Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens

Thi Minh Huyen Nguyen, Laurent Romary, Xuan Luong Vu

Une étude de cas pour l'étiquetage morpho-syntaxique de textes vietnamiens

Thi Minh Huyen Nguyen (1), Laurent Romary (1) et Xuan Luong Vu (2)

(1) LORIA
BP 239, 54506 Vandoeuvre lès Nancy
nguyen@loria.fr, romary@loria.fr
(2) Centre de Lexicographie du Vietnam
N° 67/4A, Ly Thuong Kiet Str., Hanoi, Vietnam
vuluong@vietlex.com

Résumé – Abstract

Dans cet article, nous discutons de la construction des jeux d'étiquettes pour l'analyse morpho-syntaxique du vietnamien, en prenant en compte les spécificités linguistiques de cette langue. Cette construction est inspirée du modèle MULTEXT(*) dans le but de s'orienter vers les applications multilingues ainsi que la réutilisabilité des jeux d'étiquettes. Nous allons finalement décrire une expérimentation sur l'étiquetage lexical des textes vietnamiens en utilisant QTAG (Mason et Tufis, 1998), un étiqueteur probabiliste indépendant des langues.

This paper discusses part of speech (POS) tagset construction for Vietnamese by considering linguistic specificities of this language. We take into account the schema as defined in the MULTEXT(*) model, so as to account for possible multilingual applications as well as the reusability of defined tagsets. Finally we describe our experiments on tagging Vietnamese texts using QTAG (Mason and Tufis, 1998), a language independent probabilistic tagger.

Mots Clés - Keywords

partie du discours, corpus de textes, étiquetage morpho-syntaxique, MULTEXT, normalisation, QTAG

MULTEXT, part-of-speech (POS), POS tagging, QTAG, standardization, text corpus

---

(*) Multilingual Text Tools and Corpora http://www.lpl.univ-aix.fr/projects/multext/

1 Introduction

Chaque mot d'une langue appartient potentiellement à une ou plusieurs parties du discours selon son contexte d'utilisation. L'étiquetage lexical consiste à attribuer une étiquette morpho-syntaxique pour chaque mot dans un texte. Cette tâche est essentielle pour tout traitement ultérieur comme l'analyse syntaxique, sémantique ou même pragmatique d'une langue.

La notion de mot dans une tâche d'étiquetage lexical ne correspond pas nécessairement à un mot traditionnel en raison de la segmentation aveugle des textes sans information syntaxique ou sémantique. Un mot traditionnel peut être divisé en plusieurs unités ou morphèmes (dans le cas d'amalgames ou de mots composés, par exemple). Au contraire, plusieurs mots en séquence peuvent être groupés en une seule unité : des locutions, des noms propres composés, des numéros composés, des mots composés, etc. En fonction de la définition des unités lexicales et/ou de l'application, les descriptions des classes et des étiquettes morpho-syntaxiques peuvent inclure un ou plusieurs traits comme la catégorie syntaxique, le lemme, le genre, le nombre, etc. Dans (Przepiórkowski et Woliński, 2003), les auteurs proposent une nouvelle classification purement morpho-syntaxique. Ils défendent l'idée que plusieurs jeux d'étiquettes polonais existants sont naïfs linguistiquement du fait de l'adoption directe, sans analyse critique préalable, des classes traditionnelles de parties du discours, ce qui cause un manque de réutilisabilité. (Tufis, 1998) a proposé un jeu d'étiquettes à deux couches dans le but de réduire les coûts de temps et de mémoire dans le processus d'étiquetage exploitant un jeu de plus de 700 étiquettes.

Il existe aujourd'hui différents outils pour l'étiquetage morpho-syntaxique, ainsi que d'immenses ressources de corpus annotés destinées à des traitements variés dans nombreuses langues. Les projets Treebank (http://www.cis.upenn.edu/~treebank/home.html) sont des exemples de création de larges corpus annotés. Cela suppose également l'existence de définitions variées d'unités lexicales et de jeu d'étiquettes selon l'objectif visé. Dans le cadre de Multext (Ide et Véronis, 1994) et de Multext-Est (Erjavec et al., 1996), des jeux d'étiquettes ont été définis pour une dizaine de langues avec un haut niveau de consensus au sujet de la structure de description.

