Approche pour un étiquetage morphosyntaxique du malais

Bali Ranaivo-Malançon

(1) Unit Terjemahan Melalui Komputer – Universiti Sains Malaysia
11800 Minden, Penang, Malaysia
ranaivo@cs.usm.my

Mots-clés : étiquetage morphosyntaxique, malais, corpus écrit
Keywords: POS tagging, Malay, written corpus

Résumé Dans cet article, nous proposons une méthode semi-automatique pour catégoriser les mots du lexique malais, langue officielle de la Malaisie, en partant d'un ensemble d'étiquettes très rudimentaires et d'une exploitation maximale de la structure morphologique des mots.

Abstract In this article, we propose a semi-automatic method to categorise words in Malay, official language of Malaysia. To achieve our aim, we start with a very simple tagset and take advantage of the morphological structure of words to predict their categories.

1 Introduction

Le malais, langue officielle de la Malaisie, n'est pas vraiment une langue sans ressource et un sujet relatif à cette langue pourrait ne pas apparaître dans cet atelier sur le « TAL et langues peut dotées ». Le malais est présent sur Internet (la recherche sur Google du mot « melayu » sur les pages éditées en Malaisie annonce 160000 pages) et il a fait depuis quelques années l'objet de recherche en traduction automatique dans l'unité de traduction automatique (UTMK[1]) de l'Université Scientifique de la Malaisie (USM) en coopération avec l'équipe GETA-CLIPS-IMAG de Grenoble, création de dictionnaire informatisé multilingue (FEM[2]), de base de données lexicales multilingue (Papillon[3]), traitement de la parole (principalement à l'Université Technologique de la Malaisie et l'UTMK). Toutefois, nous classons le malais dans le groupe des langues non pas peu dotées mais faiblement dotées pour la simple raison que le nombre de ressources linguistiques est très faible ou trop spécialisé (les dictionnaires utilisés par le système de traduction en ligne de MIMOS[4] sont des dictionnaires relatifs à l'agriculture et à la santé) et que les outils pour le traitement automatique du malais sont non réutilisables car implémentés sur des systèmes désuets et utilisables uniquement pour une seule application.

Nous avons adopté une approche empirique faisant appel à un corpus écrit et à des outils statistiques pour établir la liste de mots malais servant d'entrées pour un dictionnaire électronique formalisé unilingue du malais en cours de développement à l'UTMK et ajouter les catégories morphosyntaxiques à ces entrées. Ces informations grammaticales sont indispensables pour toute analyse automatique de textes malais. Aujourd'hui, aucun consensus n'existe sur la classification des mots malais et le *Dewan Bahasa dan Pustaka* (DBP), l'équivalent de l'Académie française en Malaisie, n'a toujours pas édité le dictionnaire *Kamus Dewan* [1], réactualisé avec les catégories lexicales.

Dans cet article, nous proposons une méthode semi-automatique pour commencer la classification des mots du malais en partant d'un ensemble d'étiquettes très rudimentaires et d'une exploitation maximale de la structure morphologique des mots.

---

1 UTMK : http://utmk.cs.usm.my/.

2 FEM – Français-English-Malay Dictionary : http://www-clips.imag.fr/geta/services/fem/.

3 Papillon project Web site : http://www.papillon-dictionary.org/.

4 MIMOS est une agence gouvernementale de recherche et de développement spécialisée dans le domaine des technologies de l'information et de la communication et de la microélectronique : http://www.mimos.my/.

2 Pourqoui l'étiqueteur grammatical du malais n'existe-t-il pas ?

2.1 Des projets trop orientés

Les recherches sur le traitement automatique du malais sont nombreuses et très variées en Malaisie. Le centre de traduction automatique (UTMK) de l'Université des Sciences de la Malaisie (USM) a eu un rôle de pionnier et reste la référence. Si la traduction automatique a été le point de départ de l'UTMK, actuellement ses axes de recherches se sont multipliés (communication homme machine, traitement de la parole, moteur de recherche, recherche d'information) avec une orientation vers la création de systèmes commercialisables (directive donnée par le Gouvernement malaisien) au dépend malheureusement de la recherche fondamentale. Un frémissement de recherche sur la reconnaissance de la parole du malais s'est fait sentir du côté de la Faculté d'Ingénierie Electrique de l'UTM depuis 2001. Le laboratoire d'ingénierie des langues de MIMOS, a créé 'FASIH' le premier synthétiseur malais (synthétiseur à diphones construit avec l'aide de MBROLA, Faculté Polytechnique de Mons, Belgique), et a mis en ligne un traducteur automatique malais-anglais (moteur fourni par l'UTMK).

