*TALN 2010, Montréal, 19–23 juillet 2010*

# Analyse morphologique en terminologie biomédicale par alignement et apprentissage non-supervisé

Vincent Claveau[1] Ewa Kijak[2]
(1) IRISA-CNRS, (2) IRISA-Univ. Rennes 1
Campus de Beaulieu, F-35042 Rennes, France
Vincent.Claveau@irisa.fr, Ewa.Kijak@irisa.fr

**Résumé.** Dans le domaine biomédical, beaucoup de termes sont des composés savants (composés de plusieurs racines gréco-latines). L'étude de leur morphologie est importante pour de nombreuses applications puisqu'elle permet de structurer ces termes, de les rechercher efficacement, de les traduire...
Dans cet article, nous proposons de suivre une démarche originale mais fructueuse pour mener cette analyse morphologique sur des termes simples en français, en nous appuyant sur une langue pivot, le japonais, et plus précisément sur les termes écrits en kanjis. Pour cela nous avons développé un algorithme d'alignement de termes spécialement adapté à cette tâche. C'est cet alignement d'un terme français avec sa traduction en kanjis qui fournit en même temps une décomposition en morphe et leur étiquetage par les kanjis correspondants.
Évalué sur un jeu de données conséquent, notre approche obtient une précision supérieure à 70 % et montrent son bien fondé en comparaison avec les techniques existantes. Nous illustrons également l'intérêt de notre démarche au travers de deux applications directes de ces alignements : la traduction de termes inconnus et la découverte de relations entre morphes pour la tructuration terminologique.

**Abstract.** In the biomedical domain, many terms are neoclassical compounds (composed of several Greek or Latin roots). The study of their morphology is important for numerous applications since it makes it possible to structure them, retrieve them efficiently, translate them...
In this paper, we propose an original yet fruitful approach to carry out this morphological analysis by relying on Japanese, more precisely on terms written in kanjis, as a pivot language. In order to do so, we have developed a specially crafted alignment algorithm. This alignment process of French terms with their kanji-based counterparts provides at the same time a decomposition of the French term into morphs, and a kanji label for each morph.
Evaluated on a big dataset, our approach yields a precision greater than 70% and shows its the relevance compared with existing techniques. We also illustrate the validity of our reasoning through two direct applications of the produced alignments: translation of unknown terms and discovering of relationships between morphs for terminological structuring.

**Mots-clés :** Alignement, terminologie, morphologie, analogie, traduction de terme, kanji.

**Keywords:** Alignment, terminology, morphology, analogy, term translation, kanji.

# 1 Introduction

Dans beaucoup de domaines, l'accès à l'information est guidé par des l'emploi de termes bien déterminés, formant une terminologie de ce domaine. C'est notamment le cas dans le domaine biomédical, dans lequel les nombreuses terminologies servent à structurer le savoir, et à y accéder, par exemple au travers de la terminologie MeSH qui sert pour accéder à la populaire base de documents PubMed. Savoir manipuler ces termes, les comprendre, les traduire, établir des liens sémantiques entre eux, sont donc des opérations essentielles pour les applications comme l'enrichissement de lexique bilingue, et plus généralement la traduction artificielle, la recherche d'informations...

C'est dans ce cadre que se situe le travail présenté dans cet article : nous nous intéressons à la morphologie des termes simples français du domaine biomédical comme base de l'analyse terminologique. Plus précisément, nous présentons une technique visant à décomposer morphologiquement un terme simple et en associant des connaissances sur chacun des morphes obtenus. Nous reprenons donc le cadre de travail adopté par certains travaux (Namer, 2005; Markó *et al.*, 2005, par exemple), mais en tentant de supprimer la coûteuse intervention humaine qu'ils requièrent.