Aussi se posent les questions cruciales du caractère réutilisable de ces ressources linguistiques pour un nombre croissant d'applications, leur réutilisation combinée dans un contexte multilingue, et l'adaptation d'un outil à d'autres langues. De multiples projets ont vu le jour dans cette perspective : l'évaluation des outils, la normalisation et la représentation des structures de description morpho-syntaxique (Ide et Romary, 2001).

Dans le cas des textes vietnamiens, le travail d'étiquetage est une tâche nouvelle et difficile pour les informaticiens, essentiellement du fait du désaccord sur la classification linguistique traditionnelle des mots au sein de la communauté linguistique. A ce jour il n'existe aucun standard reconnu pour les catégories des mots en vietnamien. Notre recherche vise deux objectifs principaux : en premier lieu créer des outils et des ressources linguistiques pour les applications de traitement automatique des textes vietnamiens, mais aussi assurer la disponibilité de ces outils pour les linguistes travaillant sur le vietnamien.

Après un bref état de l'art dans le domaine de l'étiquetage, nous présentons les spécificités linguistiques importantes du vietnamien dans le but de définir un jeu d'étiquettes. Pour la construction de jeu d'étiquettes, nous nous sommes volontairement basés sur le modèle

Multext intrinsèquement dédié aux applications multilingues. Ce jeu d'étiquettes sera évalué grâce à l'étiqueteur QTAG (Mason et Tufis, 1998).

2 Travaux antérieurs d'étiquetage lexical

2.1 Méthodologie et évaluation

Un état de l'art complet de l'étiquetage morpho-syntaxique est présenté dans (Paroubek et Rajman, 2000). L'étiquetage lexical s'effectue usuellement en trois étapes : segmentation du texte en unités lexicales, accès à un lexique pour récupérer toutes les étiquettes possibles pour chaque mot, et désambiguïsation pour attribuer l'étiquette correcte à chaque mot. Il existe deux approches principales pour la tâche de désambiguïsation : les méthodes à base de règles et les méthodes probabilistes.

Les méthodes à base de règles exploitent un ensemble de règles grammaticales pour résoudre le problème de l'étiquetage. Les méthodes non supervisées utilisent des contraintes produites par les linguistes et un lexique contenant pour chaque mot ses étiquettes possibles. Un tel étiqueteur s'apparente à un parseur (par ex. les systèmes GREYC et SYLEX présentés dans l'évaluation GRACE - Grammaires et Ressources pour les Analyseurs de Corpus et leur Evaluation). Les méthodes supervisées construisent les étiquettes et les règles de transformation à partir de corpus étiquetés manuellement. L'étiqueteur de Brill est l'exemple le plus connu de telles méthodes. A chaque mot est ensuite attribuée l'étiquette la plus fréquente selon le lexique. Enfin, les règles de transformation servent à la correction itérative de l'étiquetage préalable.

Les méthodes probabilistes (dont les systèmes utilisant le modèle de Markov caché) utilisent la distribution de probabilité sur l'espace des associations possibles entre les séquences de mots et les séquences d'étiquettes. Cette distribution est produite à partir du corpus d'apprentissage étiqueté ou non. La désambiguïsation entre les étiquettes d'un mot s'opère par le choix de la séquence d'étiquettes qui maximise la probabilité conditionnelle de l'association avec la séquence de mots courante. Ces méthodes reposent sur deux hypothèses. La probabilité d'association entre un mot et une étiquette est entièrement conditionnée par la connaissance de l'étiquette. Ensuite la probabilité d'occurrence d'une étiquette est conditionnée par la connaissance d'un nombre fixe d'étiquettes voisines.

La performance des systèmes d'étiquetage se mesure le plus souvent par le taux de précision (au niveau des mots) qui dépend fortement de la nature et de la taille du jeu d'étiquettes. La plupart de ces systèmes ont une performance supérieure à 90%. Le meilleur résutat obtenu dans l'évaluation du projet GRACE était de 97.8%.