Si beaucoup de projets concernant le traitement automatique du malais ont vu le jour en Malaisie, la majorité des données linguistiques et des outils créés ont été destinés à une seule application. Ainsi, les travaux sur la traduction automatique du malais auraient du avoir un étiqueteur grammatical du malais utilisable pour d'autres applications. Jusqu'à présent, chaque chercheur essaie d'adapter un étiqueteur créé pour d'autres langues sans jamais chercher à créer ou à adapter entièrement cet étiqueteur pour le malais. La raison généralement évoquée est le manque de références théoriques linguistiques. Le problème majeur que rencontre un taliste travaillant sur le malais est la rareté de ressources linguistiques. Le nombre d'années de recherche sur le traitement automatique du malais mené au sein de l'UTMK pourrait être un signe de présence de beaucoup de ressources linguistiques (dictionnaires électroniques formalisés, grammaires, bases de données lexicales, corpus) et d'outils. Malheureusement, la grande majorité de ces ressources et de ces outils est perdue ou inutilisable dû à un oubli de sauvegarde, un changement de plateforme ou tout simplement une absence totale de documentation.

2.2 Manque de linguiste informaticien

Un autre problème qui ne permet pas d'avoir un développement correct du traitement automatique du malais en Malaisie est le manque de linguiste informaticien. Il est souvent fréquent de voir dans le curriculum vitae d'un chercheur enseignant malaisien, la mention de '*natural language processing*' ou '*computational linguistics*' dans la partie 'Spécialité' bien que ces personnes n'aient eu qu'une formation partielle, parfois même inexistante, du traitement automatique des langues. Actuellement, il n'existe aucune formation complète du TAL en Malaisie. Le Département Informatique de l'USM offre deux cours de TAL en licence (cours optionnel) et en master. Le cours proposé pour la licence a été créé depuis deux ans et n'a été ouvert que cette année avec seulement six étudiants inscrits. Le cours porte sur l'introduction au TAL. Le cours en master, 'Traitement automatique de documents' (*Document Processing*), enregistre plus d'étudiants (une moyenne de vingt étudiants) mais ne crée pas plus de vocation chez les étudiants. Le Département Sciences Sociales offre un cours

de traduction automatique où les étudiants apprennent le fonctionnement des traducteurs automatiques et l'utilisation de Prolog avec des enseignants non informaticiens et non talistes. Quelques universités malaisiennes offrent un seul cours facultatif de TAL ou de linguistique computationnelle qui est généralement une brève introduction au sujet.

2.3 Rareté des références théoriques et inapplication des normes

Mis à part ces problèmes de gestion des données et de formation de spécialistes du TAL, le traitement automatique du malais souffre d'un problème politique lié à la gestion de la langue. Les normes grammaticales et orthographiques du malais sont énoncées et dictées par DBP à travers des guides pour écrire correctement le malais [2, 3, 4] ou transcrire les mots empruntés [5] et l'ouvrage de grammaire du malais [6]. Le problème est que ces règles sont rarement appliquées. L'exemple le plus frappant est l'utilisation du « di ». En tant que préposition, cette unité correspond à un mot orthographique : *di mana* 'où', *di sekolah* 'à l'école'. En tant que marqueur du passif, il est préfixé à la base : *dipakai* 'Passif-utiliser'. Les règles d'utilisation du « di » ont été énoncées depuis 1972 et expliquées clairement dans la grammaire de référence scolaire, le *Tatabahasa Dewan*. Il est encore très fréquent aujourd'hui de voir dans des lettres officielles une utilisation erronée du « di » : le préfixe est séparé d'un blanc de sa base, la préposition est jointe au mot qui le suit. Nous n'entrerons pas trop en détail dans l'utilisation des ponctuations, principalement le tiret et l'apostrophe : absence du tiret dans une forme réduplicuée (**rumahrumah* au lieu de *rumah-rumah* 'maisons') ou dans l'affixation de mots empruntés (**mengupgradekan* au lieu de *meng-upgrade-kan* 'mettre à jour'), maintien de l'apostrophe dans des mots d'origine arabe (**Juma'at* au lieu de *Jumaat* 'vendredi'), oubli de l'apostrophe dans les formes tronquées (**kan* au lieu de *'kan*, forme réduite de *akan* 'marque du futur'), etc.