L'idée originale de cet article est d'utiliser l'aspect multilingue de certaines terminologies pour utiliser une langue pivot particulière, le japonais, et plus précisément les termes écrits en kanjis, pour mener la décomposition en morphes[1] et leur associer les kanjis correspondants sans intervention humaine. Nous voulons donc faire jouer aux kanjis le rôle de représentations sémantiques des morphes. L'avantage des kanjis est que la morphologie de tels termes est une simple concaténation de mots plus élémentaires qu'il est facile de traduire en utilisant un dictionnaire généraliste. Par exemple, le terme photochimiothérapie se traduit en japonais par 光化学療法 ; la décomposition et l'alignement de ces deux termes produit : photo ↔ 光 ('lumière'), chimio ↔ 化学 ('chimie'), thérapie ↔ 療法 ('thérapie'). Notre démarche fait donc l'hypothèse que la composition des termes en kanjis suit fidèlement celle des termes simples français. Cette hypothèse peut paraître péremptoire, mais les résultats présentés dans cet article montre que c'est une hypothèse parfois mise en défaut mais raisonnable.

Notre analyse morphologique passe donc par une étape essentielle d'alignement des termes français et japonais issus d'une terminologie multilingue. Pour la mettre en œuvre, nous proposons une technique d'alignement particulièrement adaptée à la manipulation de ce type de données. Après une présentation de travaux proches en termes applicatifs ou méthodologiques, nous décrivons dans la section 3 cette technique d'alignement, mêlant algorithme *Forward-Backward* et apprentissage par analogie dans la section 3. Nous en présentons les résultats en section 4, et la section 5 illustre tout l'intérêt d'une telle notre approche avec deux applications, l'une de traduction, montrant comment des termes inconnus peuvent être manipulés, et l'autre d'analyse terminologique, soulignant l'intérêt de la décomposition en morphème pour cette tâche.

# 2 Travaux connexes

Les travaux utilisant la morphologie pour mener une analyse terminologique sont nombreux. C'est particulièrement le cas pour domaine biomédical dans lequel ces terminologies sont au cœur de nombreuses applications et où certaines opérations morphologiques, comme la composition savante, sont très pré-

[1] Dans cet article, nous distinguons à la Mel'čuk les morphes, signes (segments) linguistiques élémentaires des mots-formes, des morphèmes, classes d'équivalence de morphes de signifiés identiques et de signifiants proches.

sentes. On peut distinguer deux visions de l'usage de la morphologie pour produire des analyses de termes (ou de mots). Il y a la vision léxématique dans laquelle les relations entre termes reposent sur la forme des mots mais sans qu'il soit cherché à les découper en morphème (eg. Grabar & Zweigenbaum, 2002; Claveau & L'Homme, 2005; Hathout, 2009). À cette utilisation implicite de la morphologie s'oppose la vision morphémique dans laquelle l'étape de base consiste à découper un terme en morphe. Beaucoup de travaux ont été faits dans ce cadre. Ils adoptent soit des approches partiellement manuelles comme celles déjà citées (Namer, 2005; Markó *et al.*, 2005) dans lesquelles les morphes et les règles de combinaison sont données par un expert, soit des approches plus automatiques s'appuyant le plus souvent sur la récurrence de certaines suites de lettres dans une liste de terme pour en faire des candidats morphes. Ces dernières techniques ne permettent alors pas d'associer une valeur sémantique aux décompositions produites. À notre connaissance, aucun travail ne s'appuie sur une langue pivot pour mener automatiquement une telle analyse morphologique comme nous le proposons ici.

D'un point de vue plus technique, notre recours à l'utilisation d'une terminologie bilingue évoque également les travaux en translittération, notamment du katakana ou de l'arabe (Tsuji *et al.*, 2002; Knight & Graehl, 1998, par exemple), ou en traduction. Dans ce cadre, il convient de citer les travaux proches de Morin & Daille (2010). Ceux-ci proposent de mettre en correspondances des termes complexes japonais et français en utilisant, entre autres, une technique s'appuyant sur la morphologie, mais les règles morphologiques qu'ils emploient sont purement manuelles et ne s'appliquent qu'à un cas précis de dérivation, leur approche n'étant pas adaptée aux composés savants. Dans des travaux précédents, nous avons également proposé une technique de traduction de termes biomédicaux (Claveau, 2009) considérant uniquement les termes comme des séquences de lettres. Même si la finalité est différente, il faut noter qu'à l'instar de l'approche présentée dans cet article, ces travaux de traduction et de translittération partagent une parenté technique. En effet, tous requièrent une phase d'alignement des termes ou mots. Cet alignement se fait dans la majorité des cas à l'aide d'algorithmes 1-1, capables d'aligner un symbole (lettre ou caractère vide) de la langue source avec un symbole de la langue cible. Cependant, dans des travaux récents de phonétisation, certains auteurs ont montré que l'alignement dit *many-to-many* offrait des résultats intéressants (Jiampojamarn *et al.*, 2007).