2.2 Aspect de normalisation dans le domaine de l'étiquetage lexical

De gros efforts ont porté sur la normalisation des données, des outils et des ressources linguistiques pour favoriser leur réutilisabilité pour les recherches et les applications en traitement des langues à base de corpus. Multext en est un example significatif. Dans le cadre de ce projet, un modèle morpho-syntaxique a été développé en vue de l'harmonisation de

l'étiquetage de corpus multilingue ainsi que de la comparabilité des corpus étiquetés. Multext défend l'idée que dans un contexte multilingue, des phénomènes identiques devraient être encodés de manière similaire dans chaque langue pour faciliter les traitements dans des applications diverses (alignement automatique, extraction de terminologie multilingue, etc.).

Un principe du modèle est de séparer les descriptions lexicales qui sont en générale stables, et les étiquettes du corpus. En ce qui concerne les descriptions lexicales, le modèle utilise deux couches : un noyau commun pour des catégories communes et une couche privée contenant des informations additionelles qui sont propres à une langue ou aux applications particulières. Une solution de compromis pour les jeux d'étiquettes morpho-syntaxiques dans le noyau commun est un jeu de 11 étiquettes : Nom (N), Verbe (V), Adjectif (A), Pronom (P), Déterminant (D), Adverbe (R), Adposition (S), Conjonction (C), Numéral (M), Interjection (I), Résidu (X). L'information optionnelle de la deuxième couche est présentée par les paires attribut-valeur (structure de traits). Par exemple, un nom commun singulier est présenté par N[type = common gender = masculine number=singular case=n/a] (forme contracté Ncms-).

Or, il est évident que pour couvrir une plus grande variété de langues, il est nécessaire de présenter plus de flexibilité dans ce cadre fondamental. L'étude que nous présentons sur le jeu d'étiquettes vietnamien prouve qu'en effet quelques catégories peuvent ne pas convenir aux objets linguistiques réels de cette langue. Du point de vue de la normalisation, ceci signifie qu'une étape ultérieure serait soit de décrire une ontologie entière des catégories (comme suggéré par Farrar et al., 2002), soit d'enregistrer la variété de descripteurs possibles à travers des langues en construisant un enregistrement de méta-données (cf. Ide et Romary, 2001). Ces deux options ne sont pas nécessairement contradictoires puisque les catégories de données élémentaires peuvent se diriger aux noeuds dans l'ontologie, laissant comparer des jeux d'étiquettes à travers des langues ainsi que dans une langue donnée. De ce fait, il est important de considérer que pour une langue comme le vietnamien, un schéma d'annotation peut se fonder sur plusieurs couches de granularité d'étiquettes, et ceci devrait être pris en compte. La section suivante présente une telle stratégie, qui pourrait mener en particulier à une proposition d'un ensemble de références de descripteurs pour le vietnamien, dans le contexte du comité d'ISO TC37/SC4 (http://www.tc37sc4.org).

3 Définition d'un jeu d'étiquettes pour le vietnamien

Le vietnamien est une langue isolante, dans laquelle chaque mot a une forme unique et ne peut pas être modifié par dérivation ou flexion. Les relations grammaticales ne se manifestent pas par la flexion mais par l'ordre des mots. La classification de parties du discours n'est pas donc morphologiquement évidente.

3.1 Lexique

La langue vietnamienne a une unité spéciale appelée "tiếng" qui correspond en même temps à une syllabe du point de vue phonologique, à un morphème du point de vue syntaxique, à un sémantème du point de vue de la structure du mot, et à un mot du point de vue des constituants de la phrase. Il y a trois types de "tiếng" :

1. "tiếng" ayant un sens réel comme *sông* (rivière), *núi* (montagne), *đi* (aller), *đứng* (tenir debout), *nhớ* (se souvenir), *thương* (aimer/avoir pitié), ..., peut constituer à lui seul un constituant de phrase complet du point de vue syntaxique et sémantique. Ce type de mot est appelé **mot lexical**.