3 Constitution d'un corpus écrit

Tous les problèmes évoqués précédemment nous ont amenée à orienter nos recherches sur la classification et l'étiquetage morphosyntaxique des mots malais vers une approche empirique permettant de pallier l'absence de référence linguistique théorique et de norme.

Il y a quelques années, un projet mené à l'UTMK en collaboration avec le DBP a abouti à la création d'un corpus écrit comprenant des textes littéraires et académiques. Une grande partie de ce corpus est malheureusement égarée ou inutilisable (problème de plateforme). Nous avons regroupé les textes utilisables, les avons mis pour le moment sous format de fichier texte et fait corriger manuellement. A cet ensemble de textes littéraires, nous avons ajouté des textes journalistiques provenant d'un journal national malaisien (*Bernama*). Le tableau suivant (Figure 1) montre la composition et la taille du corpus.

Nombre de fichiers	Genre	Année de publication	Nombre d'occurrences de mots
1904	Textes journalistiques	2003	551974
11	Ouvrages littéraires	1975-1992	950398
9	Publications académiques	1986-1999	586145
2	Manuels d'utilisation	1995-2000	166486
		TOTAL =	2255003

Figure 1 : Description du corpus

Nous continuons de télécharger sur Internet des articles de journaux (*Utusan Malaysia*[5] et *Berita Harian*[6]), des textes officiels et des articles publiés par des universitaires pour pouvoir tester nos hypothèses et mettre à jour les entrées du dictionnaire (le malais ne cesse d'emprunter des termes anglais et de créer de nouvelles formes affixées). Même si la fonction 'téléchargement' est automatique (nous utilisons le logiciel *Net Transport*), le choix des types de textes que nous nous sommes imposé nécessite la lecture partielle du texte et surtout l'identification de la langue. Le malais et l'indonésien sont deux langues très proches. Les identificateurs de langues en ligne ne font pas toujours la différence entre les deux. Nous avons créé un identificateur de textes écrits malais et indonésiens qui sera disponible sur le site de l'UTMK avant la fin du mois de juin de cette année.

La correction des textes était planifiée en deux étapes : correction automatique et correction manuelle. Le test effectué avec le correcteur orthographique *DewanEja* s'est avéré non satisfaisant ce qui a ramené la correction à une tâche manuelle fastidieuse et longue. Il existe actuellement deux correcteurs orthographiques commerciaux, *EjaTepat* (Accredo Multimedia Sdn. Bhd.) et *DewanEja 3000* (The Name Technology Sdn. Bhd.). L'UTMK avait déjà créé vers la fin des années 80 un correcteur orthographique du malais, tournant sur Macintosh et Microsoft Windows sur un IBM PC. Un des projets en cours de l'UTMK est le rafraîchissement de ce correcteur orthographique trop obsolète et l'ajout d'un composant pour la correction grammaticale.

---

[5] Utusan Malaysia : http://www.utusan.com.my/.

[6] Berita Harian : http://www.bharian.com.my/

4 Etiquetage du corpus

Les entrées du dictionnaire *Kamus Dewan* (format papier et électronique) ne sont pas catégorisées. La présence des bases liées (bases simples qui n'apparaissent qu'avec un affixe ou seulement dans une composition) est une des raisons sans doute pour lesquelles ces entrées ne sont pas accompagnées de catégorie grammaticale. Le DBP est en train d'ajouter ces informations dans sa prochaine édition dont la date de publication reste incertaine. Certains dictionnaires, généralement bilingues (par exemple, *Kamus Dwibahasa Oxford Fajar*, *Collins Headstart Easy Learning English-Malay bilingual dictionary*), ont commencé à ajouter les catégories grammaticales mais comme ils ne sont pas issus du DBP, ils ne sont pas reconnus. Ils restent cependant une aide pour la classification des mots malais.

