Institut National des Langues et
Civilisations Orientales

# Pratique de la lecture en L2 : classification automatique de textes en thaï par progression lexicale

Jennifer Lewis-Wong

Master 2 Traitement Automatique des Langues, spécialité
Ingénierie multilingue

Encadré par

François Stuck et Mathieu Valette

2 octobre 2015

inalco
Institut national
des langues
et civilisations orientales

# Remerciements

Je remercie tout ceux qui m'ont soutenu dans l'élaboration de ce mémoire, en particulier mes deux directeurs de mémoire, François Stuck ฟรื่องซัว สตูก et Mathieu Valette มาทีเออ วาแลต.

Je tiens également à remercier tous mes professeurs de la section de siamois à l'INaLCO pour leur patience, savoir et enthousiasme au fil des années, en particulier Gilles Delouche ชิลล์ เดอลูช, Apisit Waraeksiri อภิสิทธิ์ วราเอกศิริ, Koson Thanadsamran โกศล ถนัดสำราญ et Pensiri Charœnpote เพ็ญศิริ เจริญพจน์.

Je remercie aussi mes professeurs et camarades de Master de leur enthousiasme contagieux pour le traitement automatique des langues.

# Table de matières

1. Introduction ..... 5
1.1. Conventions ..... 6
2. La langue thaïe ..... 8
2.1. La place du thaï dans la société thaïlandaise ..... 8
2.2. L'enseignement du thaï langue étrangère ..... 9
2.3. Caractéristiques générales du thaï ..... 10
2.3.1. Caractéristiques phonologiques ..... 10
2.3.2. Le lexique ..... 11
2.3.2.1. Les mots composés ..... 11
2.3.2.2. Les classificateurs ..... 12
2.3.3. Niveaux de langue ..... 13
2.4. Le système d'écriture ..... 13
2.4.1. La syllabe thaïe ..... 14
2.4.2. Les signes consonantiques ..... 14
2.4.3. Les signes vocaliques ..... 14
2.4.4. Les tons ..... 15
2.4.5. Règles de prononciation particulières ..... 16
2.4.6. Signes supplémentaires ..... 16
2.4.7. Conventions typographiques ..... 17
2.4.8. Difficultés du système d'écriture pour l'étudiant en thaï langue étrangère ..... 18
2.4.9. L'encodage du thaï en Unicode ..... 19
3. L'état de l'art ..... 21
3.1. Ressources pédagogiques pour la lecture en thaï ..... 21
3.1.1. Les publications visant une pratique de la lecture ..... 21
3.1.2. Les ressources électroniques ..... 22
3.1.2.1. Logiciels ..... 27
3.2. Ressources électroniques pour la lecture avec classification automatique de textes ..... 28
3.2.1. REAP ..... 28
3.2.2. TextLadder et TextGrader ..... 29
3.3. L'élaboration et l'utilisation des listes de fréquence lexicale ..... 31
3.3.1. Listes de fréquence lexicale en thaï ..... 34
3.4. La segmentation lexicale en thaï ..... 36

3.4.1. Le mot thaï ..... 36
3.4.2. Les outils de segmentation automatique du thaï ..... 38
4. Méthodologie ..... 39
4.1. Corpus ..... 39
4.2. Élaboration des listes de fréquence lexicale pour le thaï ..... 40
4.2.1. Évaluation des listes de fréquence ..... 40
4.2.2. Détermination du seuil du lexique de haute fréquence ..... 42
4.2.3. Lexique de basse fréquence ..... 44
4.3. Présentation de *ThaiTextLadder* ..... 45
5. Tests sur corpus et résultats ..... 48
5.1. Les maux de la segmentation ..... 48
5.2. Tri par quantité de vocabulaire connu ..... 49
5.3. Tri par couverture textuelle ..... 51
5.4. Homogénéité de corpus ..... 53
5.5. Test sur corpus de nouvelles ..... 55
5.6. Séquence de lecture et niveau de compétence standardisé ..... 59
6. Discussion ..... 61
7. Conclusion ..... 64
8. Table de figures ..... 65
9. Liste des tableaux ..... 66
10. Bibliographie ..... 67
11. Œuvres de référence ..... 72
11.1. Dictionnaires de thaï ..... 72
11.2. Dictionnaires de fréquence lexicale ..... 72
11.3. Manuels de thaï langue étrangère ..... 73
12. Annexes ..... 74
12.1. Les niveaux de compétences du test CU-TFL et ses équivalents ..... 74
12.2. Règles d'utilisation de l'espace typographique en thaï ..... 75
12.3. Liste des abréviations ..... 76

# 1. Introduction

Pour l'étudiant d'une langue étrangère peu enseignée et peu dotée en matériel pédagogique ou ressources d'apprentissage, l'accès que donne l'Internet aux textes authentiques écrits dans sa langue de prédilection fournirait l'occasion d'agrandir son vocabulaire par la lecture régulière, en même temps qu'il accroît ses connaissances culturelles. Les effets positifs de la lecture personnelle régulière sur tous les aspects de l'apprentissage des langues étrangères, soulignés par Krashen (2004a), inciteraient tout étudiant et professeur à étendre cette pratique. Toutefois, face à un texte authentique, même l'étudiant le plus enthousiaste se trouve souvent confronté à une réalité inconfortable : son niveau de connaissance du vocabulaire du texte s'avère insuffisant pour en tirer du sens, encore moins pour extraire du vocabulaire nouveau à partir de son contexte. L'étudiant peut perdre sa motivation, facteur non négligeable dans le processus d'acquisition d'une langue étrangère (Dörnyei, 1998). Pour l'enseignant de langues peu enseignées aussi, trouver des ressources d'enseignement pour la lecture adaptées au niveau des étudiants est également un défi : il doit baser ses choix sur sa connaissance fine de ses étudiants et le programme qu'ils ont suivi.

Comment choisir automatiquement un texte qui correspond aux compétences linguistiques de l'étudiant ? Logiquement, il faudrait non seulement classifier le niveau de difficulté des textes, mais cette classification devrait prendre en compte le profil et les compétences linguistiques de l'étudiant. Ce projet s'intéresse spécifiquement à la classification des textes par niveau de difficulté, la *lisibilité*. Même si l'évaluation des connaissances linguistiques et le profil de l'étudiant ne font pas partie de notre travail, nous tenons compte de ces aspects dans l'élaboration du projet.

Historiquement, les chercheurs en lisibilité ont recours à deux types de stratégies pour déterminer la difficulté des textes. La première est l'élaboration de formules de lisibilité qui s'appuient sur des mesures des caractéristiques superficielles des textes, telles que les proportions de mots et de syllabes dans la phrase (l'indice FKGL[1]), la fréquence des mots polysyllabiques (l'indice SMOG[2]) ou la longueur de la phrase et le pourcentage de mots complexes (l'indice GFI[3]).[4] Plus récemment, les avancées en TAL ont permis le développement de modèles statistiques de lisibilité plus complexes, basés sur des corpus de textes dont la difficulté a déjà été mesurée, comme des manuels scolaires.

Ces méthodes d'évaluation de la lisibilité de textes posent plusieurs difficultés pour la lisibilité des langues peu enseignées et pour un apprenant L2. D'abord, la plupart des travaux sur la lisibilité mesurent la difficulté des textes pour les locuteurs natifs, et il ne va pas de soi que la lisibilité d'un texte pour le lecteur en L1 soit la même que pour le lecteur en L2 (François, 2011, p196). Il faut donc développer de nouvelles formules, ou créer des modèles basés sur des corpus de textes destinés aux lecteurs en L2 déjà classifiés par niveau de difficulté. La deuxième difficulté est que ces méthodes sont développées sur une langue spécifique (jusqu'alors, la

[1] FKGL = Flesch-Kincaid Grade Level

[2] SMOG = Simple Measure of Gobbledygook. Cette formule calcule la lisibilité pour un morceau de texte de trente phrases.

[3] GFI = Gunning-Fog Index

[4] Voir DuBay (2004) pour une histoire détaillée des formules de lisibilité. L'outil en ligne de commande Unix *style* (qui s'installe avec l'installation de l'outil *diction*) permet d'évaluer la lisibilité de textes en anglais, allemand ou néerlandais au moyen de diverses formules de lisibilité.

recherche sur la lisibilité en L2 s'est concentrée sur l'anglais[5]) et ne sont pas nécessairement adaptées ou facilement adaptables à d'autres langues, bien que parfois une formule de lisibilité puisse donner des résultats satisfaisants pour des langues complètement différentes[6].

Nous nous sommes donc posé la question suivante : existe-t-il un moyen de proposer des textes appropriés au niveau du lecteur non natif qui ne dépendrait pas de ces mesures de lisibilité? L'inspiration de ce projet vient du travail de Ghadirian (2002), et son logiciel *TextLadder*, qui agence des textes d'un corpus de textes en anglais dans un ordre qui maximalise la facilité de lecture pour l'étudiant, en calculant la similarité lexicale entre les textes. Notre but était d'identifier les étapes nécessaires pour porter le logiciel *TextLadder* à une autre langue spécifique, le thaï, espérant ainsi ouvrir la voie à son adaptation à d'autres langues. Le développement de méthodes spécifiques à une langue étant souvent problématique en termes de rentabilité quand les effectifs d'apprenants sont restreints, il est intéressant de considérer comment adapter des ressources existantes ou de développer des méthodes génériques adaptables pour la création de matériels pédagogiques.

Nous commençons ce mémoire par une présentation de la langue thaïe et son système d'écriture, avec une attention particulière accordée à la langue thaïe comme langue étrangère et les difficultés spécifiques qu'elle pose aux apprenants en L2. Deuxièmement, nous dressons un état de l'art sur les publications, outils et méthodes qui concernent notre projet ; après avoir passé en revue les ressources pédagogiques pour la lecture en thaï L2, nous nous tournons vers les ressources électroniques pour la lecture avec classification automatique de textes, l'élaboration des listes de fréquence lexicale et la segmentation du texte thaï en mots. Nous décrivons ensuite les étapes dans la création de *ThaiTextLadder*, les corpus utilisés lors de sa réalisation et de nos différents tests d'évaluation, la méthodologie employée pour la création de listes lexicales dont dispose le logiciel et enfin une description du fonctionnement de *ThaiTextLadder* lui-même. Les résultats de tests sur corpus sont suivis d'une discussion d'analyse du projet.

## 1.1. Conventions

Tout locuteur d'une langue possède un sens de ce que c'est qu'un *mot* dans sa langue, mais ce terme est ambigu (Müller, 1977), car le locuteur confond sa notion abstraite du *mot* et le *mot* concret sur la page. Par *mot*, nous entendons un élément graphique identifiable et mesurable d'un texte, d'une liste. Le nombre de *mots* étant le nombre d'éléments du texte. Les différents éléments qui composent le texte sont les *vocables* du texte, son *vocabulaire*. Un même vocable peut prendre des formes différentes, les différentes flexions d'un même verbe, par exemple. Un *glossaire* (appelé aussi *lexique*) regroupe des éléments de vocabulaire d'un corpus de textes. Le *lexique* est l'ensemble de tous les vocables possibles dans une langue, les *unités lexicales* (ou *lexèmes*). Le *lemme* est la forme représentative des différentes formes d'un même lexème, utilisée comme entrée principale dans un dictionnaire.

En traitement automatique des langues, il est d'usage de travailler sur les *mots orthographiques*, les éléments du texte séparés d'une espace typographique, appelés aussi des

[5] Pour l'état de l'art de la recherche sur la lisibilité en L2, voir François (2011).

[6] Das & Roychudhury (2004; 2006), cité par Islam (2012) notent que l'indice FKGL donne des résultats satisfaisants pour le bengali, par exemple.

*tokens* (en anglais *running words*), mais comme nous verrons dans ce mémoire, il n'est pas possible de définir les mots orthographiques de cette façon dans le traitement automatique d'une langue comme le thaï.