2. "tiếng" comme *nhưng* (mais), *mà* (que), *tuy* (bien que), *nên* (alors) ..., qui ne peut pas être un constituant de phrase à lui seul, mais qui est utilisé pour composer un constituant de phrase lexical, est appelé **mot outil**.

3. "tiếng" qui vient du chinois comme *sơn* (montagne), *thuỷ* (eau), *gia* (maison), *bất* (négation) ... ou qui a un sens flou et qui est en général combiné avec une autre syllabe comme *cộ* (*xe cộ* - véhicule), *đẽ* (*đẹp đẽ* - beau), *vẻ* (*vui vẻ* - joyeux)... permet de créer des mots et peut parfois utilisé comme un mot.

Parmi les diverses définitions du mot en vietnamien, les linguistes sont parvenu à un accord et considèrent comme un mot la plus petite unité ayant un sens spécifié et une structure stable, et utilisée pour composer des constituants de phrase. Le lexique vientamien contient : *(i)* Des ***mots simples*** ou de mots monosyllabiques correspondant aux catégories 1 et 2 de "tiếng". *(ii)* Des ***mots complexes*** qui ont plus d'une syllabe. Il existe principalement trois types de combinaison des syllabes : redoublement phonétique (par ex. *trắng*/blanc - *trăng trắng*/blanchâtre), coordination sémantique (par ex. *quần*/pantalon, *áo*/chemise - *quần áo*/vêtement) et composition sémantique (par ex. *xe*/véhicule, *đạp*/pédaler - *xe đạp*/bicyclette). On note aussi l'existence de mots composés dont les syllabes ne sont plus reconnaissables (*bồ nông*/pélican). *(iii)* Enfin, ***des expressions figées et des locutions***, qui sont généralement considérées comme des unités lexicales.

A cause de la grande fréquence des mots composés, la segmentation des textes en vietnamien est assez compliquée.

3.2 Jeu d'étiquettes

Le problème de classification des catégories grammaticales en vietnamien est toujours en débat au sein de la comunauté linguistique (Hữu Đạt et al., 1998 ; Diệp et Hoàng, 1999 ; Cao, 2000). La difficulté vient de l'ambiguïté des rôles grammaticaux de nombreux mots. La mutation catégorielle verbe-nom est bien fréquente (sans aucune variation morphologique). Géneralement les déterminants peuvent être utilisés comme des noms. Même les prépositions (par ex. *trên*/sur, *trong*/dans) jouent parfois le rôle de noms (*trên*/le supérieur, *trong*/l'intérieur), etc. Dans cette section nous présentons notre approche pour définir un jeu d'étiquettes conforme aux applications de Traitement Automatique des Langues (TAL).

Les catégories grammaticales reflètent des oppositions diverses dans le système syntaxique. Le principal critère pour la définition de notre jeu d'étiquettes est donc la distribution syntaxique. Nous devrions avoir un important jeu d'étiquettes pour refléter exactement toutes les relations syntaxiques. Cependant, plus le jeu d'étiquettes est important, plus la tâche d'annotation est difficile. Aussi, nous avons besoin d'un compromis pour parvenir à un jeu d'étiquettes assez précis et de taille acceptable. Nous commençons avec un petit jeu d'étiquettes généralement admis dans la littérature (cf. Uỷ ban KHXH, 1983), et figurant dans différents dictionnaires vietnamiens. Ce jeu comporte neuf catégories: Nom (N), Verbe (V), Adjectif (A), Pronom (P),

Adjonction (J), Conjonction (C), Interjection (I), Mot de Modalité (E) et Résidu (X). Notre tâche est alors de définir un nouveau jeu d'étiquettes en subdivisant chacune de ces catégories à l'aide d'étiquettes plus spécifiques.