4.1 Premier jeu d'étiquettes

La catégorisation des mots d'une langue repose sur un ensemble d'étiquettes défini, non ambigu et si possible ni trop grand ni trop petit. Une petite recherche sur le Web en découragera plus d'un, vu le nombre de jeux d'étiquettes disponibles. Le problème est de donc de choisir le jeu d'étiquettes qui s'adapte le plus à la langue de recherche. La question reste ouverte car comment savoir lequel est le plus adapté ? Nous avons choisi de commencer avec les étiquettes « traditionnelles » présentées dans la grammaire du DBP puis de trouver une correspondance entre ces étiquettes et le jeu d'étiquettes proposé par exemple par EAGLES[7] et MULTEXT. À la fin, nous espérons proposer un jeu d'étiquettes du malais réutilisable et comparable à un jeu d'étiquettes standard.

Nous proposons un premier jeu d'étiquettes qui va permettre d'amorcer la catégorisation morphosyntaxique du corpus. L'ensemble proposé est une réorganisation des catégories proposées par le DBP. Les classes « adverbes » et « ponctuations » y ont été ajoutées.

1. Verbe (V) : Verbe transitif (VT) ; Verbe intransitif () ; Verbe actif (VA) ; Verbe passif (VP) ; Verbe transitif actif (VTA) ; Verbe transitif passif (VTP)
2. Adjectif (A)
3. Nom (N) : Nom propre (NP) ; Nom commun (NC)
4. Pronom (P)
5. Déterminant (D)
6. Adposition (AD)
7. Conjonction (C)

---

[7] EAGLES – Recommendations for the Morphosyntactic Annotation of Corpora. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

8. Interjection (I)
9. Adverbe (AV) : Négatif (AVN); Affirmatif (AVA); Interrogatif (ADI); Intensificateur (AVT)
10. Ponctuation (PO)

Nous illustrerons de manière simple les différentes étapes de notre approche à partir de trois phrases tirées du dictionnaire *Kamus Dewan*. Ces trois phrases nous servirons de fil rouge au travers de l'étiquetage des mots non ambigus, des mots affixés et à clitiques et de l'étiquetage par règles.

Exemple 1: *Buku ini baik dibaca oleh kanak-kanak.*

'Ce livre est bien pour être lu par les enfants'

Exemple 2: *Lebih baik engkau yakan sahaja akan kata-katanya.*

'Il vaut mieux pour toi que tu approuves ces remarques'

Exemple 3: *Kata-katanya itu terarah kepada lawannya.*

'Ces remarques sont dirigées contre ses opposants'

4.2 Étiquetage des mots non ambigus

Après avoir déterminé le premier jeu d'étiquettes du malais, la première étape dans l'étiquetage des mots du corpus consiste à étiqueter les mots appartenant à une seule catégorie donc non ambiguë. Cet ensemble de mots contient pour le moment les auxiliaires, les mots descriptifs, les pronoms, les déterminants, les conjonctions, les adpositions, les interjections, les adverbes et les numéraux. Au fur et à mesure que d'autres mots du corpus sont étiquetés, ils seront ajoutés à cet ensemble.

Dans les trois exemples cités précdemment, les mots non ambigus sont suivis de leur catégorie placée ici entre les deux symboles d'infériorité et superiorité.