Nous employons le mot *thaï,* qui s'accorde en genre et en nombre, pour désigner ce qui se rapporte à la Thaïlande: la société *thaïe,* le *thaï*… En français, le mot *siamois* est souvent synonyme de *thaï,* mais nous le réservons pour distinguer la langue thaïe des autres langues de la famille de langues taï.

Les mots en thaï sont suivis de leur transcription en API[7] entre barres obliques //, accompagnés, si nécessaire, de la traduction en italiques.

[7] API = Alphabet phonétique international. Ces transcriptions proviennent de l'outil de transcription phonémique disponible sur ce site http://www.thai-language.com/dict

# 2. La langue thaïe

Nous commençons avec une brève introduction à la langue thaïe, pour placer notre travail dans son contexte, avant de présenter la langue et son écriture. Nous ne présentons en détail que les aspects de la langue pertinents à cette étude ou nécessaires pour comprendre les explications qui suivent.

## 2.1. La place du thaï dans la société thaïlandaise

Le *thaï* (en thaï ภาษาไทย /pʰaːsăːtʰaj/), aussi appelé *thaï standard*, *thaï central* ou *thaiklang*[8], est la langue officielle du Royaume de Thaïlande, langue de l'éducation nationale et des médias. La promotion et le développement du thaï dans la société thaïlandaise sont assurés non seulement par l'usage exclusif de la langue dans la vie publique, mais aussi explicitement par une organisation gouvernementale, la *Société Royale de Thaïlande*[9] (en thaï ราชบัณฑิตยสภา /râːttɕʰábantʰíttàjásàpʰaː/). Parmi ses nombreuses activités, la Société fait office d'académie des lettres, à l'instar de notre Académie française. Elle publie le dictionnaire officiel de la langue, le RID ou *Royal Institute Dictionary* พจนานุกรม ฉบับราชบัณฑิตยสถาน[10] (/pʰótnaːnúkrom tɕʰàbàprâːtbantʰít jóttʰăːn/) et le système officiel de transcription du système d'écriture thaï en lettres latines[11], utilisé sur les panneaux de signalisation routière et dans les publications officielles. La *Société* publie aussi des guides concernant l'usage de la langue, tels que la prononciation et l'orthographe ou la transcription/translittération des mots étrangers en lettres thaïes[12]. Ces derniers expliquent non seulement comment intégrer les entités nommées, mais établissent des principes de correspondance phonologiques pour emprunter des mots étrangers. Notons aussi deux publications concernant des types de vocabulaire spécifiques au thaï et inexistants en français, sur les classificateurs[13] (ลักษณนาม /láksàʔnàʔnaːm/) et le vocabulaire royal[14] (ราชาศัพท์ /raːtɕʰaːsàp/).

Basé sur le dialecte central du pays, le thaï standard, bien qu'officiellement compris par tous les habitants du pays (plus de soixante-sept millions de personnes), n'est en réalité la langue maternelle que des vingt millions d'habitants de la région centrale (qui comprend aussi la capitale, Bangkok)[15], la grande majorité de la population étant des locuteurs d'autres langues de la famille de langues *taï* (Laungaramsri, 2003, p.158). À ceux-ci s'ajoutent les locuteurs des plus de soixante-dix langues des communautés considérées comme « non-Thaï » : les minorités ethniques autochtones, les immigrants (dont une communauté chinoise importante) et des

[8] En linguistique, l'appellation *siamois* permet de distinguer le thaï standard des autres langues de la famille de langues taï, mais cette appellation n'est plus utilisée dans le langage courant et de moins en moins utilisée dans la littérature scientifique.

[9] Son appellation officielle en anglais est la *Royal Society of Thailand*, autrefois connue sous le nom de *Royal Institute of Thailand* (en thaï ราชบัณฑิตยสถาน /râːttɕʰábandìttàʔjáʔsàtʰăːn/)

[10] Toutes les publications de la *Royal Society* sont publiées en version papier et sont disponibles sur l'Internet. Le dictionnaire est accessible à cette adresse : http://www.royin.go.th/dictionary/

[11] http://www.royin.go.th/wp-content/uploads/2015/03/416_2157.pdf

[12] Les langues concernées sont l'anglais, le japonais, le français, le malais, l'allemand, le russe, l'espagnol, l'arabe, l'italien, le coréen, le vietnamien, le chinois et le hindi. http://www.royin.go.th/?page_id=617

[13] http://www.royin.go.th/?page_id=641

[14] http://www.royin.go.th/?page_id=643

[15] Ethnologue http://www.ethnologue.com/country/TH/languages et http://www.ethnologue.com/language/tha

réfugiés[16] (notamment de la Birmanie). Ces communautés sont souvent désireuses de s'intégrer dans la culture dominante, et la langue thaïe joue un rôle important dans la politique officielle d'assimilation. La ligne officielle qui revendique l'homogénéité du peuple thaï commence à reconnaître ce multiculturalisme de fait, permettant des initiatives de revitalisation linguistique (Barry, 2013). Il est donc difficile d'estimer le nombre de personnes ayant le thaï comme deuxième langue et d'en juger leurs niveaux de compétence linguistique réelle. Puisque leur apprentissage de la langue thaïe se déroule dans le milieu scolaire, avec le thaï comme première langue d'enseignement, nous ne pouvons pas considérer ces populations comme les apprenants de thaï enseigné comme langue étrangère, mais comme étudiants de thaï comme langue seconde.

## 2.2. L'enseignement du thaï langue étrangère

L'apprentissage du thaï comme L2 (voire L3 ou plus) se déroule principalement à l'âge adulte et bien qu'un nombre restreint d'établissements d'enseignement supérieur étrangers proposent des programmes approfondis en langue thaïe, grâce à l'attractivité du pays comme destination touristique et de retraite, couplée avec son importance économique et sa position stratégique au sein de l'ASEAN, la langue thaïe, malgré sa difficulté apparente, attire suffisamment d'étudiants pour que l'enseignement du thaï comme langue étrangère (*Teaching Thai as a Foreign Language, TFL)* soit enseigné au niveau universitaire en Thaïlande. Il existe même un test de niveau normalisé, mis au point à l'Université de Chulalongkorn en 2012, *The Chulalongkorn University Proficiency Test of Thai as a Foreign Language*[17] (CU-TFL)[18].

Le thaï demeure toutefois une langue peu enseignée comme langue étrangère, même en Thaïlande. Les statistiques officielles qui portent sur les étudiants étrangers en Thaïlande pour l'année 2012[19] démontrent la prépondérance des programmes qui utilisent l'anglais comme langue d'instruction, avec seulement 11% des étudiants étrangers inscrits dans des programmes de langue thaïe ou langue et culture thaïes (987 étudient la langue thaïe exclusivement, 2105 en total). L'écrasante majorité d'entre eux (1958 étudiants) sont originaires de la Chine Populaire, avec des nombres significatifs originaires du Japon (34 étudiants), du Vietnam (63 étudiants) et de la Corée (27 étudiants). Il faut aussi noter que le nombre d'étudiants étrangers en thaï a été divisé par trois depuis son apogée en 2009, à la suite des troubles politiques qui sont survenus en Thaïlande depuis cette année-là.

Il existe de nombreux programmes universitaires de langue et civilisation thaïes en dehors du pays, le plus ancien étant l'enseignement du thaï (ou *siamois*) à l'INaLCO, qui a commencé en 1876 sous l'impulsion des intérêts politiques de la France dans la région; la première chaire de siamois a été inaugurée en 1899. Des cursus approfondis en études thaïes

[16] Selon le programme HCR de l'ONU, le nombre d'apatrides en Thaïlande pour l'année 2014 serait de 644,761 personnes. Source : UNHCR country operations profile - Thailand http://www.unhcr.org/pages/49e489646.html

[17] http://www.sti.chula.ac.th/academic/non-native/CU-TFL

[18] Ce test évalue les quatre compétences linguistiques classiques (lecture, écriture, compréhension et production orales). Les candidats sont classés en cinq niveaux propres au CU-TFL. Voir annexe 11.1. pour les niveaux équivalents à d'autres échelles d'évaluation. Un test d'autoévaluation utilisant les niveaux CECR, utilisé à L'Université d'Hawaï est disponible ici : http://www.hawaii.edu/thai/placement.htm

[19] สำนักงานคณะกรรมการการอุดมศึกษา, นักศึกษาต่างชาติที่ศึกษาในสถาบันอุดมศึกษาสังกัดสำนักงานคณะกรรมการการอุดมศึกษา ปีการศึกษา 2555, กรุงเทพฯ : 2557. Office of the Higher Education Commission, *Foreign Students Enrolled in Higher Education Institutions Regulated by the Higher Education Commission, Academic Year 2012*, Bangkok : 2014.

sont proposés aux étudiants en Grande-Bretagne (à la SOAS[20] et à l'Université de Leeds). Aux États-Unis, une vingtaine d'institutions d'enseignement universitaire proposent des cursus de thaï. Dix d'entre elles sont membres du *SEASSI Consortium*[21] qui (depuis 1983) permet à ses étudiants d'étudier une langue de l'Asie du Sud-est de façon intensive pendant l'été afin d'obtenir un an de crédit d'études universitaires. L'évolution récente des effectifs suit la même tendance que les inscriptions en Thaïlande : en 2013, il y avait 286 étudiants de thaï dans des institutions universitaires aux États-Unis, une légère baisse par rapport aux années précédentes (317 en 2009 et 307 en 2006)[22]. Néanmoins, si l'on regarde les tendances à long terme, les inscriptions en thaï ont augmenté de 302,8% depuis 1974. L'engouement pour les études thaïes serait en augmentation en Russie (Koldunova, 2014), où elles sont enseignées depuis 1954, mais les effectifs sont plus importants en Asie, notamment en Chine Populaire, où quatorze institutions d'enseignement supérieur proposent des programmes d'études thaïes (Luo, 2012).[23] S'il est difficile de quantifier exactement le nombre d'étudiants de thaï en dehors de la Thaïlande, toutefois, il est clair que les effectifs restent relativement modestes par rapport à d'autres langues.

Nous pouvons donc caractériser l'enseignement de thaï langue étrangère comme bien ancré, car, outre sa longévité, les effectifs semblent être en augmentation sur le long terme, en dépit de l'instabilité politique du pays. Quant à l'origine des étudiants, historiquement occidentale, elle est de plus en plus asiatique, voire chinoise. Jantharat (2012) constate que cette tendance incite les éditeurs de manuels de thaï langue étrangère à publier des matériels pédagogiques dans des langues autres que l'anglais.

## 2.3. Caractéristiques générales du thaï

D'un point de vue typologique, le thaï est considéré comme une langue *isolante*, ou *analytique*. En d'autres termes, les unités lexicales sont invariables et la catégorie lexicale est définie par l'ordre des unités dans la phrase. La syntaxe de la phrase est nécessairement rigide.

### 2.3.1. Caractéristiques phonologiques

La phonologie du thaï comporte 18 phonèmes vocaliques, 21 phonèmes consonantiques en position initiale de syllabe et 9 en position finale, plus des diphtongues. L'opposition voyelle longue/voyelle courte est pertinente, ainsi que l'opposition consonne aspirée/non-aspirée. C'est une langue tonale avec cinq tonèmes, que nous transcrivons ainsi :

[20] L'École des études orientales et africaines, l'Université de Londres

[21] Il s'agit du *Southeast Asian Studies Summer Institute*. Les membres actuels sont : Cornell University, Michigan State University, Northern Illinois University, Ohio University, University of California-Berkeley, University of California-Los Angeles, University of Hawaii-Manoa, University of Michigan, University of Washington et University of Wisconsin-Madison.

