# Procédure de reconnaissance des mots et des phrases thaï

Krit KOSAWAT

**Abstract:** This work proposes a solution for using INTEX with the Thai alphabet, a linguistic system quite different from the various Latin alphabets. The main problems arise first from the need to recognize words in texts, a trivial question in European languages. Then the problem of sentence boundaries will be presented.

Based on an alphabetical system, Thai has its own characters, of which some are written above or below the others. The problems of word recognition are mainly due to the fact that there is no space between words. The full stop at the end of sentences is also missing. Besides, there is no inflexion: the category of a word is determined by its position in the sentence. Such are the principal causes of ambiguities and errors in lexical interpretation while analyzing a document by a computer system that cannot understand the sense of vocabulary.

**Keywords:** Word Recognition, Sentence Recognition, Thai, INTEX, Electronic Dictionary.

**Mots clés :** Reconnaissance de mot, reconnaissance de phrase, Thaï, INTEX, dictionnaire électronique.

## 1. Introduction

Le thaï (siamois) est la langue officielle du royaume de Thaïlande. Il appartient à la famille du taï, un sous-groupe de la famille du kadaï (ou kam-taï). Un certain nombre de linguistes considèrent maintenant le kam-taï, avec l'austronésien, comme une branche de l'austro-taï, bien que cette hypothèse demeure controversée (HUDAK 1987, p. 757).

✉ Krit KOSAWAT, LADL (Laboratoire d'Automatique Documentaire et Linguistique)
Université de Marne-la-Vallée, Institut Gaspard-Monge,
5, Bd Descartes, Champs-sur-Marne, F-77454 Marne-la-Vallée Cedex 2
e-mail : kosawat@univ-mlv.fr

## 2. Système d'écriture du thaï

Basé sur un système alphabétique, le thaï possède ses propres caractères, dont certains s'écrivent au-dessus ou au-dessous des autres. Il s'écrit horizontalement de gauche à droite, et principalement, sans espace entre les mots.

### 2.1. Alphabet thaï

L'alphabet thaï ne connaît pas de différence entre majuscule et minuscule. Les caractères sont classés en 5 catégories, à savoir :

2.1.1. *Consonnes*

Il existe 44 graphèmes consonantiques dont 2 sont obsolètes[1].

ก ข (ฃ) ค (ฅ) ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ
ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ
ส ห ฬ อ ฮ

2.1.2. *Voyelles*

La langue thaï est caractérisée par un très grand nombre de voyelles. Il y a 19 graphèmes qui peuvent se composer entre eux pour former 32 phonèmes vocaliques différents.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Graphèmes[2] | -ะ | -ั | -า | -ำ | -ิ | -ี | -ึ | -ื |
| | -ุ | -ู | เ- | แ- | โ- | ใ- | ไ- | |
| | ฤ | ฤๅ | (ฦ) | (ฦๅ) | | | | |
| Phonèmes | -ะ | -า | -ิ | -ี | -ึ | -ื | -ุ | -ู |
| | เ-ะ | เ- | แ-ะ | แ- | โ-ะ | โ- | เ -าะ | |
| | -อ | เ -อะ | เ -อ | เ - ียะ | เ - ีย | | | |

[1] Les caractères entre parenthèses sont obsolètes.

[2] Le trait d'union représente la position de consonne.

เ - ือะ เ - ือ -ัวะ -ัว ฤ
ฤๅ (ฦ) (ฦๅ) -ำ ใ- ไ- เ -า

2.1.3. *Marques de ton*

Les tons sont une des caractéristiques de la langue thaï. Il s'agit de traits distinctifs affectant des phonèmes. Les différents tons associés à un même mot peuvent correspondre à des sens différents. En thaï, il existe 5 tons dont 4 graphèmes à savoir :

| | |
|---|---|
| (aucun graphème) | ton moyen |
| -่ | ton bas |
| -้ | ton descendant |
| -๊ | ton haut |
| -๋ | ton montant |

2.1.4. *Chiffres*

Les caractères de chiffres thaï sont utilisés moins fréquemment que les chiffres arabes dans la vie quotidienne, à l'exception des documents officiels. Ils vont de zéro jusqu'à neuf :

๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙

2.1.5. *Caractères spéciaux*

| | |
|---|---|
| -์ | signe muet |
| -ๆ | signe de phonème court |
| ๆ | signe répétitif |
| ฯ | signe abréviatif |

## 2.2. Syllabe

Une syllabe est un ensemble composé de consonnes, de voyelles et d'un ton (SIRIBOONMA 1999, p. 172). Celle-ci se prononce par une seule émission de voix. En analysant des syllabes thaï, on constate que la séquence ci-dessous est bien respectée. Il est à noter que les caractères entre crochets sont facultatifs.

[Vi]Ci[Vd][T][Vt][Ct]

Ci : Consonne initiale
Ct : Consonne terminale
Vi : Voyelle initiale
Vd : Voyelle diacritique
Vt : Voyelle terminale
T : Marque de ton

2.2.1. *Consonne initiale*

C'est un élément de base exigé pour construire une syllabe en thaï. La consonne initiale se place vers le début de syllabe. Toutes les consonnes thaï peuvent être utilisées comme consonne initiale. Certaines consonnes peuvent se combiner avec une autre pour former une double consonne initiale.

2.2.2. *Consonne terminale*

C'est la consonne qui termine la syllabe. Elle est facultative. Toutes les consonnes thaï peuvent s'utiliser en position terminale à l'exception des 9 suivantes :

ฃ ฅ ฉ ฌ ฝ ฟ ห อ ฮ

Par ailleurs, il existe des doubles consonnes terminales.

2.2.3. *Voyelle initiale*

C'est la voyelle qui s'écrit toujours devant la consonne initiale. Elle est obligatoirement suivie par une consonne initiale (éventuellement double).

2.2.4. *Voyelle diacritique*

Ce sont les voyelles qui s'écrivent toujours au-dessus ou au-dessous de la consonne initiale. En codes de caractère informatique, ces voyelles se placent toujours derrière la consonne initiale.

Pour utiliser ce type de voyelles, une consonne initiale (éventuellement double) est exigée.

2.2.5. *Voyelle terminale*

C'est la voyelle qui s'écrit toujours derrière la consonne initiale. Une consonne initiale (éventuellement double) est également exigée.

2.2.6. *Marque de ton*

À l'exception du ton moyen dont le graphème n'existe pas, la marque de ton s'écrit au-dessus de la consonne initiale ou de la voyelle diacritique si cette dernière existe. Son code informatique se place toujours derrière les deux dernières

### 2.3. Mot

Un mot en thaï est composé par une ou plusieurs syllabes. La plupart des mots thaï sont monosyllabiques, mais il existe un certain nombre de mots polysyllabiques qui ont des origines différentes telles que pālie, sanskrite, khmère, chinoise, européennes, etc.

Les mots thaï sont invariables, il n'y a aucune flexion. Les verbes ne se conjuguent pas. Les noms et les adjectifs sont également invariables en genre et en nombre.

### 2.4. Phrase

La structure d'une phrase simple en thaï est composée de 3 éléments :

Sujet + Verbe + Complément

## 3. INTEX version thaï

Le thaï est déjà intégré dans INTEX grâce à ses alphabets ASCII étendus à 8 bits (UTF-8). Un dictionnaire de mots simples (DELAS) pour le thaï a été créé. Il contient 33 666 mots classés en 7 catégories à savoir : Nom, Pronom, Verbe, « Adjectif-verbe », Préposition, Conjonction et Interjection.

Étant donné que les mots thaï sont invariables, le dictionnaire des formes fléchies (DELAF) n'existe pas. Par ailleurs, faute de séparateur de mots, les mots composés sont considérés comme des mots simples et il n'y a donc pas de DELAC.

Remarque : Conformément à la grammaire thaï, l'adjectif et l'adverbe sont classés dans la même catégorie appelée ici « l'adjectif-verbe » (ADJV).

La figure 1 montre le fonctionnement d'INTEX pour le thaï, quand des séparateurs de mots et de phrases ont été ajoutés manuellement dans le corpus.