Ex. 1: *Buku ini*<D> *baik dibaca oleh*<AD> *kanak-kanak*.<PO>

Ex. 2: *Lebih*<AVT> *baik engkau*<P> *yakan sahaja*<AV> *akan kata-katanya*.<PO>

Ex. 3: *Kata-katanya itu*<D> *terarah kepada*<AD> *lawannya*.<PO>

4.3 Étiquetage des mots affixés et des mots à clitiques

La liste des mots tirés du corpus a d'abord fait l'objet d'une correction manuelle afin d'éliminer les mots étrangers et d'isoler les abréviations et les noms propres. Le regroupement des mots par sous-chaînes partagées suivi d'une correction manuelle a permis

non seulement de mettre à jour toutes les formes dérivées d'une base, mais aussi de déterminer la propriété transitive d'un verbe actif préfixé par « me- ». Le dictionnaire *Kamus Dewan* ne cite pas comme sous-entrée la forme passive préfixée par « di- ». Si la famille morphologique d'une base contient à la fois les deux formes préfixées par « me- » et « di- », ces deux formes sont des verbes transitifs, le premier étant la forme active et le second la forme passive.

Les mots affixés malais peuvent être donc catégorisés par la reconnaissance des affixes qui les composent. Par exemple, les mots contenant le préfixe « me- » sont des verbes actifs et les mots contenant à la fois le préfixe « me- » et l'un des suffixes « -kan » ou « -i » sont des verbes actifs transitifs. La décomposition morphologique des mots est obtenue par l'analyseur de l'affixation du malais développé par Ranaivo [7]. Comme cet analyseur travaille sur des mots hors-contexte et que la liste des bases permettant la validation des découpages n'est accompagnée d'aucune information linguistique, certains mots affixés sont classés avec plusieurs étiquettes.

Dans les exemples 1 à 3, les mots affixés sont *dibaca*, *yakan* et *terarah*. Les deux premiers ont été reconnus par l'analyseur comme étant des verbes transitifs au passif (VTP). Le dernier a été analysé avec deux catégories possibles : verbe intransitif (VI) et adjectif (A).

Ex. 1 : *Buku ini*<D> *baik dibaca*<VTP> *oleh*<AD> *kanak-kanak.*<PO>

Ex. 2 : *Lebih*<AVT> *baik engkau*<P> *yakan*<VTP> *sahaja*<AV> *akan kata-katanya.*<PO>

Ex. 3 : *Kata-katanya itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya.*<PO>

La classification de certaines bases simples peut se faire par l'identification de certains clitiques. Par exemple, les mots servant de support aux clitiques « -ku » et « -mu » sont soit une de ces prépositions, *bagi*, *kepada*, *pada*, *oleh*, *untuk*, soit des noms.

- *Tidak terpikul olehku benda yang berat itu.*

  Négation – pouvoir porter – par moi – objet – Relatif – lourd – Déterminant

  Je n'ai pas pu porté cet objet lourd.
- *Adikku ialah seorang yang baik.*

  Frère/Sœur moi – est – une personne – Relatif – bien

  Mon frère / Ma sœur est une personne bien.

4.4 Étiquetage par règles

L'étiquetage du reste des mots du corpus peut s'effectuer soit en créant un étiqueteur soit en adaptant un étiqueteur existant. Cette deuxième solution n'est pas vraiment facile à réaliser. Les étiqueteurs utilisant un dictionnaire et des grammaires ne sont pas utilisables car le malais ne possède pas encore ces données linguistiques. Les étiqueteurs demandant un corpus étiqueté préalablement et manuellement ne sont pas non plus applicables car cela implique que le jeu d'étiquettes du malais soit déterminé alors que nous essayons de le construire. Un

autre obstacle à l'adaptation des outils existants est l'incompatibilité des plateformes et la difficulté d'utilisation car l'outil est écrit avec un langage de programmation inconnu du linguiste informaticien. Pour l'instant, la solution à ce problème n'est pas très claire car nous sommes encore dans la phase de test des étiqueteurs disponibles sur Internet.

Afin d'étiqueter le maximum de mots dans le corpus, nous avons établi quelques simples règles contextuelles.

- Règle 1 : Un mot précédant *ini* ou *itu* est un nom

Cette règle permet d'étiqueter les mots *buku* et *kata-katanya*.

Ex. 1 : *Buku*<N> *ini*<D> *baik dibaca*<VTP> *oleh*<AD> *kanak-kanak*.<PO>

Ex. 3 : *Kata-katanya*<N> *itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya*.<PO>

- Règle 2 : Un mot devant une adposition est un nom

Cette deuxième règle permet d'étiqueter les mots *kanak-kanak* et *lawannya*.