# 3 Approche

Notre approche d'alignement est basée sur un algorithme de type *Expectation-Maximization* (EM) dont nous présentons brièvement le fonctionnement dans la sous-section suivante (le lecteur intéressé peut se reporter à Jiampojamarn *et al.*, 2007, pour plus de détails et un exemple d'application). La deuxième sous-section explique les modifications apportées à cet algorithme afin qu'il puisse gérer naturellement et automatiquement la variation morphologique inhérente à notre problème de découpage en morphes.

## 3.1 Alignement EM

L'algorithme à la base de notre approche est relativement standard : il s'agit d'un algorithme de type *Baum-Welch*, étendu pour mettre en correspondance des sous-séquences de symboles et non plus seulement des alignements 1-1. Dans notre cas, il prend en entrée des termes français et leurs traductions en kanjis, par exemple extraits d'une terminologie multilingue. La longueur maximale des sous-séquences pouvant être alignées pour les kanjis et les lettres est paramétrée par $maxX$ et $maxY$. Pour chaque paire de

termes $x^T, y^V$ à aligner ($T$ et $V$ sont les longueurs des termes en kanjis et en lettres), l'algorithme EM (algorithme 1) calcule tout d'abord dans une table $\gamma$, les comptes partiels de toutes les correspondances possibles entre sous-séquences de kanjis et de lettres (*Expectation*). Ces comptes sont ensuite utilisés pour estimer les probabilités d'alignement dans $\delta$ (*Maximization*).

La phase *Expectation* (algorithme 2) utilise une approche *forward-backward* : elle calcule les probabilités *forward* $\alpha$ et *backward* $\beta$. Pour chaque position $t, v$ dans les 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 chaînes jusqu'au point courant, en fonction des probabilités d'alignement $\delta$ courantes (cf. algorithme 4). $\beta(t, v)$ est calculé de manière analogue en considérant $(x_t^T, y_v^V)$. Ces probabilités sont alors utilisées pour le calcul des comptes $\gamma$. Dans cette version de l'algorithme, l'étape de *Maximization* (algorithme 3) consiste simplement à construire les probabilités d'alignement $\delta$ en normalisant les comptes de $\gamma$.

**Algorithme 1** *Algorithme EM* (d'après Jiampojamarn *et al.*, 2007)

Entrées : liste de paires $x^T$ , $y^V$ , maxX, maxY
**tant que** changements dans $\delta$ **faire**
 initialisation de $\gamma$ à 0
 **pour tout** paire ($x^T$ , $y^V$ ) **faire**
  $\gamma$ = Expectation($x^T$ , $y^V$ , $maxX$ , $maxY$ , $\gamma$)
 $\delta$ = Maximization($\gamma$)
**renvoit** $\delta$

**Algorithme 2** *Expectation*

Entrées : $x^T$ , $y^V$ , $maxX, maxY, \gamma$
$\alpha :=$ Forward-many2many( $x^T$ , $y^V$ , $maxX, maxY$ )
$\beta :=$ Backward-many2many( $x^T$ , $y^V$ , $maxX, maxY$ )
**si** $\alpha_{T,V} > 0$ **alors**
 **pour** $t = 1...T$ **faire**
  **pour** $v = 1...V$ **faire**
   **pour** $i = 1...maxX$ st $t - i \geq 0$ **faire**
    **pour** $j = 1...maxY$ st $v - j \geq 0$ **faire**
     $\gamma(x_{t-i+1}^t, y_{v-j+1}^v) \mathrel{+}= \frac{\alpha_{t-i,v-j}\delta(x_{t-i+1}^t, y_{v-j+1}^v)\beta_{t,v}}{\alpha^{T,V}}$
**renvoit** $\gamma$