[22] *Enrollments in Languages Other Than English in United States Institutions of Higher Education*, 2013. Modern Language Association. http://www.mla.org/pdf/2013_enrollment_survey.pdf

[23] Parmi les institutions asiatiques, on note la présence de cursus approfondis en Chine populaire (à l'Université de Pékin, à l'Université des langues étrangères de Pékin, à l'Université des minorités du Guangxi, à l'Université des minorités du Yunnan, à l'Université des études internationales de Shanghai, et à l'Université chinoise de Hong Kong, entre autres), au Japon (Tokyo University of Foreign Studies) et à Singapour (National University of Singapore).

| | |
|---|---|
| ton égal moyen | *pas de diacritique* |
| ton bas | ◌̀ |
| ton descendant | ◌̂ |
| ton haut | ◌́ |
| ton montant | ◌̌ |

### 2.3.2. Le lexique

La même unité lexicale peut appartenir à des catégories lexicales différentes selon la position de l'unité dans la phrase. Prenons par exemple le lexème ได้ /dâj/ *obtenir*, qui s'est grammaticalisé et prend des sens différents selon sa position :

(1) **ได้**ไป
/dâj paj/
*ACCOMPLI aller*
*Il est allé.*

(2) กิน**ได้**
/kin dâj/
*manger POSSIBILITÉ*
*Il se peut qu'il mange.* / *Il peut manger.*

Toutefois, le seul changement de position d'un élément ne suffit pas nécessairement pour changer sa catégorie, car en thaï il existe certains préfixes qui peuvent la modifier. Par exemple, le verbe ศึกษา /sùksǎ:/ (*étudier*) peut se transformer en การศึกษา /ka:nsùksǎ:/ (*éducation*) ou bien นักศึกษา /náksùksǎ:/ (*étudiant*).

Une grande partie du lexique de base est composée de monosyllabes, bon nombre des unités lexicales polysyllabiques étant des emprunts ou des composés. Le lexique du thaï s'est enrichi des emprunts de sources divers, particulièrement du pâli et du sanskrit, mais aussi du khmer, des dialectes chinois, et plus récemment de l'anglais. Ce sont les emprunts qui constituent une grande partie des lexèmes simples polysyllabiques, lexèmes dont les composants ne contribuent pas au sens.

#### 2.3.2.1. Les mots composés

Les lexèmes composés sont formés à partir de lexèmes monosyllabiques thaïs, en combinant des emprunts ou une combinaison des deux. Les composés formés à partir de lexèmes sanskrits ou pâlis ont leurs propres règles de combinaison (Kosawat 2003). Hoonchamlong (2013) identifie quatre sortes de mots composés :

1. Les composés réguliers, comme รถไฟ /rótfaj/ *train*, composé de รถ /rót/ *véhicule* et ไฟ /faj/ *feu.*
2. Les doublets sémantiques, comme บ้านเรือน /bâ:nrɯ:an/ *maison*, composé de บ้าน /bâ:n/ *maison* et เรือน /rɯ:an/ *maison*.
3. Les doublets euphoniques, dont la deuxième syllabe dépourvue de sens n'a qu'une fonction phonétique, comme dans มากมาย /mâ:k ma:j/ *beaucoup*

4. Les expressions complexes, souvent à quatre syllabes, et formées sur un composé, comme pour l'expression เจ็บไข้ได้ป่วย /tɕèpkʰâjdâjpùaj/ *tomber malade*. Cette expression est composée de เจ็บ /tɕèp/ *douleur*, ไข้ /kʰâj/ *fièvre*, ได้ /dâj/ *obtenir* et ป่วย /pùaj/ *malade*, et basée sur le composé เจ็บป่วย /tɕèppùaj/ *être malade*.

Parmi ces composés, se trouvent les termes génériques. Nous avons déjà vu l'exemple de นักศึกษา /náksɯ̀ksăː/ (étudiant). Le terme générique นัก /nák/ s'utilise pour créer les noms de professions, des « experts » dans un domaine donné. Par exemple :

(3) นักฟุตบอลล์ (/nákfútbɔːn/ footballeur) นัก + ฟุตบอลล์ (football)

(4) นักโทษ (/nákt$^{h}$ôːt/ prisonnier) นัก + โทษ (punir)

(5) นักเขียน (/nákk$^{h}$ɨ̌ːan/ écrivain) นัก + เขียน (écrire)

Les termes génériques sont utilisés pour désigner les noms de lieux. Par exemple ประเทศฝรั่งเศส /pràʔt$^{h}$êːtfàràŋsèːt/ est composé de ประเทศ /pràʔt$^{h}$êːt/ (pays) et ฝรั่งเศส /fàràŋsèːt/ (France, français).

Notons deux procédés de formation de lexèmes abstraits au moyen des préfixes ความ /k$^{h}$waːm/ (+VERBE pour former des noms abstraits) et การ /kaːn/ (+NOM *affaires de* ou +VERBE *acte de*) :

(6) ความ /k$^{h}$waːm/ + รัก /rák/ (*aimer*) ⇒ ความรัก /k$^{h}$waːm rák/ *l'amour*

(7) ความ /k$^{h}$waːm/ + คิด /k$^{h}$ít/ (*penser*) ⇒ ความคิด /k$^{h}$waːm k$^{h}$ít/ *une idée*

(8) การ /kaːn/ + เมือง /mɯːaŋ/ (*pays, ville*) ⇒ การเมือง /kaːnmɯːaŋ/ *la politique*

(9) การ /kaːn/ + เงิน /ŋɤ̄ːn/ (*argent*) ⇒ การเงิน /kaːnŋɤ̄ːn/ *la finance*

(10) การ /kaːn/ + ศึกษา /sɯ̀ksăː/ (*étudier*) ⇒ การศึกษา /kaːnsɯ̀ksăː/ *l'éducation*

Dans la langue parlée le préfixe การ /kaːn/ est habituellement omis.

Dans le cas de composés, il est parfois difficile de trancher s'il s'agit d'une unité lexicale ou deux, non seulement pour les apprenants, mais aussi pour les natifs (voir la section 3.4 sur la segmentation lexicale en thaï).

#### 2.3.2.2. Les classificateurs

Le thaï, tout comme le chinois et le japonais, utilise des classificateurs pour spécifier les substantifs quand ils sont accompagnés de démonstratifs ou de numéraux (sauf sans le cas où le substantif est son propre classificateur).

(11) คอมพิวเตอร์สองเครื่อง
/k$^{h}$ɔːmp$^{h}$iwtɤː sɔ̌ːŋ k$^{h}$rɯ̂ːaŋ/
*ordinateur deux CL*
*deux ordinateurs*

(12) แมวตัวนี้
/mɛːw tuːa níː/
*chat* CL *ce*
*ce chat*

Ce dernier exemple illustre aussi la position du déterminant. Le déterminant (นี้ /níː/ *ce*) suit le déterminé (แมว /mɛːw/ *chat*).

### 2.3.3. Niveaux de langue

Un système sophistiqué de pronoms personnels (ou plutôt de *référents personnels*) et différents niveaux de langue reflètent les relations entre locuteurs et leurs places respectives dans la société. Le même verbe, par exemple, peut s'exprimer par une gamme de mots, du plus vulgaire au très poli. Voici des traductions de *manger*[24] :

| | | |
|---|---|---|
| (13) | ยัดห่า /játhàː/ | (niveau le plus vulgaire) |
| | สวาปาน /sàwaːpaːn | (niveau vulgaire péjoratif) |
| | แดก /dɛ̀ːk/ | (niveau familier - *bouffer*) |
| | กิน /kin/ | *manger* |
| | ทาน /tʰaːn/ | (niveau poli - *se nourrir*) |
| | รับประทาน /rápprà?tʰaːn/ | (niveau soutenu - *se sustenter*) |
| | ฉัน /tɕʰǎn/ | (niveau soutenu, vocabulaire religieux) |
| | เสวย /sàwɤ̌ːj/ | (niveau soutenu, vocabulaire royal) |

Les deux derniers termes relèvent des vocabulaires spécifiques aux moines et à la famille royale respectivement. Ces deux sphères de la société thaïe possèdent des verbes et des substantifs propres qu'il convient d'employer quand on fait référence à des personnes qui relèvent de ces catégories. Par exemple, le mot pour désigner le chat du roi วิเชาร์ /wíʔlaː/ est différent du mot qui désigne les autres chats แมว /mɛːw/.

## 2.4. Le système d'écriture

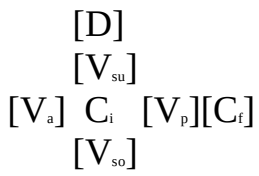
La première attestation d'une écriture proprement thaïe date du XIIIe siècle (Fels, 1993, p18) et selon les études de Ferlus (1999) qui ajoute un composant phonétique aux études épigraphiques, l'existence d'une écriture thaïe antérieure n'est pas à exclure. Depuis ces premières écritures basées sur une forme d'ancienne écriture khmère, le système d'écriture du thaï s'est modifié, pour d'abord être unifié sous le règne du roi Narai (en 1680), puis plus largement répandu quand le thaï s'est imposé comme langue d'instruction pendant la réforme des monastères de 1898 sous le règne du roi Chulalongkorn. À part quelques modifications mineures, c'est le système d'écriture du roi Narai qui est utilisé de nos jours (Dānwiwat, 1987).

Le système d'écriture thaïe, tout comme les autres systèmes d'écriture de l'Asie du Sud-est dérivés de l'écriture brāhmī indien, est un *alphasyllabaire* ou *abugida* (appellation dérivée du mot éthiopien classique pour alphasyllabaire). La phrase se lit de gauche à droite, syllabe par syllabe, mais les signes accompagnant la consonne initiale pour former une syllabe sont placés à gauche, à droite, en dessous et au-dessus de cette consonne.

[24] Exemple pris du cours de Gilles Delouche donné à l'INaLCO intitulé *Vocabulaire spécifique du siamois*.

### 2.4.1. La syllabe thaïe

La consonne initiale qui constitue le noyau de la syllabe est dotée d'une voyelle inhérente non écrite. Les autres signes ont la fonction de remplacer (pour les signes vocaliques) ou modifier (pour les signes indiquant les tons) la voyelle inhérente. Le schéma ci-dessous représente la forme de base de la syllabe écrite en thaï.




$C_i$ signe(s) consonantique(s) initial(aux)
$C_f$ signe(s) consonantique(s) final(s)
$V_a$ signe vocalique antéposé
$V_p$ signe(s) vocalique(s) postposé(s)
$V_{su}$ signe vocalique suscrit
$V_{so}$ signe vocalique souscrit
D signe diacritique (marque de ton)

*Figure 1 La forme de base de la syllabe écrite en thaï (d'après Kosawat, 2003) Seule la consonne initiale est obligatoire, sont facultatifs les signes entre [ ]*

### 2.4.2. Les signes consonantiques

Chacune des quarante-quatre signes consonantiques sont toujours citées avec la voyelle inhérente /ɔː/. Par exemple, la consonne ม seule se prononce /mɔː/.

La voyelle inhérente d'une consonne en fin de syllabe est toujours muette. Exemple (la voyelle muette en question est en gras dans l'exemple) :

(14) ร /rɔː/ + ◌ั /a/ + บ /bɔ**ː**/ → รับ /ráp/ (*recevoir*)

Deux consonnes ensemble peuvent aussi supprimer la voyelle inhérente de la première:

(15) ก /kɔː/ + ล /lɔː/ + า /aː/ + ง /ŋɔː/ → กลาง /klaːŋ/ (*milieu*)

Une consonne écrite est muette si le signe ◌์ (appelé ทัณฑฆาต /tʰantʰákʰâːt/ ou การันต์ /kaːran/) est placé au-dessus, mais ce signe n'est pas présent dans tous les cas de consonnes muettes (en **gras** dans les exemples) :

(16) ม /mɔː/ + เ-า /aːw/+ ◌้ (ton2) + ส /sɔː/ + ◌์ → เม้าส์ /máːw/ (*souris d'ordinateur*)

(17) พ /pʰɔː/ + เ /e/ + ช /tɕʰɔː/ + ร /rɔː/ → เพชร /pʰét/ (*diamant*)

Les groupes de consonnes en position initiale sont possibles, mais elles n'ont pas toujours une correspondance directe phonème-graphème. À titre d'exemple, la combinaison ทร (ท /tʰɔː/ + ร /rɔː/) peut se prononcer /s/, comme dans ทราบ /sâːp/ (*savoir*).