Ex. 1 : *Buku*<N> *ini*<D> *baik dibaca*<VTP> *oleh*<AD> *kanak-kanak*<N>.<PO>

Ex. 3 : *Kata-katanya*<N> *itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya*<N>.<PO>

- Règle 3 : Un mot devant un adverbe-intensificateur est un adjectif ou un verbe

Cette règle permet d'étiqueter le mot *baik*.

Ex. 2 : *Lebih*<AVT> *baik*<A ;V> *engkau*<P> *yakan*<VTP> *sahaja*<AV> *akan kata-katanya*.<PO>

Ces trois étapes (étiquetage des mots non ambigus, étiquetage des mots affixés et à clitique, étiquetage par règles) sont répétées jusqu'à ce que tous les mots du texte soient étiquetés ou que plus aucune règle ne soit applicable. Cette répétition du processus va permettre d'étiqueter les mot *baik* (exemple 1) et *kata-katanya* (exemple 2). Dans nos trois exemples, seul le mot *akan* (exemple 2) n'a pas pu être étiqueté et les deux mots, *baik* (exemples 1 et 2) et *terarah* (exemple 3), sont avec deux catégories.

Ex. 1 : *Buku*<N> *ini*<D> *baik*<A;V> *dibaca*<VTP> *oleh*<AD> *kanak-kanak*<N>.<PO>

Ex. 2 : *Lebih*<AVT> *baik*<A;V> *engkau*<P> *yakan*<VTP> *sahaja*<AV> *akan kata-katanya*<N>.<PO>

Ex. 3 : *Kata-katanya*<N> *itu*<D> *terarah*<VI;A> *kepada*<AD> *lawannya*<N>.<PO>

5 Conclusion

Nous avons présenté dans cette communication une méthode permettant la catégorisation des mots d'une langue, dans ce travail le malais. Cette méthode consiste tout d'abord à définir un jeu d'étiquettes morphosyntaxiques très grossier et établir une liste des mots dont la catégorie n'est pas ambiguë à partir de ces catégories prédéfinies. Puis, les mots affixés et à clitique sont étiquetés en utilisant un analyseur morphologique. Les mots restants sont étiquetés par l'application de quelques règles contextuelles. A la fin du traitement d'un texte, tous les mots étiquetés sont ajoutés à la liste des mots dont la catégorie n'est pas ambiguë. Ces différentes étapes sont répétées jusqu'à ce que tous les mots du corpus aient obtenus une étiquette ou que plus aucune règle ne soit applicable.

Les premiers résultats de nos travaux sont la création d'un corpus partiellement annoté avec des catégories morphosyntaxiques, un jeu d'étiquettes morphosyntaxiques du malais et une catégorisation morphosyntaxique des bases simples et affixées.

L'étape suivante sera de développer la méthode destinée à compléter l'étiquetage de tous les mots en utilisant un corpus partiellement étiqueté.

Remerciements

Je remercie mes deux relecteurs anonymes et Christian Boitet pour leurs commentaires constructifs.

Références

[1] *Kamus Dewan – Edisi Ketiga*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 1994.

[2] *General guidelines for Malay Spelling*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 1992.

[3] ISMAIL BIN DAHAMAN. *Pedoman Ejaan dan Sebutan Bahasa Melayu*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 2000.

[4] ISMAIL BIN DAHAMAN. *Pedoman Ejaan Rumi Bahasa Melayu*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 2000.

[5] *General guidelines for the formation of terms in Malay*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 1992.

[6] NIK SAFIAH KARIM, FARID M. ONN, HASHIM HJ. MUSA, ABDUL HAMID MAHMOOD. *Tatabahasa Dewan. Edisi Baharu*, Kuala Lumpur, Dewan Bahasa dan Pustaka. 2004.

[7] RANAIVO B. *Analyse automatique de l'affixation en malais*, Thèse de doctorat, Institut National des Langues et Civilisations Orientales, Paris. 2001.