**Algorithme 3** *Maximization*

Entrée : $\gamma$
**pour tout** sous-séquence $a$ tq $\gamma(a, \cdot) > 0$ **faire**
 **pour tout** sous-séquence $b$ tq $\gamma(a, b) > 0$ **faire**
  $\delta(a, b) = \frac{\gamma(a,b)}{\sum_x \gamma(a,x)}$
**renvoit** $\delta$

**Algorithme 4** *Forward-many2many*

Entrée : ($x^T$ , $y^V$ , $maxX, maxY$ )
$\alpha_{0,0} := 1$
**pour** $t = 0...T$ **faire**
 **pour** $v = 0...V$ **faire**
  **si** $(t > 0 \vee v > 0)$ **alors**
   $\alpha_{t,v} = 0$
  **si** $(v > 0 \wedge t > 0)$ **alors**
   **pour** $i = 1...maxX$ st $t - i \geq 0$ **faire**
    **pour** $j = 1...maxY$ st $v - j \geq 0$ **faire**
     $\alpha_{t,v} \mathrel{+}= \delta(x_{t-i+1}^t, y_{v-j+1}^v)\alpha_{t-i,v-j}$
**renvoit** $\alpha$

Ce processus EM est itéré jusqu'à stabilisation des probabilités d'alignement $\delta$. Une fois la convergence atteinte, l'alignement consiste alors simplement à trouver le découpage maximisant $\alpha(T, V)$. Outre l'alignement, nous conservons aussi les probabilités d'alignement $\delta$ finales qui serviront à traiter de nouveaux termes (cf. section 5.1).

Cette technique n'est pas fondamentalement différente de celles utilisées en traduction statistique. À ce titre quelques remarques s'imposent : cette approche permet de gérer la fertilité c'est-à-dire l'alignement avec un morphe vide, la version simplifiée présentée ici n'en rend pas compte par soucis de place ; en revanche, la distorsion ne peut pas être prise en compte sans modification majeure de l'algorithme.

## 3.2 Normalisation morphologique automatique

La phase de maximisation de l'algorithme construit simplement les probabilités de traduction d'une suite de kanjis en une suite de lettres. Par exemple, pour le kanji 菌 (*'bactérie'*), il peut y avoir une entrée de $\delta$ l'associant à la suite de lettres bactérie, une autre pour bactério (comme dans bactério/lyse) et encore une pour bactéri (dans myco/bactéri/ose), chacune avec une certaine probabilité. Cette dispersion des probabilités, évidemment néfaste à l'algorithme, est donc causée par la variation morphémique : bactério, bactérie, bactéri étant trois morphes d'un même morphème, on souhaite que leurs probabilités se renforcent. L'adaptation que nous apportons vise à ce que la phase de maximisation soit capable de regrouper automatiquement les morphes relevant d'un même morphème.

Pour ce faire, nous utilisons une approche simple mais bien adaptée à ce problème s'appuyant sur le calcul d'analogies. Une analogie est une relation entre 4 éléments que l'on note $a : b :: c : d$ et qui se lit comme « $a$ est à $b$ ce que $c$ est à $d$ » (Lepage, 2003, pour plus d'informations sur les analogies). L'analogie a été utilisée dans de nombreux travaux de TAL, notamment pour la traduction de phrases (Lepage, 2003) ou de termes (Langlais & Patry, 2008). C'est aussi un calcul d'analogie qui est utilisé dans les travaux de structuration de terminologie précédemment cités (Claveau & L'Homme, 2005). Nous nous inspirons de ces derniers travaux pour formaliser notre problème. Dans notre cadre, une telle analogie sera par exemple dermato : dermo :: hémato : hémo. Sachant que dermato et dermo appartiennent à un même morphème, on peut en déduire que hémato et hémo aussi. La mise en œuvre pratique de cette analogie consiste à apprendre une règle de réécriture de préfixe et de suffixe permettant de passer de dermato à dermo et de vérifier que cette règle s'applique bien à hémato-hémo.