### 2.4.3. Les signes vocaliques

Les 29 voyelles et diphtongues sont représentées par des dix-huit signes, seuls ou en combinaison : seize signes vocaliques et deux signes consonantiques (อ /ɔː/ et ย /yɔː/). Les exemples qui suivent illustrent le fonctionnement de ces combinaisons :

Remplacement de la voyelle inhérente :

(18) ม /mɔː/ + ◌ี /iː/ → มี /miː/ (*avoir*)

(19) ม /mɔː/ + ◌ิ /i/ → มิ /míʔ/ (négation soutenue)

Modification du ton d'une même syllabe avec des marques de ton :

(20) ขาว /kʰăːw/ (*être blanc*) → ton montant

(21) ข่าว /kʰàːw/ (*nouvelles*) → ton bas

(22) ข้าว /kʰâːw/ (*riz*) → ton descendant

Signes disposés autour d'une consonne pour écrire une diphtongue :

(23) บ /bɔː/ + เ ีย /iːa/ + ร์ /rɔː/ (muet) → เบียร์ /biːa/ (*bière*)

(24) ข /kʰɔ̌ː/ + เ า /aw/ → เขา /kʰăw/ (*il*)

Les signes utilisés ensemble n'ont pas la même valeur que quand ils apparaissent seuls. Ainsi la diphtongue de l'exemple (21) เ ีย /iːa/ est composée de เ /e/, ี /iː/ et la consonne ย /yɔː/. De même, la diphtongue de l'exemple (22) เ า /aw/ se décompose en เ /e/ et า /aː/.

### 2.4.4. Les tons

La complexité du système d'indication des tons constitue un obstacle majeur pour le débutant en lecture de thaï. Afin de bien prononcer une syllabe, l'apprenant doit reconnaître la « catégorie » de la consonne initiale et identifier la « nature » de la syllabe pour pouvoir identifier la valeur de la marque de ton ou son absence. Pour résumer le système de base[25], le ton à prononcer est indiqué par une combinaison de plusieurs facteurs:

- La classe de la consonne. Chaque consonne est considérée comme appartenant à une de trois catégories: celles des consonnes *hautes*, *moyennes* ou *basses*. La voyelle inhérente des consonnes moyennes et basses ont un ton inhérent égal moyen, celle des hautes un ton inhérent montant.
- La présence ou absence de marque de ton : ◌่, ◌้, ◌๊ ou ◌๋. Les marques de ton se placent au-dessus de la consonne initiale de la syllabe, ou si la consonne est dotée d'une voyelle suscrite, au-dessus de cette voyelle.

S'il n'y a pas de marque de ton, il faut prendre en compte:

- La nature de la syllabe. Chaque syllabe est classifiée *morte* (ou *fermée*), si elle se termine par une occlusive /p/, /t/, /k/ ou le coup de glotte /ʔ/, ou *vivante* (ou *ouverte*) le cas échéant.
- La longueur de la voyelle. Pour une syllabe *morte* avec une consonne initiale *basse*, il faut prendre en compte s'il s'agit d'une voyelle longue ou courte.

S'il y a présence d'une marque de ton, il faut l'interpréter, car les marques de ton n'ont pas la même valeur pour les syllabes à la consonne initiale basse que pour les autres consonnes. Les exemples suivants portent tous les deux la marque de ton 2 ◌้ mais sont prononcés avec des tons différents :

Consonne haute ข /kʰɔ̌ː/

(25) ข้า /kʰâː/ (*je, esclave, domestique*) → ton descendant

[25] Voir Pooput & Conjeaud (2010) *Pratique du thaï Volume 1* pour un tableau schématique des tons.

Consonne basse ค /kʰɔː/

(26) ค้า /kʰáː/ (*commerce*) → ton haut (emphatique)

Nous constatons à partir de ces exemples que le phonème consonantique /kʰ/ s'écrit avec des lettres de classes différentes selon le ton de la syllabe. Tous les phonèmes consonantiques ne sont pas représentés dans toutes les classes de consonnes. Pour modifier la classe des consonnes ne possédant pas d'équivalent dans les autres classes, deux lettres (ห /hɔː/, une haute, et อ /ʔɔː/, une moyenne, muette en l'occurrence) sont antéposées à la consonne. Par exemple, la consonne น est une basse, mais l'ajout de la haute ห, lui donne la qualité haute et au lieu du ton égal moyen, l'ajout de la consonne initiale ห muette permet d'écrire une syllabe avec น /nɔː/ à l'initiale avec un ton montant.

Sans ห muette à l'initiale, น considérée comme une basse :

(27) นา /naː/ (*rizière*) Le ton inhérent est le ton égal moyen.

Avec ห muette à l'initiale, น compte comme une haute :

(28) หนา /nǎː/ (*épais*) Le ton inhérent est le ton montant.

Prenant en compte toutes les combinaisons que nous avons vues, il y a quinze types de syllabes possibles.

### 2.4.5. Règles de prononciation particulières

À ce système de base, il faut aussi compter avec des mots d'emprunt qui ont parfois des règles de prononciation particulières. Par exemple, dans เสมียน /sàmǐːan/ (*commis, greffier*), mot d'origine khmère, un /a/ court est inséré entre les deux consonnes initiales ส et ม, et la diphtongue qui les entoure เ-ียน /ǐːa/ prend le ton de la consonne initiale haute ส /sɔ̌ː/, mais elle est prononcée après la deuxième consonne ม. Parfois des mots étrangers sont prononcés avec un ton qui n'est pas écrit.

### 2.4.6. Signes supplémentaires

Outre ces jeux de signes consonantiques et vocaliques existent d'autres signes spécifiques au thaï, dont les plus courants sont :

| | | |
|---|---|---|
| ฯ | appelé ไปยาลน้อย /pajjaːnnɔ́ːj/ | symbole d'abréviation par convention[26] |
| ๆ | appelé ไม้ยมก /máːjjámók/ | symbole de répétition |
| ฿ | | symbole monétaire pour le baht |

ฯ est utilisé en combinaison avec ล pour créer ฯลฯ (ไปยาลใหญ่ /pajjaːnjàj/) qui veut dire *et cetera*.

Certains signes de ponctuation ne sont plus utilisés :

| | | |
|---|---|---|
| ๏ | โคมูตร /kʰoːmûːt/ | symbole de fin d'une histoire |
| ๚ | ฟองมัน /fɔːŋman/ ou ตาไก่ /taːkàj/ | début de ligne, paragraphe ou strophe |

[26] L'usage le plus courant de ce symbole est pour abréger le nom de la capitale de la Thaïlande, Bangkok, en กรุงเทพฯ /kruŋtʰêːp/ car le nom complet le voici est d'une longueur impressionnante : กรุงเทพมหานคร อมรรัตนโกสินทร์ มหินทรายุธยา มหาดิลกภพ นพรัตนราชธานีบูรีรมย์ อุดมราชนิเวศน์มหาสถาน อมรพิมานอวตารสถิต สักกะทัตติยวิษณุกรรมประสิทธิ์ /kruŋtʰêːpmáhǎːnkʰɔːn ʔamonrátánákoːsĭn máhǐntʰáraːjútajaː máhǎːdilokpʰòp nóppʰárátrâːtɕʰátaːniːbuːriːrom ʔùtomárâːtɕʰáníʔwêːtmáhàːsatʰǎːn ʔamonpʰíʔmaːnʔàwátaːnsàtʰìt sàkkàʔtʰáttijáʔwítsanúkamráʔsìt/

๏ อังคั่นคู่ /ʔaŋkʰânkʰûː/ fin de section ou histoire, fin de strophe

Le même symbole ๏ que nous avons vu ci-dessus (appelé ไปยาลน้อย /pajjaːnnɔ́ːj/) était utilisé autrefois pour marquer la fin d'une phrase ou d'une strophe (dans ce cas, il est appelé อังคั่นเดี่ยว /ʔaŋkʰândìːaw/, คั่นเดี่ยว /kʰândìːaw/ ou ขั้นเดี่ยว /kʰândìːaw/).

On constate aussi l'emploi de la ponctuation occidentale (guillemets, parenthèses, virgules, points d'interrogation…), le point étant utilisé essentiellement pour des abréviations.

À ceux-ci s'ajoutent les chiffres traditionnels (de 0 à 9) : ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙. Les chiffres arabes sont aussi utilisés.

Pour une explication plus détaillée en français du système d'écriture thaï, voir Conjeaud & Pooput (2010) et Brown (1991).

### 2.4.7. Conventions typographiques

Tous les systèmes d'écriture de l'Asie du Sud-est qui utilisent des alphasyllabaires, ainsi que le chinois et le japonais, sont écrits en continu sans séparation entre les mots (ce style s'appelle *scriptio continua*). Entre ces systèmes, le thaï moderne a la particularité de ne pas posséder de signe de ponctuation spécifique à la délimitation de la phrase. Il n'y a pas non plus de distinction majuscule/minuscule qui permettrait d'identifier facilement le début de phrase et les entités nommées. Pour le traitement automatique de ce style de texte, une étape de division de l'énoncé en unités lexicales est nécessaire avant tout autre traitement, par le biais d'un outil de segmentation.

Pour illustrer ce style d'écriture, voici en bas un exemple de paragraphe en thaï, accompagné de sa traduction en anglais[27]. On note le nombre d'unités lexicales apparentes en anglais et le peu d'espaces en thaï.

| | |
|---|---|
| วันหนึ่ง เขาตื่นลืมตาขึ้นมาจากความฝันอันสับสน รู้สึกอ่อน เพลีย จากการพักผ่อนไม่เพียงพอ เขาเดินโซเซไปชนกองหนังสือในห้องรับแขกที่ตั้งไว้สูงจนเป็นกำแพงกระดาษ หนังสือเล่มใหญ่เกือบสิบเล่มหล่นลงมาทับเท้าทำเอาเขาร้องโอดโอย แล้วเซไปเตะซีดีที่กองไว้อีกมุมห้องจนกระจัดกระจาย ต้องเดินกะเผลกๆ ผ่านห้องหับอันมัวซัวไปเข้าห้องน้ำล้างหน้าล้างตา | One day, he woke up from a confused dream, feeling weak from not enough rest. Walking un-steadily he bumped into a pile of books in the living room so high it formed a wall of paper. More than half a dozen big tomes fell on his feet, making him cry out in pain. Then as he staggered he kicked a pile of CDs in a corner of the room, scattering them. He had to limp past a dim room to enter the bathroom to wash his face. |

*Figure 2 Échantillon de texte en thaï avec sa traduction en anglais*

Les espaces dans les textes en thaï ont d'autres fonctions que la séparation des unités lexicales. Une œuvre de référence publiée par la *Société Royale de Thaïlande* donne des principes d'espacement, utiles à comprendre pour cerner la problématique de la définition de l'unité lexicale et de la phrase en thaï.

En principe, il a deux types d'espace: l'espace simple (qui a la largeur de la lettre ก) et l'espace double (deux fois plus large), cette dernière servant uniquement pour délimiter la phrase. La distinction entre ces deux espaces semble assez méconnue. En effet, des articles sur

[27] Il s'agit du premier paragraphe de la nouvelle ของ /kʰɔ̌ːŋ/ *Les choses*, de l'auteur ตะเฆ่สัน /tàʔkʰêʔsăn/ (nom de plume de Vip Burapadecha) traduit par Marcel Barang. Texte et traduction tirés du blog du traducteur https://thaifiction.wordpress.com/ apparu pour la première fois dans Chor Karrakeit 55, 2011.

l'identification automatique de la phrase en thaï n'en font pas mention. Selon Aroonmanakun (2007) la phrase « n'a pas de délimiteurs explicites » et pour Sornlertlamvanich (2000), « il n'existe pas de marqueur de fin de phrase explicite ». Elle est parfois respectée dans le monde de l'édition, mais l'utilisation de l'espace simple pour terminer une phrase n'est pas considérée comme une faute.