## 4. Reconnaissance des mots thaï

À cause de l'ambiguïté des consonnes qui peuvent être initiales ou terminales, il est difficile de reconnaître correctement les syllabes qui forment les mots ; considérons par exemple :

ตากลม → $CiVtC_1C_2Ct$

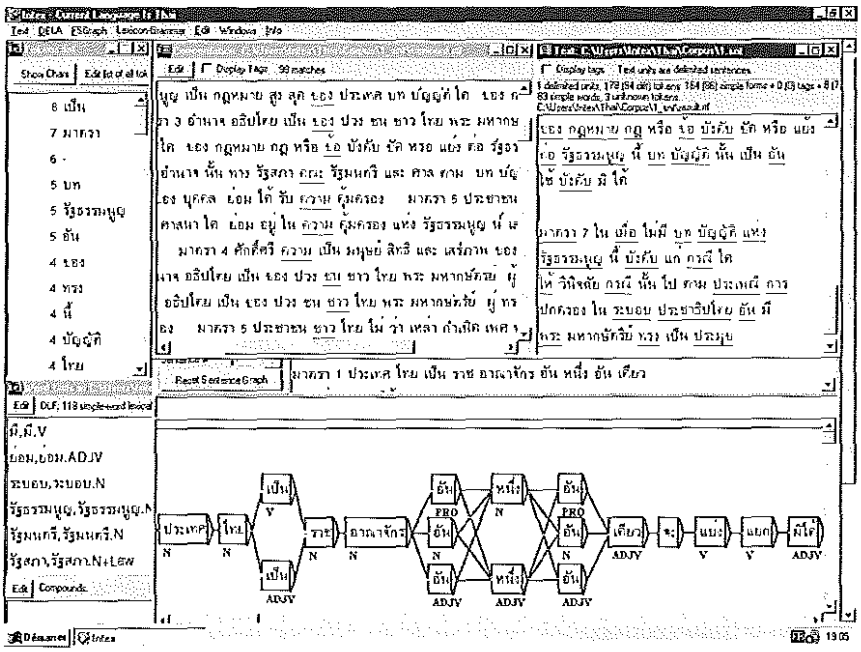

*Fig. 1 : La première version d'INTEX pour le thaï*

$C_1$ et $C_2$ sont ambigus, car deux segmentations de mot sont possibles dans ce cas. Soit $C_1$ est la consonne terminale d'un premier mot, soit $C_1$ et $C_2$ sont la double consonne initiale d'un second mot. Les deux segmentations donnent deux sens différents :

ตาก/ลม → CiVtCt/CiCt → sécher/vent

ตา/กลม → CiVt/CCiCt → œil/rond

### 4.1. Méthode de lexème à caractère unique

En analysant les mots composés en français, INTEX est capable de reconnaître les mots « pomme de terre » et « terre cuite » dans la séquence de lexèmes « pomme de terre cuite » selon le schéma présenté ci-dessous :