En nous inspirant des principes de construction du modèle Multext, nous avons élaboré des spécifications lexicales pour le vietnamien dans un schéma comparable à ce modèle. Les différences du jeu d'étiquettes ci-dessus avec celui de Multext sont les suivantes : les numéraux et les déterminants de Multext se retrouvent dans la catégorie des noms du vietnamien, les adpositions de Multext se retrouvent dans la catégorie des conjonctions du vietnamien ; la catégorie des adjonctions du vietnamien contient des adverbes et des adjonctions de noms (noms pluralisant) de Multext ; la classe des modaux est propre au vietnamien. Pour rester conforme à Multext, nous avons ajouté la catégorie des numéraux à ce jeu d'étiquettes. Cette classe comprend les cardinaux et les ordinaux de la classe des noms et les adjonctions de noms de la classe des adjonctions. Par conséquent, seuls les adverbes restent dans la classe des adjonctions. Nous n'avons pas récupéré les classes Déterminant et Adposition du modèle Multext à cause de particularité de la grammaire vietnamienne. En définitive, nous obtenons le jeu d'étiquette de premier niveau suivant : Nom (N), Verbe (V), Adjectif (A), Pronom (P), Adverbe (J), Conjonction (C), Numéral (S), Interjection (I), Particule Modale (M), Résidu (X). Ci-dessous, nous présentons des spécifications lexicales de base pour chaque catégorie, fondées sur les combinaisons lexicales possibles.

- **Nom** : Seul l'attribut **type** (commun ou propre) dans Multext est approprié pour le vietnamien. Par contre, nous définissons de nouveaux attributs dont les valeurs sont entre crochets : **collective** [yes (*cây cối*/végétation), no (*cây*/plante)**]**, **sense** [object (*nhà*/maison), plant (*lúa*/riz), animal (*mèo*/chat), human (*học sinh*/élève), material (*sắt*/fer), abstract (*tình cảm*/sentiment), fact (*sự*/fait), space (*trong*/intérieur), time (*ngày*/jour), senses (*màu*/couleur), style (*giáo sư*/professeur)], **countable** [absolute (*cái*/chose[1]), partial (*bàn*/table), no (*nhân dân*/peuple)], et **unit** [classifier (*cái*/le-un), collective (*bộ*/ensemble), exact measurement (*lít*/litre), rough measurement (*nắm*/poignée)].

- **Verbe** : Comme les verbes du vietnamien ne sont pas flexionnels, aucun attribut défini dans Multext n'est approprié ici. Nous avons donc créé de nouveaux attributs propres au vietnamien : **transitive** [yes (*viết*/écrire), no (*ngủ*/dormir)], **sense** [psychology (*tin*/croire), discourse (*nói*/dire), direction (*lên*/s'elever), movement (*chạy*/courir), existence (*mất*/perdre), transformation (*trở thành*/devenir), volition (*muốn*/vouloir), acceptation (*bị*/subir), comparison (*bằng*/égaler), residual (*viết*/écrire)]. De plus, il existe un verbe spécial "*là*/être", qui est son étiquette lui-même.

- **Adjectif** : Le seul attribut pour les adjectifs du vietnamien est **type** [quality, quantity] (par ex. *đẹp*/jolie, *cao*/haut).

- **Pronom** : L'attribut principal intéressant de cette catégorie est son **type**, car il n'y a pas de cas, de genre ou de nombre dans la grammaire vietnamienne. L'ensemble de

[1] "cái" est un nom classificateur. Par exemple : *cái*/chose *bàn*/table = la table, *một*/un *cái*/chose *bàn*/table = une table, *cái*/chose *này*/cette = cette chose

valeurs appropriées à cet attribut est : personal (par ex. *tôi*/je, *chị*/vous-soeur), temporal (par ex. *bây giờ*/maintenant), demonstrative (par ex. *đây*/ici, *này*/ce), quantitative (par ex. *tất cả*/tout, *bấy nhiêu*/autant), predicative (par ex. *thế*/cela), et interrogative (par ex. *ai*/qui, *gì*/quoi).