Par ailleurs, dans les logiciels d'édition et de traitement de texte varient la valeur de l'espace varie pour les polices à chasse variable, rendant la différence entre l'espace simple et l'espace double difficile à discerner. Nous n'avons pas trouvé d'emploi systématique de l'espace double dans nos corpus.

L'espace simple, à part son utilisation courante comme délimiteur de fin de phrase, a de nombreuses fonctions. Une liste de règles d'utilisation de l'espace simple, publiée par la *Société Royale* figure en annexe. Entre autres, elle sert à indiquer le début de certaines propositions coordonnées, distinguer les noms propres du texte environnant, les éléments de la date et l'heure, les éléments d'une liste et indiquer le début d'une citation directe.

Le guide de la *Société Royale* dit que ce ne sont que des recommandations; dans l'usage, les Thaïs disposent d'une certaine latitude pour insérer ou non l'espace dans leurs écrits selon ce qu'ils estiment élégant ou utile. L'emploi de l'espace typographique peut donner des pistes pour améliorer les résultats de la segmentation automatique des textes.

### 2.4.8. Difficultés du système d'écriture pour l'étudiant en thaï langue étrangère

Le système d'écriture que nous venons de brièvement décrire présente des difficultés spécifiques pour l'apprenant en L2. Malgré un niveau de correspondance graphème-phonème relativement élevé par rapport à d'autres systèmes d'écriture, l'orthographe du thaï est considérée comme une orthographe complexe (Kasisopa, 2011), en vertu du fait que le nombre de phonèmes est considérablement plus restreint que le nombre de graphèmes. En effet, Berment (p135, 2004) calcule qu'il y a plus d'un million de syllabes possibles avec ce système d'écriture. L'étudiant identifie rapidement deux des principales difficultés comme l'écriture en continu et la gymnastique mentale requise pour déchiffrer la notation des tons. En réalité, l'apprentissage de la lecture en thaï est encore plus complexe, car l'étudiant (surtout celui habitué aux systèmes d'écriture latine) est confronté à un texte qui lui fournit peu de repères familiers – il n'y a ni distinction entre minuscules et majuscules pour indiquer le début d'une phrase ou des entités nommées, ni ponctuation pour indiquer la fin de syntagme (virgule) ou de phrase (point, ou point d'interrogation), par exemple.

Hoonchamlong (2013) identifie deux types de difficultés en lecture de thaï en L2. Des difficultés de bas niveau : de déchiffrage des graphèmes et l'identification de frontières de syllabes; et des difficultés de haut niveau : d'identification de frontières de mots et de sens des mots, de déchiffrage de l'organisation du texte (syntagmes, phrases) et l'identification d'ellipses. Elle explique que dans d'autres langues, la ponctuation donne des informations concernant l'organisation du texte, alors qu'en thaï ce sont des connecteurs (comme des conjonctions) et certaines unités lexicales qui ont cette fonction. Ce qui est système d'écriture dans une langue devient syntaxe ou lexique dans l'autre.

En ce qui concerne l'acquisition des compétences de lecture de bas niveau, c'est l'identification des tons et de frontières de mots qui posent problème. L'étudiant doit d'abord s'entraîner à identifier la syllabe avant d'identifier le mot, mais le nombre considérable de

combinaisons possibles peut décourager le débutant. Pour faciliter la tâche, les écoliers thaïs commencent l'apprentissage de la lecture avec des textes segmentés en mots par des espaces. Certains manuels de thaï langue étrangère proposent des textes dans ce format (Hoonchamlong, 2007a, 2007b, par exemple) et certains outils électroniques d'aide à la lecture (voir la section 3.1.2) proposent une segmentation des textes en mots (et parfois aussi en syllabes).

Cette introduction en douceur à la *scriptio continua* se justifie par divers travaux de recherche. Kasisopa (2011) note que pour les langues qui séparent les unités lexicales, la lecture par les locuteurs natifs de textes sans espaces devient environ 35% moins rapide. Une différence est aussi perceptible pour les locuteurs natifs du thaï lisant des textes segmentés et non segmentés, mais, à la différence des non-Thaïs, il n'y a pas de différence perceptible dans leurs mouvements oculaires, les natifs étant capables d'identifier le noyau du mot avec ou sans espaces dans le texte. Ceci s'explique par le fait que le lecteur natif possède d'autres connaissances qui lui permettent d'anticiper ce qui va venir.

Ce sont ces connaissances que le lecteur en L2 doit acquérir pour pallier les difficultés de haut niveau. Il doit pouvoir interpréter l'organisation du texte en phrases et syntagmes par le biais des espaces, connecteurs (qui peuvent se composer de deux termes séparés « comme des lunettes ») et l'identification de mots clés comme le verbe principal. Ensuite, il faut développer des connaissances en morphologie des mots composés pour pouvoir anticiper les frontières des mots.

Le lecteur en L2, à la différence du lecteur en L1, doit développer des stratégies d'identification de frontières de mots inconnus. Il nous semble donc, très utile de créer un système qui introduit du vocabulaire nouveau progressivement, non seulement pour que l'étudiant puisse deviner les sens des mots à partir du contexte, mais aussi pour lui permettre de pratiquer ces stratégies d'identification de frontières de mots inconnus confortablement.

### 2.4.9. L'encodage du thaï en Unicode

Le bloc Unicode pour le thaï s'étend de U+0E00 à U+0E7F. La norme Unicode pour le thaï est basée sur la norme *Thai Industrial Standard 620-2529*, et sa version actualisée 620-2533. Ce même bloc est aussi utilisé pour l'écriture en lettres thaïes d'autres langues telles le pâli et le sanskrit, quelques signes étant spécifiques à ces langues.

Si la représentation en Unicode de la plupart des alphasyllabaires dérivés du brāhmī indien préconise le stockage en ordre logique, le thaï, comme le laotien, est stocké dans un ordre visuel (Gillam, 2003, p611). Plus précisément, le principe de stockage logique veut que la consonne initiale soit toujours stockée avant la voyelle à laquelle elle est reliée, quelle que soit la position des signes vocaliques. Cela veut dire que l'ordre de frappe sur le clavier est identique à l'ordre de prononciation. Le principe d'ordre de stockage visuel qui utilisé pour le thaï ne tient pas compte de la prononciation ; les signes sont stockés dans leur ordre d'apparition. Si une voyelle apparaît avant sa consonne initiale, elle est stockée en mémoire avant la consonne (en **gras** dans les exemples suivants).

Exemple thaï de stockage selon l'ordre visuel :

| | | | | | |
|---|---|---|---|---|---|
| (29) | แมว | /mɛ:w/ | *chat* | | |
| | | est stocké | แ | ม | ว |
| | | | 0E41 | **0E21** | 0E27 |

Exemple birman de stockage selon l'ordre logique :

(30) ခြင် /kyaung/ *chat*

| est stocké | ခ | ြ | ေ | ာ | င | ် |
|---|---|---|---|---|---|---|
| | **1000** | 103C | 1031 | 102C | 1004 | 103A |

Simple en apparence, le choix de l'ordre visuel de stockage complique le processus de tri, qui est toujours selon l'ordre logique. Voici une liste de vocabulaire thaï triée en ordre alphabétique :

กก
กิโล
ไก่
งก
งง
ไง

Les éléments qui commencent par ไ n'apparaissent pas l'un après l'autre, car ไ est une voyelle qui se positionne à gauche de la consonne initiale. Elle n'est prise en compte dans le tri qu'après le signe consonantique qui suit (la première lettre prononcée dans le mot).

Le principe de stockage visuel a été choisi pour ne pas compromettre l'interopérabilité entre la norme Unicode et les normes nationales déjà en vigueur, afin d'assurer l'adoption généralisée de la norme Unicode. Cet ordre de frappe est parfois déroutant pour le débutant en thaï habitué au clavier d'une autre langue dotée d'un système d'écriture similaire ou un système de saisi basé sur la prononciation (comme la méthode de saisi pinyin pour le chinois), mais paraît naturel pour les étudiants habitués à l'alphabet latin.

# 3. L'état de l'art

Cette partie concerne les travaux antérieurs pertinents pour notre projet. Tout d'abord, nous cherchons à situer notre recherche dans le contexte de l'enseignement et l'apprentissage de la lecture du thaï langue étrangère, examinant à la fois les manuels et les ressources et outils électroniques. Nous abordons ensuite les ressources électroniques pour la lecture dans d'autres langues qui proposent des textes classifiés par niveau de difficulté, terminant avec celle qui nous intéresse particulièrement, *TextLadder*.

Comme nous allons le découvrir, *TextLadder* et le projet *Candle* qui reprend la même idée ont recours à des listes lexicales élaborées pour les lecteurs de l'anglais en L2. Depuis quelques années, des chercheurs en linguistique de corpus ont publié des listes de mots pour de nombreuses langues basées sur la fréquence lexicale. Après avoir détaillé le contenu et la création de ces listes avec une attention particulière pour les listes de fréquence lexicales thaïes existantes, nous terminons cette section avec un exposé sur la segmentation en mots et les outils de segmentation lexicale des textes en thaï.

## 3.1. Ressources pédagogiques pour la lecture en thaï

### 3.1.1. Les publications visant une pratique de la lecture

La longue histoire d'enseignement de la langue thaïe à l'INaLCO a engendré la publication de nombreux manuels de thaï (Delouche, 2009a, 2009b et Conjeaud & Pooput, 2010a, 2010b), riches en matériels pédagogiques pour la lecture, accompagnés de CDs des pistes audio des textes et exercices. Ces livres, aussi utiles pour l'autodidacte que l'étudiant en salle de classe, ne visent pas exclusivement le développement de compétences en lecture. À part le livre de Brown (1990) pour l'apprentissage du système d'écriture, il y a peu d'autres ressources pédagogiques destinées aux étudiants francophones.

D'autres publications destinées aux étudiants anglophones ont été publiées aux États-Unis et en Grande-Bretagne, bien souvent pour les étudiants au niveau débutant[28]. Certaines de ces publications et ressources pédagogiques sont dédiées à l'apprentissage de la lecture, notamment le fruit du travail précurseur de Haas (1964) sur un dictionnaire destiné aux étudiants de thaï langue étrangère. Elle a publié un livre pour la lecture du thaï *Thai Reader* (1978) pour les étudiants de niveau intermédiaire, disponible sur le site *SEAsite* enrichi avec pistes audio et exercices[29]. Ce même site donne accès à des livres de lecture gradués par niveau destinés aux écoliers thaïs utilisés dans les écoles primaires de 1978 à 1994[30], aussi enrichis d'un concordancier, des pistes audio et d'un dictionnaire contextuel qui donne la traduction d'un mot lorsqu'on passe le curseur de la souris dessus[31]. Le site donne une liste de fréquence pour le vocabulaire (234 éléments) pour le premier livre de la série. Le livre de Jones et coll.

[28] Une liste de manuels de thaï langue étrangère est fournie à la fin du mémoire (voir section 11.3).

[29] Basé au Center for Southeast Asian Studies, Northern Illinois University http://www.seasite.niu.edu/Thai/thaireader/frameset.htm

[30] Lek Changply, *The Return of "Maanee Maana", a classic Thai textbook for foreigners learning Thai edition*, Chiang Mai Post, September 20-30, 2012

[31] Il s'agit des *Maanii Readers* (*มานะ มานี ปิติ ชูใจ*). Bien qu'il s'agisse de textes simples, sur le plan du contenu ils peuvent aussi être considérés comme des textes « authentiques » pour des étudiants de thaï langue étrangère, car ils n'ont pas été écrits dans le seul but d'enseigner que la lecture; l'auteur (รัชนีศรี ไพรวรรณ /rátnîːtriː pʰrajwan/) a voulu inculquer les valeurs culturelles thaïes chez les écoliers et ces livres font partie de l'expérience collective des Thaïs.