La différence principale avec (Claveau & L'Homme, 2005), outre le fait que nous travaillons cette fois sur des morphes et non plus sur des termes complets, est que nous ne disposons pas préalablement d'exemples de morphes en relation (comme dermato et dermo dans l'exemple précédent). Nous utilisons donc une technique d'amorçage qui consiste à supposer que pour une séquence de kanjis donnée, si deux morphes sont répertoriés dans $\gamma$ comme traductions possibles de cette séquence de kanjis et que ces deux morphes partagent une sous-chaîne commune plus longue qu'un certain seuil, ils appartiennent à un même morphème. De ces amorces sont alors construites les règles de préfixation et suffixation permettant de vérifier les analogies. Une règle trouvée plusieurs fois ayant plus de chances d'être correcte, nous conservons à chaque itération de l'algorithme EM toutes les règles générées et leur nombre d'occurrences et n'appliquons que les plus fréquemment trouvées. L'ensemble du processus est donc entièrement automatique. Cette nouvelle étape de *Maximization* est résumée dans l'algorithme 5. Elle assure que tous les morphes reconnus comme relevant du même morphème aient des probabilités d'alignements égales et renforcées.

**Algorithme 5** *Maximization* avec normalisation par analogie

Entrées : $\gamma$
**pour tout** sous-séquence $a$ tq $\gamma(a,\cdot) > 0$ **faire**
 **pour tout** $m_1, m_2$ tq $\gamma(a,m_1) > 0 \wedge \gamma(a,m_2) > 0 \wedge$ plus_longue_sous_chaîne_commune$(m_1,m_2)$ > seuil **faire**
  trouver la règle $r$ de préfixation et suffixation entre $m_1, m_2$
  incrémenter le score de $r$
 **pour tout** sous-séquence $b$ tq $\gamma(a,b) > 0$ **faire**
  constituer l'ensemble $\mathcal{M}$ des morphes liés à $b$ à l'aide des $n$ règles analogiques les plus fréquentes de l'itération précédente
  $$\delta(a,b) = \frac{\sum_{c\in\mathcal{M}} \gamma(a,c)}{\sum_x \gamma(a,x)}$$
**renvoit** $\delta$

# 4 Expérimentations

## 4.1 Données

Les données que nous utilisons sont issues du métathésaurus de l'UMLS (Tuttle *et al.*, 1990), réunion de plusieurs terminologies pour plusieurs langues. Le métathésaurus associe aux termes un identifiant indépendant des langues qui nous permet d'extraire facilement des paires de termes japonais/français. Seuls les termes composés uniquement de kanjis sont conservés pour le japonais, et seuls les termes simples sont conservés pour le français, auxquels on ajoute un marqueur de fin de chaîne (' ;'). Cela permet de former 8000 paires.

Parmi ces 8000 paires qui constituent nos données à aligner, nous en avons manuellement aligné 1600 sélectionnées aléatoirement. Elles vont permettre d'évaluer les résultats de notre technique d'alignement.

## 4.2 Résultats d'alignement

L'évaluation de notre approche est faite au travers de la précision, calculée sur l'alignement complet du terme japonais avec le terme français (c'est donc le pendant du *sentence error rate* utilisé en traduction artificielle).

Pour chaque paire, l'algorithme EM indique la probabilité de l'alignement proposé. En classant les alignements des plus sûrs aux moins sûrs, on peut calculer une précision en fonction du nombre de termes alignés. La figure 1 présente les résultats obtenus sur les 1600 paires de test. Nous y indiquons les courbes obtenues par l'algorithme EM avec et sans normalisation morphémique. À titre de comparaison, nous indiquons également les résultats obtenus par GIZA++ (Och & Ney, 2003), outil de référence en traduction artificielle. Pour ce dernier, différents modèles IBM et paramètres ont été testés ; les résultats reportés sont les meilleurs obtenus (avec le modèle IBM 4).