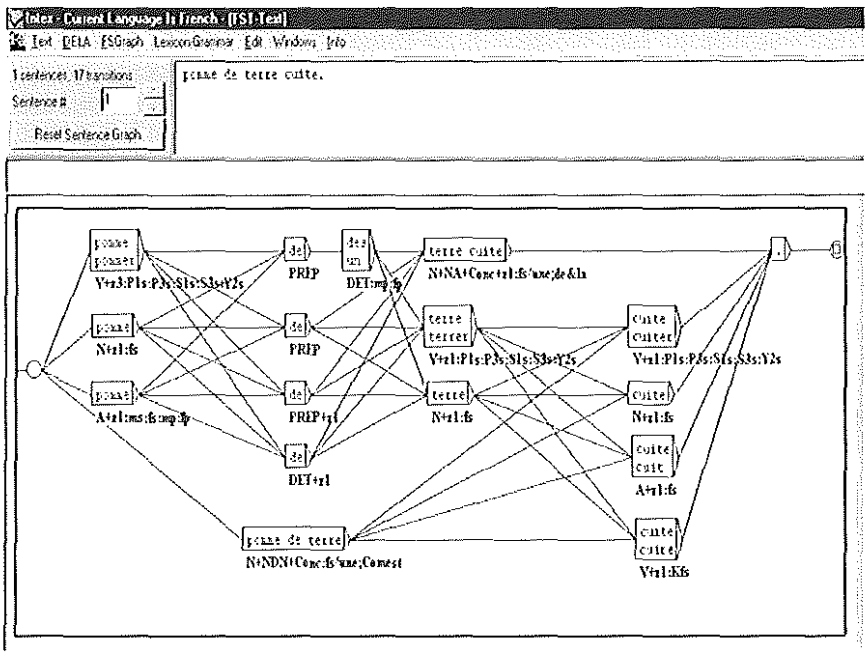


*Fig. 2 : La reconnaissance des mots composés en français*

La procédure de reconnaissance de mot composé en français fonctionne par lexème de telle manière que plusieurs lexèmes qui se suivent peuvent former un mot composé. Si l'on considère que chaque lexème ne contient qu'un seul caractère et que plusieurs lexèmes (à un caractère) qui se suivent peuvent former un mot, INTEX serait égale-

ment capable dans ce cas de reconnaître les mots thaï. Pour cela, il suffit d'ajouter automatiquement un espace blanc entre chaque caractère thaï dans le corpus ainsi que dans tous les mots du dictionnaire et de placer ce dernier dans le répertoire des dictionnaires de mots composés d'INTEX. Désormais, INTEX considère que tous les mots thaï sont des mots composés et que chaque caractère thaï est un mot simple.

La figure 3 montre comment, grâce à cette méthode, INTEX est capable de reconnaître tous les mots qui se trouvent dans le dictionnaire. Cependant, INTEX reconnaît aussi un certain nombre de mots orthographiquement corrects, mais qui n'existent pas dans le corpus. Ces mots-là sont considérés comme du « bruit ».

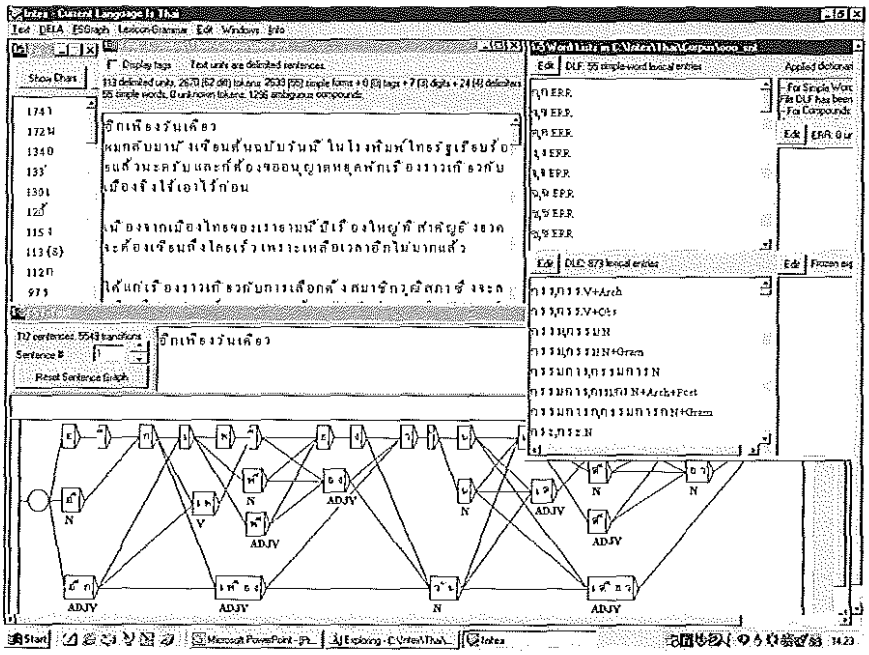

*Fig. 3 : La reconnaissance des mots thaï par lexème à caractère unique*

Ce « bruit » peut être illustré par un exemple en français comme suit :

a u **c o m m e n c e m e n t** d u m o i s...