(1976), *Thai Cultural Reader Book 1*, est destiné aux étudiants de niveau intermédiaire ayant déjà lu les textes de *Thai Reader* (Haas 1978), les textes portant sur une grande variété de sujets, ne sont pas présentés dans un ordre de difficulté croissante. Le glossaire en fin de l'ouvrage contient environ 3 500 éléments. Le livre de Brown (1986) *AUA Language Center Thai Course, Reading and Writing : Mostly Reading*, est très détaillé sur l'explication du système d'écriture du thaï, mais contient aussi de nombreux textes pour la pratique de la lecture, en écriture manuscrite et d'écritures différentes. Si le déchiffrage de l'écriture manuscrite n'est plus une compétence utile, cela soulève le problème de la lisibilité inégale des différentes polices de caractères. La plupart des manuels de thaï mentionnés ici utilisent une police facile à lire, et en général les exemples de matériels pédagogiques authentiques écrits avec des polices différentes et difficiles à déchiffrer sont rares. Ceci pose toujours un problème pour l'étudiant confronté aux polices très stylisées.

### 3.1.2. Les ressources électroniques

À ces ouvrages s'ajoutent des ressources exclusivement disponibles sur le web. ***SEAsite***, mentionné ci-dessus, propose des textes[32] classés « débutant », « intermédiaire » ou « avancé », tous accompagnés d'une traduction et des pistes audio. En plus des sources déjà mentionnées, il s'agit essentiellement de textes authentiques soigneusement choisis : des paroles de chansons, des textes littéraires ou des dessins humoristiques.

Le site ***Self Study Thai***[33] propose 84 articles de la VOA[34], tous assortis d'une piste audio au format MP3, de glossaires du vocabulaire difficile et de traductions des phrases entières. Les rubriques proposées sont : USA, Asia, World, Business, Health, Science, Society, Lifestyle, Sports & Entertainment. Certains de ces articles sont disponibles au format du *Thai Text Reader* (voir plus loin). L'archive de la rubrique ***Learning Thai***[35] du quotidien thaï anglophone propose une trentaine d'articles du même format que *Self Study Thai*.

Parmi les 88 leçons de thaï du système d'apprentissage des langues étrangères développé pour le Centre des langues étrangères de l'institut de langues de la Défense[36] des États-Unis, appelé ***GLOSS***[37], se trouvent 37 leçons pour la lecture basées sur des textes authentiques de thèmes variés. Les sujets proposés portent sur la culture, l'économie, l'environnement, la géographie, la politique, la science, la société et la technologie; la particularité de *GLOSS* étant de proposer aussi des textes militaires et sur la sécurité. Des exercices autour de l'apprentissage de la lecture, du vocabulaire et des locutions sont proposés, tout comme des traductions des textes et leurs pistes audio. Les textes, destinés aux étudiants intermédiaires et avancés, sont classifiés selon l'échelle ILR[38] de 1 à 2+ (donc sur quatre niveaux).

---

[32] http://www.seasite.niu.edu/Thai/language/reading.htm
[33] http://www.selfstudythai.com/index.html
[34] VOA, Voice of America, est le service de radiodiffusion internationale officiel des États-Unis. Le service thaï est disponible ici : http://www.voathai.com/ Le site web de VOA, qui existe dans 41 langues différentes, constitue une ressource intéressante pour la création de corpus de langues peu enseignées.
[35] http://www.bangkokpost.com/learning/learning-from-news/333366/learning-thai-with-post-today-archive
[36] Le DLIFLC, Defense Language Institute - Foreign Language Center http://www.dliflc.edu/
[37] GLOSS = *Global Language Online Support System* https://gloss.dliflc.edu/
[38] L'échelle *ILR* (*Interagency Language Roundtable*) est une échelle d'évaluation standard pour mesurer les compétences linguistiques développée pour les besoins d'évaluation du personnel du gouvernement des États-

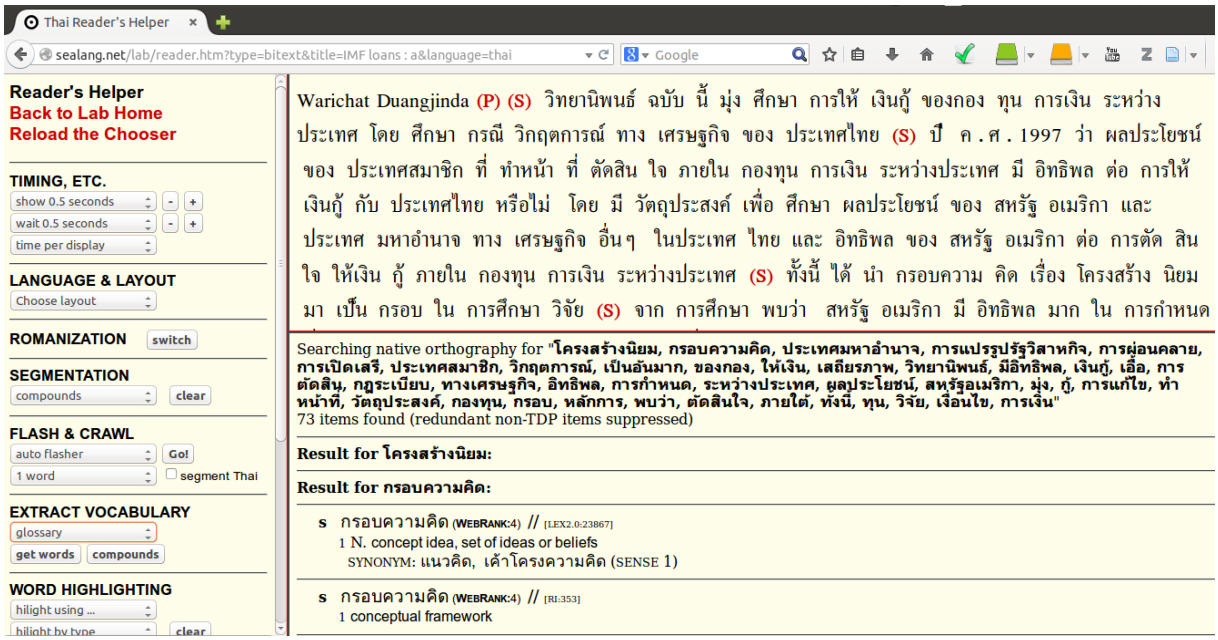



*Figure 3 SEAlang Lab Reader's Helper. Texte d'un résumé de thèse sur les prêts du FMI. Extraction automatique de mots composés et séparation automatique du texte en mots composés.*

Cette échelle ILR est aussi utilisée pour classer certains des textes du ***SEAlang***[39] ***Lab*** par niveau de difficulté, pour l'instant sur trois niveaux : 0/1, 1+/2, 2+/3. En dehors de la présentation en bitexte (texte aligné avec sa traduction), le *Lab* fournit des outils pour faciliter la lecture du thaï appelé *Reader's Helper*, tels que la segmentation (en mots, mots composés ou syntagmes), le surlignage des mots par fréquence lexicale ou la romanisation. Le contenu du *Lab* est très riche, avec des textes alignés par paragraphe ou par phrase (au choix) de textes littéraires, d'une centaine d'articles de presse et des résumés de thèse. Il reprend des textes du *Thai Reader* (Haas 1978), des textes de *LangNet*[40] (site américain d'apprentissage de langues non disponible en libre accès) et du *Thai Reader Project.* Le lecteur peut aussi extraire automatiquement des listes de vocabulaire du texte, avec ou sans des définitions tirées du dictionnaire en ligne *Thai SEAlang*[41] et rajouter ses propres annotations. Le site de *SEALang* utilise une stratégie de segmentation à base de dictionnaire qui privilégie un minimum de mots. Les erreurs inévitables de segmentation sont corrigées à la main et le dictionnaire est augmenté en continu avec de nouveaux mots[42].

Unis. L'échelle de 11 niveaux va de 0 (aucune compétence) à 5+ (locuteur natif ou bilingue). Le descriptif des niveaux est disponible ici: http://www.govtilr.org/index.htm

[39] http://sealang.net/lab/index.htm Le site *SEAlang* regroupe un nombre de projets dirigés par Doug Cooper, axés sur l'Asie du Sud-est, qui constituent des ressources pour l'apprentissage des langues et la recherche. Le financement pour le développement de ces projets provient de sources variées, essentiellement de subventions d'institutions et ministères américains. Le *SEALang Lab* a été développé en conjonction avec le *Center for Research and Language Acquisition* (CARLA) de l'Université de Minnesota qui est spécialisé dans les ressources pour les LCTLs, acronyme de *Less Commonly Taught Languages* (les langues plus rarement enseignées), désignation officielle aux États-Unis pour les langues enseignées dans l'enseignement public à part le français, l'allemand et l'espagnol.

[40] http://www.langnet.org/

[41] Une compilation de plusieurs dictionnaires de référence. Pour plus de détails, voir http://sealang.net/thai/dictionary.htm

[42] Communication personnelle de Doug Cooper.

Les bitextes de littérature du *SEAlang Lab* proviennent essentiellement du traducteur Marcel Barang, dont le blog ***เรื่องสั้นไทย | thai to english fiction***[43] (/rûːaŋsântʰaj/ *nouvelles thaïes*) regroupe une centaine de nouvelles d'auteurs divers, chacune alignée à sa traduction commentée.

Le ***Thai Reader Project***[44] est un ensemble de cours qui se concentrent uniquement sur l'acquisition des compétences liées à la lecture. Bien que disponibles exclusivement sur le web, il s'agit d'une collection de fichiers au format PDF, destinés à être imprimés et exploités dans une salle de classe. Chaque leçon est composée d'un texte authentique et d'exercices de compréhension. Les auteurs qualifient les textes d'authentiques dans la mesure où ils sont réalistes, mais il se peut que certains aient été simplifiés à des fins pédagogiques.

Dans la même veine, mais compilé par un étudiant plutôt qu'un professionnel de didactique de thaï, nous avons ***Advanced Thai Reading and Vocabulary Building***[45] deux livres téléchargeables au format PDF, chacun des cinquante cours étant composé d'un texte avec une liste de vocabulaire difficile commenté, suivit d'exercices de traduction.

Le site de ***Scola***[46] associe des journaux entiers numérisés avec la traduction automatique des articles. Le système utilise un traitement de reconnaissance optique de caractères multilingues combiné avec un traitement de traduction automatique (MacRostie et coll., 2010). L'utilisateur est averti sur le site que la traduction résultante comporte de nombreuses erreurs, et elle est fournie dans le but d'aider le lecteur à comprendre l'essentiel du texte (*gisting* en anglais) et lui permettre de faire des recherches (guidées par la traduction). Actuellement, 162 numéros de journaux thaïs (qui datent de 2010 à 2015) sont proposés, comme ไทยรัฐ /tʰajrát/ *Thai Rath*, อู่ข้าว /ʔùːkʰâːw/ *Rice Bowl*, ตลาดลูกหนัง /tàlàːtlûːknăŋ/ *Siam Sports Daily*, ASTV ผู้จัดการรายวัน /pʰûːtɕàtkaːn raːjwan/ *Manager Daily*. *Scola* fournit des matériels authentiques dans 175 langues et dialectes.

[43] https://thaifiction.wordpress.com/ Ce site de littérature mondiale qui présentait de nombreux bitextes français en traduction thaïe *Wanakam.com World Classics in Thai* n'est plus à son adresse d'origine, mais il est disponible via le moteur de recherche *Internet Archive's Wayback Machine :*
http://web.archive.org/web/20071211105624/http://www.wanakam.com/

[44] Développé par les professeurs de thaï sous la direction de Robert J. Bickner à l'Université du Wisconsin à Madison http://readingthai.wisc.edu/thai-reader-site-home.html

[45] L'auteur s'appelle Hugh Leong, son site : http://www.retire2thailand.com/retire2-reading-thai.php

[46] http://www.scola.org/Homesearch.aspx?searchterm=ewe&language=Thai Un accès temporaire gratuit au site est possible sur simple demande.