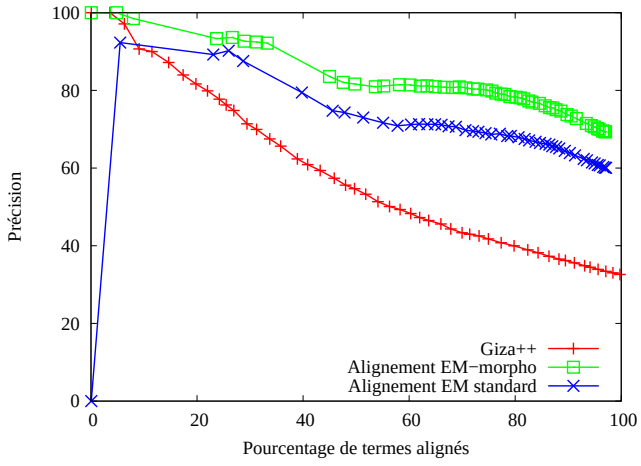


FIG. 1 – Précision de l'alignement en fonction du nombre de termes alignés

Comme attendu, on constate sur cette figure l'intérêt de la technique d'alignement embarquant la normalisation morphémique, avec 70 % de précision au pire cas. Cette normalisation permet en effet une amélioration de 10 % de la précision quel que soit le nombre de mots alignés.

Un examen manuel des résultats montre que la principale cause d'erreur est due à l'invalidation de notre hypothèse de travail : certaines paires de termes français-kanjis ne sont pas décomposables de la même manière. Par exemple le terme français anxiolytiques peut être rendu en kanjis par une séquence signifiant littéralement 'médicament pour la dépression'. Parmi ces erreurs, certaines relèvent de termes qui ne sont pas des composés savants et sont donc indécomposables, soit dans une des langues, soit dans les deux (*eg.* méninges se traduit par 脳膜 'membrane de cerveau'). L'autre part d'erreurs est simplement causée par un problème de couverture des données : certains morphes ou certaines séquences de lettres apparaissent une fois seulement, ou toujours combinés avec un même autre morphe, ce qui conduit à de mauvaises segmentations.

# 5 Utilisation des alignements morphes/kanjis

Dans cette section, nous présentons deux applications que notre approche d'analyse morphologique par alignement rend possible. La première montre comment traduire et décomposer un terme inconnu et la deuxième illustre les possibilités d'analyse offertes par ces alignements.

## 5.1 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). 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 & Zweigenbaum, 2002; Fung & 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 & Patry, 2008; Claveau, 2009). 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 traductions 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. Il est donc segmenté en morphe et les morphes sont étiquetés par les kanjis correspondants. Cela produit donc une analyse morpho-sémantique 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 $\delta$. 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 éléments 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 les cas d'erreurs, les premières traductions proposées sont souvent des paraphrases correctes bien que non attestées dans l'UMLS (mais attestées sur le Web par exemple).

## 5.2 Analyse des morphes

Une fois tous les termes alignés, on peut étudier les correspondances récurrentes de morphes français avec les kanjis. Ces correspondances peuvent être mise en lumière par différentes techniques : des treillis de Galois (les kanjis représentant l'intention et les morphes l'extension), ou comme pour une analyse distributionnelle, des analyses en termes de graphes *petit-monde*, de composantes connexes du graphe, de cliques... Dans cet article, nous utilisons une représentation sous forme de graphe : les nœuds sont les morphèmes (détectés grâce aux règles analogiques produites lors de l'alignement ; on les représente ici sous la forme d'un morphe représentatif) et les kanjis, et les arcs sont valués selon le nombre de fois que tel morphème est aligné avec telle séquence de kanjis lors de l'alignement des 8000 paires de l'UMLS.

Cela nous permet d'explorer facilement les différents types de voisinage d'un morphème : son nœud reçoit une quantité d'énergie qui est transmise inversement proportionnellement aux valeurs des arcs. Les figures 2 et 3 présentent respectivement les kanjis (traduits manuellement) et les morphèmes atteints, sous forme de nuage de mots (la taille et la couleur représentent donc l'énergie atteignant ce nœud) pour le morphème ome (cancer). Les nœuds atteints ainsi sont ceux conceptuellement proches, censés exhiber des relations de traduction ou de synonymie, comme on peut le voir dans ces deux exemples.