- **Adverbe** : Les adverbes du vietnamien sont des mots outils très importants pour exprimer le temps, changer le degrée du prédicat, etc. d'une phrase. Nous définissons un nouvel ensemble de valeur pour l'attribut **type** : time (par ex. *đã*/temps passé, *sẽ*/temps futur), degree (par ex. *rất*/très, *quá*/trop), continuation/similarity (par ex. *cũng*/aussi, *vẫn*/toujours), negation (par ex. *không*/ne pas) et imperative (par ex. *hãy*/particule impératif, *đừng*/ne pas faire). De plus, un autre attribut est ajouté : **position** [pre, post] (par ex. *đã*/déjà, *rồi*/déjà).

- **Conjonction** : Cette catégorie emploie l'attribut **type** avec deux valeurs : subordinating (par ex. *của*/de, *do*/à cause de, *để*/pour) et coordinating (par ex. *và*/et, *nhưng*/mais, *nếu* ... *thì*/si ... alors).

- **Numéral** : Les valeurs de l'attribut **type** de cette classe sont : cardinal (*một*/un), ordinal (*nhất*/premier), adjunct (*những*/pluralisant).

- **Interjection**: Aucun attribut n'est associé à cette classe.

- **Mot de modalité** : Nous distinguons deux types de mots dans cette catégorie : particle correspondant aux mots ajoutés à une phrase afin de changer son intensité, et copulative correspondant aux mots ajoutés au début ou à la fin d'une phrase afin d'exprimer le sentiment de l'orateur.

- **Résidu** : Ce sont les unités lexicales et les expressions qui n'ont pas de classification spécifique.

D'autres traits pourront être ajoutés pour des objectifs différents (information sur la forme composée ou sur la forme redoublée, etc.). Pour ces spécifications lexicales, nous assignons 48 étiquettes à un jeu de deuxième niveau. Dans la section suivante, nous présentons un étiqueteur stochastique de textes vietnamiens avec deux jeux d'étiquettes définis.

4 Processus de l'étiquetage

Aucune recherche au sujet de l'étiquetage de partie du discours vietnamien n'a été publiée à ce jour. Nous avons démarré le travail en construisant un lexique vietnamien, dans lequel on associe chaque mot à ses étiquettes possibles dans les jeux d'étiquettes mentionnés précédemment. Comme nous l'avons discuté (section 3.2), le jeu d'étiquettes du deuxième niveau que nous avons choisi est un compromis afin d'éviter un jeu d'étiquettes trop grand. Pour valider ce choix, nous appliquons ce jeu d'étiquettes sur des corpus dont nous vérifierons à terme la distribution syntaxique. Un outil pour étiqueter automatiquement un corpus avec un jeu d'étiquettes donné est indispensable. Nous nous servons de l'étiqueteur QTAG à cette fin.

QTAG est un étiqueteur stochastique indépendant des langues. Il crée le lexique, le jeu d'étiquettes, les probabilités lexicales et contextuelles à partir du corpus manuellement étiqueté.

Grâce à cette base d'apprentissage, l'étiqueteur peut trouver les étiquettes possibles avec leur fréquence pour les assigner à chaque unité lexicale dans un nouveau corpus déjà segmenté. Si la recherche d'une unité dans le lexique échoue, l'étiqueteur essaie de lui trouver les étiquettes possibles par sa forme morphologique. Au pire des cas, cette unité se voit attribuer toutes les étiquettes existantes. Enfin, l'étiqueteur effectue la tâche de désambiguïsation en utilisant les distributions probabilistes apprises à partir du corpus.

On supprime le prédicteur morphologique de QTAG, puisque le vietnamien est une langue sans variation morpho-syntaxique. Nous nous concentrons maintenant sur la construction du lexique et du corpus d'apprentissage et puis évaluons les résultats obtenus.

4.1 Ressources langagières et corpus d'entraînement

En nous appuyant sur le Dictionnaire Vietnamien (Hoang Phe, 2002), nous construisons un lexique de 37454 unités lexicales, dont chaque unité a ses propres étiquettes. Ce lexique inclut des termes usuels du lexique de la vie quotidienne et des journaux, des termes fréquents en littérature, des termes dialectaux fréquemment utilisés, des termes scientifiques ou techniques dans les documents scientifiques populaires, des expressions usuelles, des syllabes spéciales seulement utilisées pour la composition des mots, et des abréviations d'usage courant. Le lexique est graduellement enrichi avec de nouveaux mots apparus dans les corpus traités.

