the 20th International Conference on Computational Linguistics, COLING'04*, Geneva, Switzerland, 2004.

Stroppa N., Yvon F., « An analogical learner for morphological analysis », *Proceeedings of the 9th CoNLL*, Ann Arbor, MI, USA, p. 120-127, 2005.

Trieschnigg D., Kraaij W., de Jong F., « The influence of basic tokenization on biomedical document retrieval », *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, ACM, New York, NY, USA, p. 803-804, 2007.

Tsuji K., Daille B., Kageura K., « Extracting French-Japanese Word Pairs from Bilingual Corpora based on Transliteration Rules », *Proc. of the 3rd International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Spain, 2002.

Tuttle M., Sherertz D., Olson N., Erlbaum M., Sperzel D., Fuller L., Neslon S., « Using Meta-1 – the 1st Version of the UMLS Metathesaurus », *Proc. of the 14th annual Symposium on Computer Applications in Medical Care (SCAMC)*, Washington, USA, p. 131-135, 1990.

Voorhees E., *C. Fellbaum (ed.), WORDNET : An Electronic Lexical Database*, The MIT Press, chapter Using WORDNET for Text Retrieval, p. 285-303, 1998.