c e,c e.DET+z1:ms
c e,c e.PRO+z1:3s
c e,c e.PRO+z1:3s:3p
c o m m e,c o m m e.ADV+z1
c o m m e,c o m m e.CONJS+z1
c o m m e n c e,c o m m e n c e r.V+z1:P1s:P3s:S1s:S3s:Y2s
c o m m e n c e m e n t,c o m m e n c e m e n t.N+z1:ms
e n,e n.PREP+z1
e n,e n.PRO+z1
m e,m e.PRO+z1:1s
m e n t,m e n t i r.V+z1:P3s

En appliquant cette méthode uniquement au mot « commencement » qui se trouve dans le texte en français modifié ci-dessus, INTEX trouve 7 décompositions du mot orthographiquement correctes alors qu'il n'existe qu'un seul mot. Les 6 autres interprétations sont considérées comme du bruit.

Afin de réduire le bruit, certaines contraintes sont nécessaires.

## 4.2. Contraintes orthographiques

### 4.2.1. *Bigramme inexistant*

Quand on analyse tous les mots du dictionnaire, on constate que certains bigrammes n'apparaissent jamais. Par exemple, en français, il n'y a aucun mot dans lequel un caractère « b » est suivi par un « f ». Ce qui veut dire que, si l'on trouve un bigramme « bf » dans une séquence de lettres qui se suivent, on est certain que ces deux caractères n'appartiennent pas au même mot et qu'on peut ajouter un séparateur entre eux.

Cette méthode dépend évidemment du degré de complétude du dictionnaire. À chaque fois que des nouveaux mots sont ajoutés, l'analyse doit être refaite afin de vérifier si la liste des bigrammes est toujours valide.

### 4.2.2. *Syllabe acceptable*

Représentons la syllabe thaï, en ajoutant un espace blanc entre chaque caractère conformément à la méthode des lexèmes à caractère unique, la structure de syllabe prend la forme suivante :

[Vi] Ci [Vd] [T] [Vt] [Ct]

En analysant la caractéristique de chaque unité, on remarque que la plupart des consonnes sont ambiguës, car elles peuvent être initiales ou terminales. Par contre, les Vi, Vd, T et Vt ne le sont pas parce que leur position a été déjà déterminée par la forme de structure présentée ci-dessus. De plus, l'apparition de ces caractères non-ambigus réclame l'existence de la consonne initiale. Autrement dit, elles ne peuvent pas former isolément une bonne syllabe sans consonne initiale. En conséquence, les espaces blancs entre les Vi, Vd, T, Vt et la consonne initiale peuvent être supprimés afin de former une syllabe acceptable.

[Vi]Ci[Vd][T][Vt] [Ct]

En supprimant les espaces blancs derrière la Vi et devant les Vd, T et Vt dans tous les mots du corpus ainsi que dans tous les mots du dictionnaire, on arrive à créer des groupes de syllabes acceptables et chaque groupe est séparé par un espace blanc. Ce qui veut dire que le lexème ne contient plus un seul caractère, mais plutôt une syllabe. Grâce à cette méthode, les bruits sont notablement réduits et le graphe est beaucoup moins complexe (voir la figure 4).

## 5. Reconnaissance des phrases thaï

Faute de point indiquant la fin de phrase, la reconnaissance des phrases thaï n'est pas évidente. D'autant plus que l'emploi du point-virgule, du point d'exclamation et du point d'interrogation servant à déterminer la fin de phrase n'est pas non plus obligatoire. En effet, deux phrases différentes sont généralement séparées par un simple espace blanc. L'espace blanc est donc ambigu, puisqu'il peut également être utilisé en tant que séparateur de mot.

En fait, la reconnaissance des phrases thaï se fait plutôt par la structure de phrase, donc après avoir reconnu clairement les mots et leur partie du discours. Cela exige une analyse complète de la structure de la langue thaï. Cependant, à cause de leur invariabilité, certains mots thaï ont une même forme pour le nom, le verbe et l'adjectif-verbe. En l'absence de particule ou de postposition servant à indiquer des rapports grammaticaux, comme cela existe dans les langues coréenne et japonaise, la partie de discours des mots thaï n'est pas

déterminée de manière univoque. Cette circonstance entraîne une difficulté de reconnaissance des phrases, d'autant plus que la structure des phrases thaï est très variée. Il existe, par exemple, des cas où le sujet et le verbe sont inversés et même des cas où le sujet et le verbe sont facultatifs.