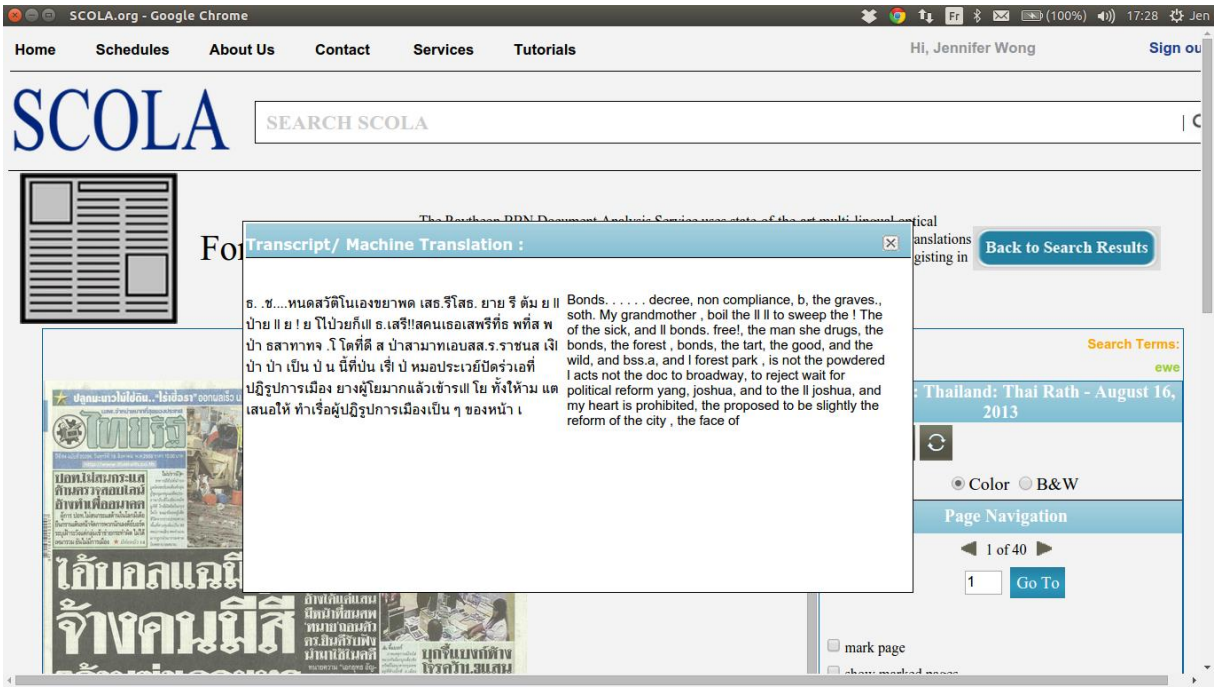



*Figure 4 Scola - Extraction automatique d'un article de presse (ici d'un numéro du journal Thai Rath) avec traduction automatique.*

***Readlang***[47], une aide à la lecture conçue pour les tablettes et smartphones aussi bien que les ordinateurs, est composée d'un site web couplé d'une extension pour les navigateurs *Chrome* ou *Safari*, appelé *Web Reader*. L'utilisateur précise sa langue maternelle et choisit la langue qu'il voudrait apprendre. Actuellement, seize langues sont disponibles – le thaï figure parmi les quarante langues en cours de développement disponibles en version bêta. *Web Reader* permet à l'utilisateur de trouver les définitions des mots dans n'importe quel site web de son choix. L'interface permet de choisir de remplacer le mot inconnu par son équivalent dans la langue maternelle, ou d'afficher la traduction au-dessus (voir la figure 3.3) avec la possibilité d'intégrer un dictionnaire web au choix. En plus de ces fonctionnalités, le site de *Readlang* propose à l'utilisateur de choisir parmi les textes existants ou de charger ses propres documents. Ainsi, le site fonctionne comme une banque de textes partagés par les utilisateurs. Pour certaines langues les textes partagés sont classés par niveau CECR[48] à l'aide d'une mesure de lisibilité (l'indice ARI[49]) et le pourcentage de mots qui apparaissent dans des listes de fréquence lexicale pour les 2000 premiers mots. Les niveaux ont été calibrés approximativement suivant le jugement du concepteur.[50]

L'utilisateur enregistré sur *Readlang* peut suivre son progrès dans l'apprentissage du vocabulaire, en utilisant des cartes créées automatiquement au fur et à mesure qu'il consulte des mots dans le dictionnaire lors de ses lectures. L'interface d'auto-interrogation à répétition

[47] http://readlang.com/landing-page

[48] CECR = Cadre européen commun de référence pour les langues http://www.coe.int/t/dg4/linguistic/Source/Framework_FR.pdf

[49] ARI = Automated Readability Index prend en compte la longueur des mots en caractères et la longueur des phrases en mots.

[50] Communication personnelle du concepteur de *Readlang*, Steve Ridout.

espacée permet également d'exporter ces cartes (y compris les phrases où les mots apparaissent) dans le format Anki[51], logiciel multiplateforme de révision de cartes-mémoire. Ces systèmes d'auto-interrogation à répétition espacée utilisent des listes de fréquence lexicale pour privilégier l'apprentissage du vocabulaire le plus fréquent.

Comme nous l'avons évoqué, la version de *Readlang* pour l'apprentissage du thaï est actuellement en version bêta. Puisqu'il n'y a pas de segmentation des mots, l'utilisateur doit faire glisser la souris sur le mot pour le sélectionner au lieu de simplement cliquer dessus.

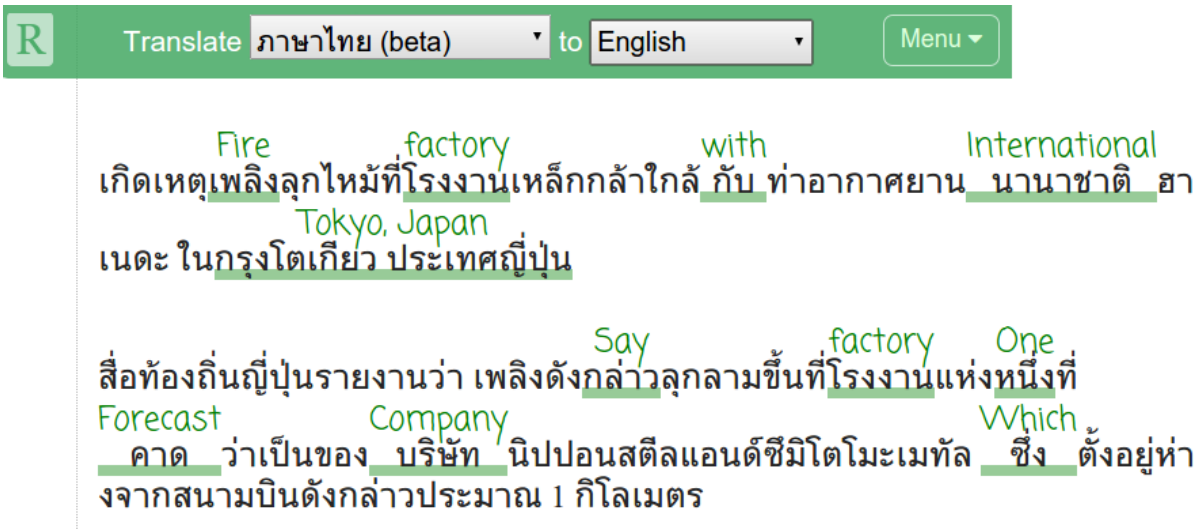



*Figure 5 Readlang avec un texte thaï. La définition du mot souligné apparaît au-dessus du mot recherché.*

Le site ***Thai2English***[52] est essentiellement un dictionnaire qui permet à l'utilisateur d'entrer du texte, qui est ensuite reproduit avec une glose qui comprend son sens en anglais, sa prononciation et parfois des précisions concernant son emploi. Le dispositif est aussi disponible en téléchargement.

***AideMoi*** est un dispositif web d'aide à la lecture en cours de développement au centre de recherche ER-TIM[53]. L'idée est de favoriser, lors de la lecture d'un texte, l'observation et l'intuition du lecteur en lui procurant d'une part des annotations morphosyntaxiques automatiques, d'autre part des fonctionnalités de navigation lexicale (occurrences, concordances, définitions,...) au sein du texte en cours de lecture ou parmi ceux déjà lus, et d'éviter au maximum le recours au dictionnaire.

Dans sa version actuelle, *AideMoi* est capable d'annoter à la volée des textes français, hindis, hongrois et thaïs, mais ne propose pas encore de navigation lexicale.

L'étiquetage morphosyntaxique des textes en thaï utilise le service REST d'étiquetage morphosyntaxique par ORCHID, développé par Poltree et Saikaew (2011)[54] (voir figure 3.4). L'équipe prévoit l'intégration de textes supplémentaires par le lecteur lui-même. L'objectif de

[51] http://ankisrs.net/
[52] http://www.thai2english.com/
[53] Équipe de Recherche Textes, Informatique, Multilinguisme http://www.er-tim.fr/
[54] http://www.thaisemantics.org/service/swath/index (lien mort)

l'équipe est de proposer une méthodologie d'intégration dans *AideMoi* d'autres langues suffisamment pourvues en outils de TAL.


*Figure 6 AideMoi en cours de développement avec un texte thaï tiré du Thai Reader Project.*

### 3.1.2.1. Logiciels

***Thai Text Reader***[55] (TTR) est une variante du *Foreign Language Text Reader*[56] (FLTR) adapté spécifiquement pour le thaï. Il comprend un outil de segmentation de thaï couplé avec un dictionnaire et des ressources de lecture préparées, tirées des *Maanii Readers*, de la rubrique *Learning Post*[57] du journal *Bangkok Post*, des textes de la *VOA* tirés du site *Self Study Thai*, mentionné ci-dessus, ainsi que des transcriptions du site *Thai Recordings*[58]. Le logiciel permet la création d'un dictionnaire personnalisé, ainsi que l'intégration de jusqu'à trois dictionnaires web. L'utilisateur peut modifier les entrées du dictionnaire personnalisé interne (dans la fenêtre *Edit Term*) dont les entrées de base sont prises des dictionnaires intégrés et il peut aussi rajouter des annotations sur ses connaissances des mots (avec une note de 1 à 5), annotations qui déterminent la couleur de la surbrillance des mots dans le texte. Le site du logiciel contient un outil[59] en ligne qui permet de préparer des textes pour la lecture avec *Thai Text Reader*, de segmenter le texte en mots, en choisissant l'espace ou l'espace invisible comme délimiteur.