FIG. 2 – Nuage de kanjis pour ome

En étudiant les cooccurrences des morphèmes dans les termes français, on peut également étudier les affinités de premier ordre (morphèmes fréquemment associés à tel morphème) et celles de second ordre (morphèmes ayant les mêmes cooccurrants que tel morphème). Cela doit permettre de grouper des morphèmes par paradigme. C'est le cas par exemple dans la figure 4 où l'on constate que les morphèmes automatiquement associés à gastro relèvent pour la plupart d'organes, les plus fortement associés étant proches anatomiquement. Ces quelques informations de natures différentes (bien d'autres exploitations de

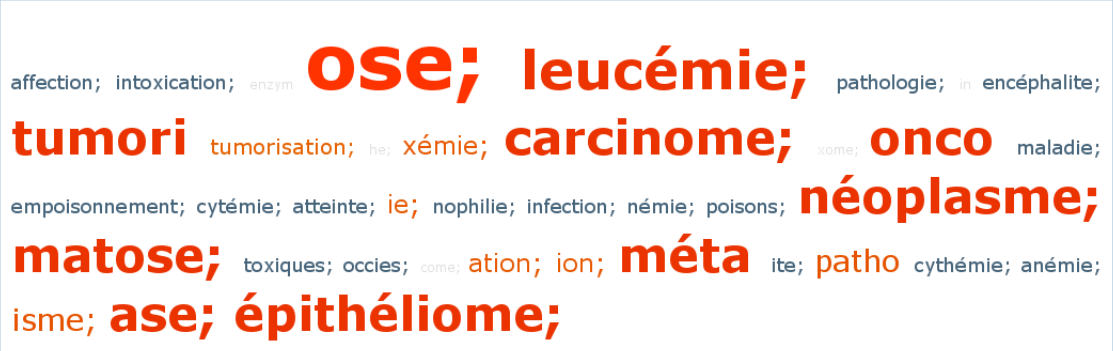


FIG. 3 – Nuage de morphèmes pour ome

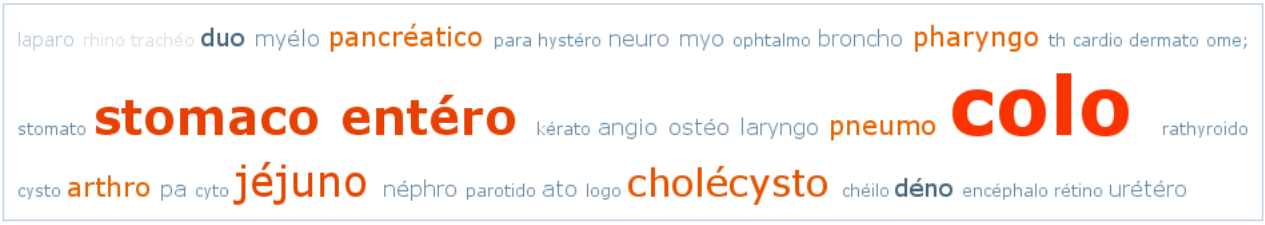


FIG. 4 – Nuage des affinités de second ordre du morphème gastro

ces alignements sont possibles) permettent ainsi d'identifier des relations entre termes, ou de construire des termes proches, ou encore d'explorer la base terminologique sur ces bases morphologiques.

# 6 Conclusion

L'approche d'alignement que nous avons présentée et l'idée originale qu'elle met en œuvre — utiliser une autre langue pour guider la décomposition en morphes et leur analyse — offrent de nombreuses possibilités de manipulation des termes à moindre coût. Du fait de cette entière automatisation, nous sommes bien sûr loin de la capacité de systèmes d'analyse comme celui de Namer (2005) qui permet par exemple de hiérarchiser la décomposition d'un terme mais dont l'aspect manuel limite le développement. Cependant, beaucoup de perspectives et d'améliorations sont envisageables. Concernant la linéarité du découpage, on pourrait par exemple exploiter la nature de certains kanjis dont le fonctionnement syntaxique, comme l'attente d'un argument, est connu. Dans le même ordre d'idée, on pourrait aussi exploiter les relations sémantiques entre kanjis, assez faciles à trouver dans un dictionnaire généraliste.