À cause de ces restrictions, il est préférable d'envisager d'autres procédures.

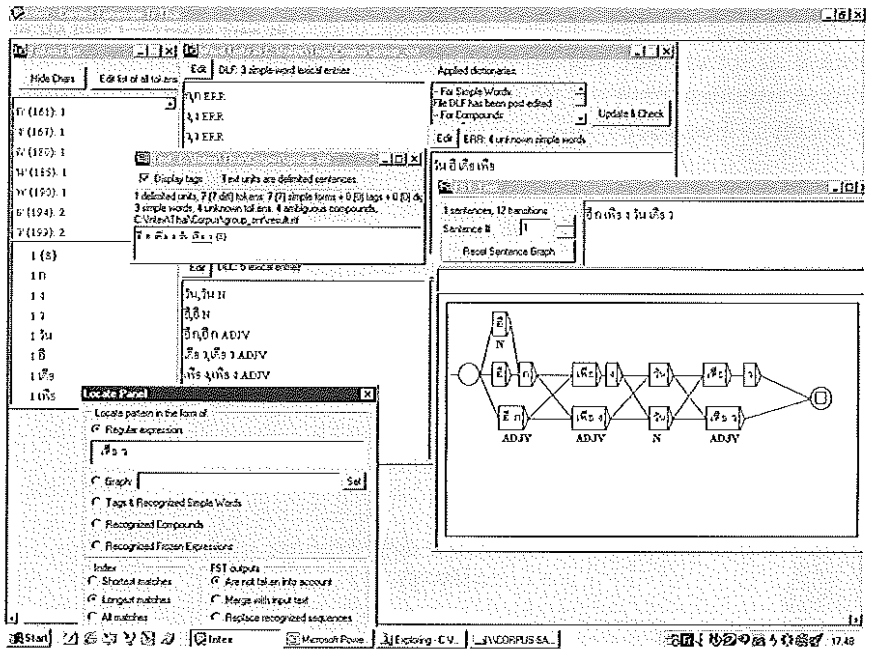


*Fig. 4 : La reconnaissance des mots thaï par syllabe acceptable*

## 5.1. Reconnaissance des paragraphes

Un bon style de rédaction d'un texte en thaï favorise une phrase longue et complexe avec beaucoup de conjonctions. Il arrive souvent qu'un paragraphe contienne seulement une ou deux phrases. Si l'on arrive à détecter toutes les fins de paragraphe, on pourra reconnaître un certain nombre de fins de phrase.

Le « Preprocessing » d'INTEX ne connaît pas le caractère « Retour à la ligne » (\n) servant à séparer les paragraphes. Pour cela, il faut

traiter le corpus hors INTEX en insérant, devant le « \n », un signe qui sera remplacé lors du « Preprocessing » par un séparateur de phrase.

### 5.2. Reconnaissance du début et de la fin de phrase par mots clés

Compte tenu du fait que certains mots en thaï sont écrits toujours au début ou à la fin de la phrase, ils peuvent être employés comme des mots clés pour diviser une séquence de mots en deux phrases différentes, en ajoutant entre elles, un séparateur de phrases.