Avant l'étiquetage proprement dit (manuel ou automatique), le premier pas est la segmentation, i.e. l'identification des unités lexicales dont la définition est donnée dans la section 3.1. Le vietnamien est monosyllabique, mais les mots composés sont fréquents. Cela ne permet pas une simple segmentation par les espaces dans un texte. Pour résoudre ce problème, nous avons adopté les automates d'états finis pour identifier des segmentations possibles pour chaque phrase (délimitée par des ponctuations). En pratique, la segmentation correcte la plus probable est le chemin le plus court dans le graphe. Dans le cas ambigü (plusieurs chemins de la même longueur), une intervention humaine est nécessaire. Cette solution simple s'avère efficace dans la plupart des cas. Quelques améliorations de cette méthode pourraient être faites dans le futur proche (par ex. l'identification des formes redoublées, la désambiguïsation utilisant l'information de partie du discours, etc.). Une autre approche de segmentation est présentée dans (Dinh Dien et al., 2001).

Ensuite, le corpus segmenté destiné à l'apprentissage de l'étiqueteur est manuellement annoté après le passage d'un outil d'étiquetage préalable. En vue de l'expérimentation, nous avons annoté un corpus de 74753 unités dont 63733 unités lexicales (à peu près de 10000 unités lexicales différentes, sans compter des ponctuations). Un cinquième de ce corpus est composé de textes journaux, le reste, de textes littéraires.

4.2 Evaluation

Notre étiqueteur modifié prend en compte le lexique construit (section 4.1). Nous avons entrepris 6 essais sur deux jeux d'étiquettes définis avec une taille croissante du corpus d'entraînement. Le texte restant dans le corpus manuellement étiqueté est employé pour le but d'évaluation.

Voici un exemple du résultat de l'étiquetage pour la phrase "**hồi / lên / sáu / , / có / lần / tôi / đã / nhìn / thấy / một / bức / tranh / tuyệt / đẹp**" qui est traduite mot à mot en ***"quand / monter / six / , / avoir / fois / je / déjà / regarder / voir / un / [classificateur] / image / extrême / beau"*** (Lorsque j'avais six ans, j'ai vu, une fois, une magnifique image) :

<w pos="Nt"> **hồi**</w> <w pos="Vto"> **lên** </w> < w pos="Sc"> **sáu** </w>
<w pos=","> **,** </w> <w pos="Vte"> **có** </w> <w pos="Nt"> **lần** </w>
<w pos="Pp"> **tôi** </w> <w pos="Jt"> **đã** </w> <w pos="Vtx"> **nhìn** </w>
<w pos="Vtx"> **thấy** </w> <w pos="Sc"> **một** </w>
<w pos="Nc"> **bức** </w> <w pos="No"> **tranh** </w>
<w pos="Jd"> **tuyệt** </w> <w pos="Aa"> **đẹp** </w>

dans lequel : Nt - nom temporaire, Vto - verbe transitif de direction, Sc - nombre cardinal, Pp - pronom personnel, Jt - adverbe temporaire, Vtx - verbe transitif (résidu), Nc - nom de classificateur, No - nom d'objet, Jd - adverbe de degrée, Aa - adjectif de qualité.

L'expérimentation confirme que, plus le corpus d'entraînement est volumineux, plus le résultat est précis. Le meilleur taux de précision pour le jeu d'étiquettes du premier niveau est d'environ 94% (9 étiquettes lexicales et 10 ponctuations) avec un corpus d'entraînement d'environ 50000 unités lexicales (60000 unités au total). Pour celui du deuxième niveau, le meilleur taux de précision est d'environ 85% (48 étiquettes lexicales et 10 ponctuations) avec le même corpus d'entraînement. Sans utiliser le lexique construit ci-dessus, ces taux de précision sont à peu près de 80% et 60% respectivement. Une petite partie d'erreurs est due aux erreurs d'étiquetage dans les données d'entraînement. Bien que le résultat soit plutôt modeste en apparence, particulièrement au deuxième niveau, il n'est pas décourageant car la taille du corpus d'entraînement est encore très petite (50000 unités en comparaison aux centaines de milliers d'unités des corpus d'entraînement dans d'autres travaux sur l'étiquetage).