Concernant les aspects d'analyse illustrés dans la dernière section, beaucoup de possibilités sont là-aussi à l'étude. Les liens entre morphes tels que nous les produisons n'étant pas typés, l'emploi d'heuristiques (comme l'inclusion de chaînes utilisée par Grabar & Zweigenbaum (2002)) ou de techniques issues de l'analyse distributionnelle pourrait apporter des réponses. La question de l'évaluation pour ce type de travaux, et plus particulièrement de la construction de la vérité terrain se pose alors. Enfin, une adaptation de ces principes pour les termes complexes est à l'étude. La principale difficulté est de gérer les réordonnancements des mots composant ces termes, et donc de gérer la distorsion dans l'algorithme d'alignement.

# Références

CHIAO Y.-C. & ZWEIGENBAUM P. (2002). Looking for french-english translations in comparable medical corpora. *Journal of the American Medical Informatics Association*, **8**(suppl).

CLAVEAU V. (2009). 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.

CLAVEAU V. & L'HOMME M.-C. (2005). Structuring terminology by analogy-based machine learning. In *Proc. of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*, Copenhague, Danemark.

FUNG P. & YEE L. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of 36th Annual Meeting of the Association for Computational Linguistics ACL*, Montréal, Canada.

GRABAR N. & ZWEIGENBAUM P. (2002). Lexically-based terminology structuring : Some inherent limits. In *Proc. of International Workshop on Computational Terminology, COMPUTERM*, Taipei, Taiwan.

HATHOUT N. (2009). Acquisition morphologique à partir d'un dictionnaire informatisé. In *Actes de la 16e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles*, Senlis, France.

JIAMPOJAMARN S., KONDRAK G., & SHERIF T. (2007). Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Proc. of the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, USA.

KNIGHT K. & GRAEHL J. (1998). Machine transliteration. *Computational Linguistics*, **24**(4), 599–612.

LANGLAIS P. & PATRY A. (2008). Enrichissement d'un lexique bilingue par apprentissage analogique. *Traitement automatique des langues*, **49**(1), 13–40.

LEPAGE Y. (2003). *De l'analogie : rendant compte de la communication en linguistique*. Thèse d'habilitation (hdr), Université de Grenoble 1.

LU W.-H., LIN S.-J., CHAN Y.-C. & CHEN K.-H. (2005). Semi-automatic construction of the chinese-english mesh using web-based term translation method. In *Proc. of AMIA annual symposium*.

MARKÓ K., SCHULZ S. & HAN U. (2005). Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, **44**(4).

MORIN E. & DAILLE B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation (LRE)*, **44**.

NAMER F. (2005). Morphosémantique pour l'appariement de termes dans le vocabulaire médical : approche multilingue. In *Actes de la conférence TALN*, p. 63–72, Dourdan, France.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

SCHULZ S., MARKÓ K., SBRISSIA E., NOHAMA P. & HAHN U. (2004). Cognate Mapping - A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. In *Proc. of the 20$^{th}$ International Conference on Computational Linguistics, COLING'04*, Genève, Suisse.

TSUJI K., DAILLE B. & KAGEURA K. (2002). Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proc. of the 3$^{rd}$ International Conference on Language Resources and Evaluation, LREC'02*, Las Palmas de Gran Canaria, Espagne.

TUTTLE M., SHERERTZ D., OLSON N., ERLBAUM M., SPERZEL D., FULLER L. & NESLON S. (1990). Using meta-1 – the 1$^{st}$ version of the UMLS metathesaurus. In *Proc. of the 14$^{th}$ annual Symposium on Computer Applications in Medical Care (SCAMC)*, p. 131–135, Washington, États-Unis.