Afin d'éviter une double insertion dans le cas où la première phrase se terminerait par un mot clé de fin de phrase et la deuxième phrase commencerait par un mot clé de début de phrase, il faut utiliser le graphe « Sentence » présenté ci-dessous :

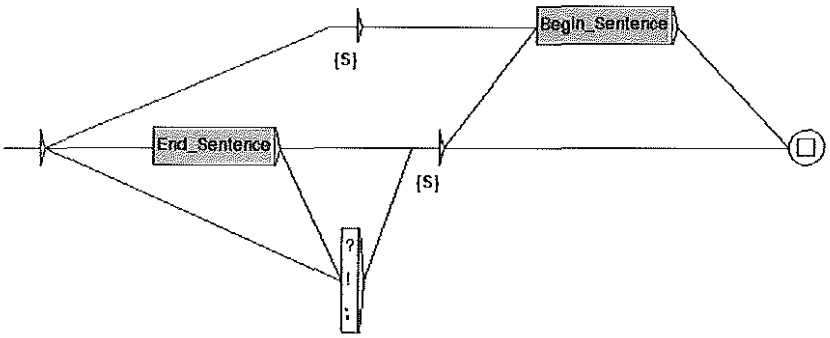


*Fig. 5 : Le graphe « Sentence » dans la phase « Preprocessing » d'INTEX*

## 6. Conclusion et perspective

Malgré toutes les difficultés discutées dans cet article, INTEX arrive à reconnaître relativement bien les mots et les phrases thaï. Augmenter le nombre des mots du dictionnaire ainsi que le nombre des mots clés sera un des moyens efficaces pour perfectionner le programme. Néanmoins, n'étant pas prévu pour le traitement des langues non européennes, INTEX mériterait quelques modifications.

Ajouter des espaces blancs dans le corpus n'est pas une manière naturelle de procéder. En créant un graphe ou en cherchant une expression régulière, il est nécessaire d'ajouter un blanc entre chaque

caractère ou entre chaque syllabe selon la méthode utilisée. Ceci complique son utilisation. En outre, dans certaines langues comme l'allemand et le coréen, il existe autant des mots isolés non-ambigus que des mots agglutinants. Bien que l'insertion automatique des blancs permette au programme de reconnaître des mots agglutinants, elle crée simultanément de nouvelles ambiguïtés aux mots isolés.

Afin de résoudre ce problème, un nouveau dictionnaire intitulé « Dictionnaire Electronique des Mots Agglutinants » (DELAA) sera proposé prochainement. Il contiendra, pour le thaï, des mots aux formes simples, mais probablement pour des autres langues flexionnelles, des mots aux formes fléchies. Ce dictionnaire servira de dictionnaire principal pour les langues agglutinantes comme le chinois, le japonais et le thaï. Dans d'autre cas, il sera un supplément s'appliquant seulement aux mots non reconnus par les DELAS et DELAC dans l'hypothèse où ces mots-là pourraient être des mots agglutinants.

Une des créations majeures serait le programme de consultation de dictionnaire de mots agglutinants dit « dicoa.exe » qui adopterait la technique appliquée aux mots composés, soit « dicoc.exe » (SILBERZTEIN 2000, p. 178). Les deux programmes se distinguent par le fait que les mots sont vérifiés caractère par caractère dans le premier cas, lexème par lexème dans le deuxième cas. L'utilisation du DELAA et du « dicoa » permettrait d'analyser des textes contenant des mots agglutinants sans ajouter de blancs.

Cependant, en l'absence des contraintes orthographiques, le problème du bruit (signalé au paragraphe 4.1) réapparaîtra inévitablement. Étant donné que chaque langue possède des contraintes différentes, il serait préférable de donner à chaque utilisateur d'INTEX un moyen de définir lui-même ses propres contraintes.

En fait, la structure d'INTEX la mieux adaptée pour le thaï devrait ressembler à celle de la figure 6 ci-dessous.

En conclusion, les mots et les phrases thaï pourraient être bien davantage reconnus grâce à l'addition des 3 modules spécifiques illustrés dans la figure 6.

- En ajoutant des contraintes orthographiques plus concrètes, les syllabes thaï seraient mieux identifiées.

- Ensuite, l'application des dictionnaires, en particulier ceux contenant des mots agglutinants, rendrait le logiciel apte à la reconnaissance des mots thaï.
- Enfin, grâce à la recherche plus complète sur la structure de la langue thaï, INTEX serait, dans la phase de désambiguïsation, capable de bien déterminer la limite des phrases thaï.

Il faut noter que, contrairement au français, la reconnaissance des phrases thaï devrait être réalisée, de préférence, dans une phase tardive de l'analyse des textes.