L'étiqueteur est disponible sur la page http://www.loria.fr/equipes/led/outils.php (vnQTAG) ainsi que les ressources linguistiques (lexique, corpus d'apprentissage).

5 Conclusions

Nous avons présenté notre travail sur l'étiquetage lexical du vietnamien. Puisque les chercheurs vietnamiens se sont très récemment impliqués dans le domaine de TAL, nous avons dû construire toutes les ressources linguistiques nécessaires et définir toutes les structures de données à partir de zéro. Néanmoins, nous tirons bénéfice de quelques avantages : beaucoup de méthodologies existantes pour l'annotation morpho-syntaxique et une forte conscience de la tendance de normalisation. Le jeu d'étiquettes défini pourrait être facilement ré-ajusté et étendu grâce à des descriptions lexicales. Ces descriptions sont en plus comparables à celles d'autres langues prises en compte dans le cadre du projet Multext. Avec l'aide de l'étiquetage automatique, nous pouvons facilement augmenter la taille du corpus annoté. Les résultats obtenus constituent une base pour d'autres recherches dans le domaine de TAL pour le vietnamien : analyse syntaxique, recherche d'information, alignement multilingue, traduction automatique, etc.

Références

Uỷ ban Khoa học Xã hội Việt Nam (1983), *Ngữ pháp tiếng Việt (Vietnamese Grammar)*, Hanoi, NXB Khoa học Xã hội.

Ide N., Véronis J. (1994). MULTEXT: Multilingual Text Tools and Corpora. *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94,* Kyoto, Japan, 588-92.

Erjavec T., Ide N., Petkevic V., Véronis J. (1996) Multext-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages. *Proceedings of the First TELRI European Seminar,*87-98

Hữu Đạt, Trần Trí Dõi, Đào Thanh Lan (1998), *Cơ sở tiếng Việt (Basis of Vietnamese)*, Hanoi, NXB Giáo dục.

Mason O., Tufis D. (1998), Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger, Proceddings of *First International Conference on Language Resources and Evaluation (LREC),* Granada (Spain), 28-30 May 1998, p.589-596.

Tufis D. (1998), Tiered Tagging, in *International Journal on Information Science and Technology*, vol. 1, no. 2, Editura Academiei, Bucharest, 1998.

Diệp Quang Ban, Hoàng Văn Thung (1999), *Ngữ pháp tiếng Việt (Vietnamese Grammar, vol. 1-2)*, Hanoi, NXB Giáo dục.

Cao Xuân Hạo (2000), *Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa (Vietnamese - Some Questions on Phonetics, Syntax and Semantics)*, Hanoi, NXB Giáo dục.

Paroubek P., Rajman M. (2000), Etiquetage morpho-syntaxique, *Ingénierie des langues* (p. 131-150) Paris, HERMES Science Europe.

Dinh Dien, Hoang Kiem, Nguyen Van Toan (2001), Vietnamese Word Segmentation, Proceedings of *the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001)*, Tokyo (Japan), 27-30 November 2001, p. 749-756.

Ide N., Romary L. (2001), Standards for Language Resources, *Proceedings of the IRCS Workshop on Linguistic Databases,* Philapdelphia, 141-9.

Farrar, S., W. D. Lewis, D. T. Langendoen (2002), An Ontology for Linguistic Annotation, *AAAI '02 Workshop: Semantic Web Meets Language Resources*.

Hoàng Phê (2002), *Từ điển tiếng Việt (Vietnamese Dictionary)*, Vietnam Lexicography Centre, NXB Đà Nẵng.

Przepiórkowski A., Woliński M. (2003, to appear), The Unbearable Lightness of Tagging* A Case Study in Morphosyntactic Tagging of Polish, Proceedings of *4th International Workshop on Linguistically Interpreted Corpora (LINC-03*), Budapest (Hungary), 13-14 April 2003.