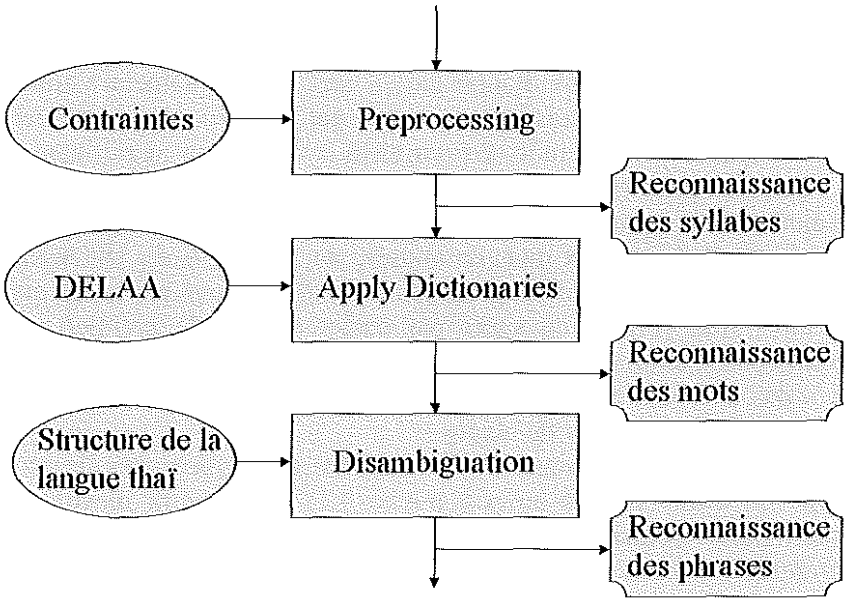


*Fig. 6 : La structure d'INTEX pour le thaï*

**Remerciements**

Je remercie vivement Cédrick Fairon, Max Silberztein, Blandine Courtois et Cristina Mota pour leurs idées et leurs explications. Je remercie également Surasak Na Nakorn et Cédrick Fairon pour la correction de cet article.

## Bibliographie

COMRIE (Bernard) : 1987, ed. *The World's Major Languages* (London : Croom Helm.).

GREFENSTETTE (Gregory) and TAPANAINEN (Pasi) : 1994, « What is a word, What is a sentence? Problems of Tokenization », *Proceedings of the 3rd International Conference on Computational Lexicography and Text Research (COMPLEX'94)*, p. 79–87.

GROSS (Maurice) : 1989, « La construction de dictionnaires électroniques », *Annales des Télécommunications*, Tome 44, n° 1–2, p. 4–19.

HUDAK (Thomas John) : 1987, « Thai », in COMRIE (Bernard) : 1987, p. 757–775.

MAUREL (Denis) : 1989, « Reconnaissance de séquences de mots par automates, adverbes de date du français », Thèse de Doctorat en informatique (Paris : Université Paris VII).

NAGATA (Masaaki) : 1999, « A Part of Speech Estimation Method for Japanese Unknown Words using a Statistical Model of Morphology and Context », *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, p. 277–284.

PANUPONG (Vichin) : 1982, *The Structure of Thai* [in Thai] (Bangkok : Ramkhamhaeng University).

SILBERZTEIN (Max) : 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX* (Paris : Masson).

SILBERZTEIN (Max) : 2000, *INTEX* (Paris : LADL).

SIRIBOONMA (Aurapin) : 1999, « Erreurs et Difficultés en français rencontrées par des étudiants thaïlandais au niveau universitaire, à partir de l'analyse des productions écrites », Thèse de Doctorat en Linguistique (Paris : Université Paris V).

SPROAT (Richard), SHIH (Chilin), GALE (William), CHANG (Nancy) : 1996, « A Stochastic Finite-State Word-Segmentation Algorithm for Chinese », *Computational Linguistics*, 22,3, p. 377–404.

THAWARANON (Kobkool) : 1978, « Spacing in the Thai Writing » [in Thai], Master Thesis of Arts, Department of Thai Language (Bangkok : Chulalongkorn University).