[55] http://thaitextreader.com/
[56] Archive du site de développement : http://web.archive.org/web/20120805054341/http://code.google.com/p/fltr/
[57] http://www.bangkokpost.com/learning/learning-together/333366/learning-thai-with-post-today-archive
[58] http://thairecordings.com/
[59] http://thaitextreader.com/thaiparser/index.html

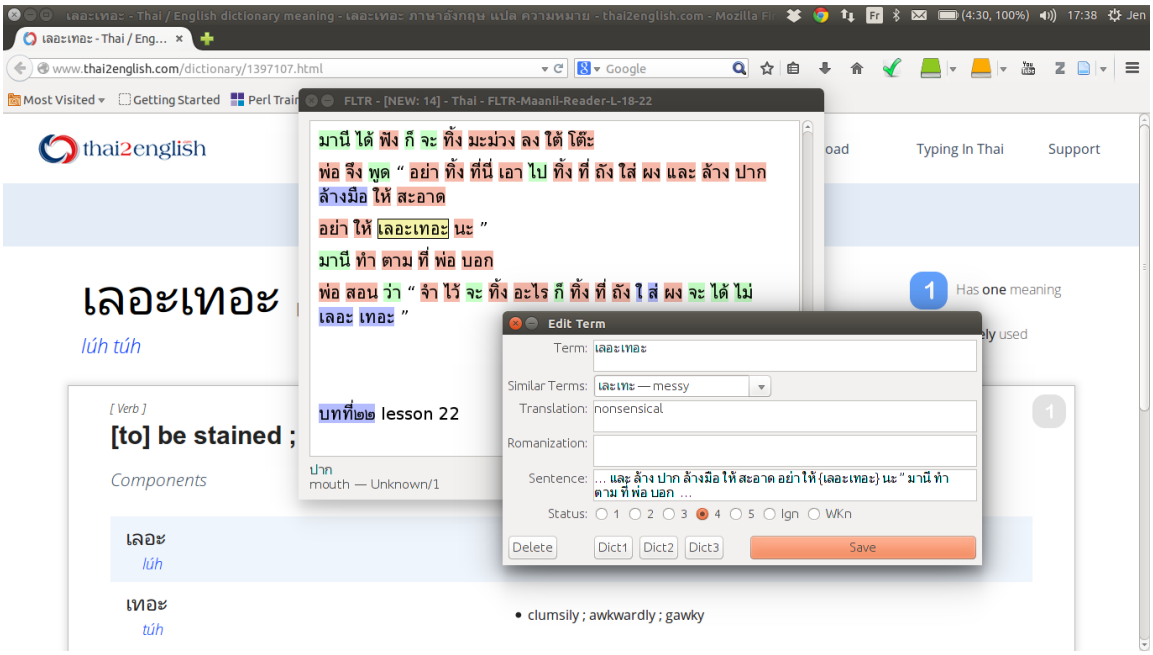



*Figure 7 Thai Text Reader avec dictionnaire web intégré (ici, Thai2English en fond d'écran). La couleur de la surbrillance diffère selon le niveau de connaissance des vocables précisé par le lecteur (options Status dans la fenêtre Edit Term, de 1 à 5).*

***Learning with Texts***[60] (LWT) fournit les mêmes fonctionnalités que *Thai Text Reader*, mais avec la possibilité d'intégrer les pistes audio. Le système a l'avantage d'avoir une interface plus conviviale, mais il est assez fastidieux à installer, car il exige l'installation d'un serveur local.

## 3.2. Ressources électroniques pour la lecture avec classification automatique de textes

### 3.2.1. REAP

***REAP***[61] est un « système intelligent de tutorat » de compréhension de la lecture et d'apprentissage lexical. Développé au sein de l'Institut des technologies linguistiques à l'Université Carnegie Mellon, le système est une aide pédagogique pour des professeurs d'anglais enseigné comme langue L1 ou L2. Le système a la particularité de prendre en compte les connaissances spécifiques de l'étudiant et son niveau d'études dans le choix des textes. Ces informations permettent la personnalisation du parcours de lecture. *REAP* a été porté en portugais[62] (voir Marujo et coll., 2009) et en français (en L1).

Après l'ouverture d'une session, le système demande à l'étudiant de préciser son niveau d'études, puis propose un test de 20 mots afin de déterminer son niveau de connaissances lexicales, avant de lui demander de remplir un questionnaire sur ses centres d'intérêt. Dix thèmes sont proposés : Culture, Entreprise, Informatique, Jeux, Santé, Maison, Loisirs,

[60] http://lwt.sourceforge.net/index.php

[61] Le projet *REAP* est sous-titré *Reader-Specific Lexical Practice for Improved Reading Comprehension*. Voir Brown & Eskenazi (2004). Site du projet : http://reap.cs.cmu.edu/

[62] Site du projet REAP.PT http://www.cmuportugal.org/tiercontent.aspx?id=1560

Sciences, Société, Sport. Ensuite, l'étudiant se voit proposer un éventail de textes authentiques issus du web, parmi ceux de la base de données *REAP* qui correspondent à ses goûts et son niveau. Le texte choisi par l'étudiant apparaît, avec la possibilité de rechercher des mots dans le dictionnaire (*Cambridge Advanced Learner's Dictionary* pour l'anglais, *le Littré* pour le français). Les mots identifiés comme mots cibles par le professeur sont soulignés. À la fin de la lecture *REAP* pose une série de questions de types différents pour évaluer l'apprentissage du vocabulaire et le niveau de l'étudiant est réévalué. *REAP* pose aussi des questions sur l'intérêt de l'étudiant pour le texte et la difficulté du texte.

*REAP* offre aux chercheurs et professeurs un moyen de suivre le progrès des étudiants. Pour chaque étudiant, *REAP* fournit des informations non seulement sur les résultats des tests, mais aussi concernant le nombre de textes lus et le nombre de mots recherchés dans le dictionnaire (et la liste de mots) dans chaque texte. Pour chaque texte, on peut voir les mots cibles. Dans les vidéos de démonstration, il y a très peu de mots cibles par texte (trois maximum).

La classification des textes par niveau de difficulté utilise une stratégie hybride, combinant modèles de langue basés sur la statistique lexicale[63] avec une classification par difficulté grammaticale modélisée sur des constructions grammaticales trouvés dans des manuels d'anglais langue étrangère de niveaux différents. La recherche sur laquelle *REAP* est basée (Heilman et coll., 2007) a démontré que les modèles de langues basés sur la statistique lexicale sont plus efficaces que l'approche qui utilise la difficulté grammaticale seule, mais une combinaison des deux est encore plus précise à la fois pour des corpus L1 et L2. Il semblerait que la difficulté grammaticale soit plus décisive pour L2 que pour L1. Ceci serait dû au fait que l'apprentissage du vocabulaire et l'apprentissage de la grammaire se déroulent en même temps en L2, alors que l'acquisition du vocabulaire continue après l'acquisition de la grammaire en L1.

### 3.2.2. TextLadder et TextGrader

***TextLadder***[64] est un logiciel conçu par Ghadirian (2002) qui classe des textes en anglais dans un ordre de lecture afin de faciliter l'acquisition du vocabulaire de manière progressive par la répétition lexicale. Le logiciel permet à l'utilisateur d'entrer son propre corpus de textes; une version web sur le site *ReadingEnglish*[65] propose un corpus des textes d'anglais simplifié tirés du site de la VOA classifiés de la même manière que *TextLadder*. Le site du projet taïwanais d'apprentissage Internet *Candle*[66] inclut un module de pratique de la lecture appelé ***TextGrader***, conçu par Huang & Liou (2007) et basé sur le travail de Ghadirian (2002). Les deux projets ne diffèrent pas sensiblement, mais seuls Huang & Liou (2007) cherchent à mesurer l'efficacité de l'apprentissage de vocabulaire par cette méthode.

Les textes sont classifiés à l'aide de listes lexicales compilées pour l'anglais par d'autres chercheurs, la *GSL* (*General Service List*) de West (1953) et l'*UWL* (*University Word List*) de Xue & Nation (1984). Ces listes représentent le vocabulaire le plus fréquent en anglais et le

[63] Le site du projet propose un outil d'évaluation de la lisibilité basé sur la statistique lexicale : http://reap.cs.cmu.edu/demo/readability2012/

[64] http://www.readingenglish.net/software/

[65] http://www.readingenglish.net/students/

[66] http://candle.cs.nthu.edu.tw/newcandle/Home_C.asp

vocabulaire supplémentaire le plus fréquent des textes académiques respectivement. Notons qu'il ne s'agit pas simplement de regrouper les unités lexicales en lemmes (différentes formes d'une même unité lexicale) mais de regrouper les unités lexicales apparentées dans des *familles de mots* (des unités lexicales qui ont la même racine et une ressemblance sémantique), environ 2000 familles de mots pour la GSL, 800 pour l'UWL. À ces listes, Huang et Liou ajoutent la liste HSF (High School Frequency Word List), utilisée dans l'élaboration des manuels scolaires à Taïwan, et leur propre liste spécifique au corpus choisi. Ghadirian utilise aussi une liste de base de vocabulaire spécifique à son corpus, la *VOA Special English Word List*[67].

En premier lieu, ces listes sont utilisées pour filtrer les textes appropriés au lecteur. Seuls les textes couverts à au moins 95% par ces listes sont retenus, suivant le principe que le lecteur en L2 doit connaître 95% des mots d'un texte avant de pouvoir déduire le sens des mots inconnus (Liu & Nation 1985). Ensuite sont créées deux listes : une de vocabulaire cible, et l'autre de vocabulaire connu. Ghadirian n'utilise comme liste de vocabulaire connu que du vocabulaire connu des débutants : une partie de la GSL (les 176 premières familles de mots) augmentée avec du vocabulaire de base. Huang et Liou utilisent la totalité de la GSL plus la liste HSF comme liste de vocabulaire connu. Les autres listes composent la liste de vocabulaire cible dans chaque cas. Huang et Liou fournissent la taille et le pourcentage de couverture du corpus pour chacune de leurs listes : 9712 mots pour la liste de vocabulaire connu seule avec une couverture de 92,01%, et 13 614 mots pour les deux listes ensemble avec une couverture de 96,61%.

Après élimination de textes qui ne sont pas couverts à 95% par les listes lexicales initiales (les textes sont aussi sélectionnés pour leur taille), le système choisit le texte le plus facile. C'est le texte avec le plus grand nombre de mots de la liste de vocabulaire connu et le moins de mots de la liste de vocabulaire cible. Ce texte devient le premier texte dans la séquence de textes à lire. Le vocabulaire cible identifié dans ce premier texte est ensuite rajouté à la liste de vocabulaire connu. Seul le vocabulaire de la liste de vocabulaire cible est rajouté, le vocabulaire inconnu qui ne figure pas dans la liste de vocabulaire cible n'est pas pris en compte. Cette nouvelle liste est utilisée pour identifier le deuxième texte de la séquence de lecture et ainsi de suite. Le système de *TextGrader* diffère légèrement de *TextLadder* dans la mesure où Huang et Liou ont conçu un algorithme qui favorise aussi la répétition de vocabulaire cible. Le vocabulaire cible rencontré n'est pas mis directement dans la liste de vocabulaire connu, mais dans une troisième liste, de vocabulaire cible exposé, le processus de classification favorisant le vocabulaire de cette liste.

Les textes de *TextGrader* sont présentés avec le vocabulaire cible en surbrillance, avec variation de la couleur de surbrillance selon qu'il s'agit de la première occurrence d'un mot ou qu'il s'agit d'une répétition. Les deux systèmes ont intégré un dictionnaire (*Google translate* pour *TextLadder* et un dictionnaire anglais-chinois pour *TextGrader*), *TextLadder* permettant au lecteur de rechercher la définition des tous les mots du texte, alors que *TextGrader* ne glose que le vocabulaire cible.

Ghadirian (2002) décrit l'utilisation de *TextLadder* pour créer un programme de lecture utilisant un corpus de 266 textes en anglais simplifié de *Voice of America*. Dans la première

[67] Cette liste est disponible sur le site de la VOA à l'adresse suivante : http://docs.voanews.eu/en-US-LEARN/2014/02/15/7f8de955-596b-437c-ba40-a68ed754c348.pdf

version de *TextLadder*, le pourcentage de vocabulaire cible initial est très élevé et baisse à 1% après 80 textes. Le lecteur doit lire 253 textes afin de rencontrer tout le vocabulaire cible. Dans sa version modifiée l'algorithme inverse la stratégie de sélection de textes si le pourcentage de vocabulaire cible baisse à 5%. Cette changement de stratégie a pour but de rendre plus constant le taux de vocabulaire cible d'un texte à l'autre et parvient à assurer que le pourcentage de vocabulaire cible baisse plus lentement, 1% n'étant pas atteint qu'après la lecture de 144 textes, tout le vocabulaire cible apparaissant dans les premiers 216 textes.

Huang & Liou (2007) ont utilisé *TextGrader* pour créer un programme de lecture en anglais expérimental de douze semaines pour des étudiants sinophones taïwanais. Le corpus de 5008 articles en anglais du magazine taïwanais *Sinorama* était choisi pour s'assurer que les connaissances autres que linguistiques des étudiants correspondraient bien au contenu des textes. Des 124 textes résultant du tri initial, seuls les 16 premiers textes étaient retenus pour le programme de lecture. Les chercheurs démontrent que leur système qui favorise un rapprochement entre les occurrences de vocabulaire cible en réduisant les intervalles de répétition est plus performant que celui de Ghadirian (2002).

Les chercheurs qui ont développé ces deux systèmes ont des buts légèrement différents. Comme nous avons évoqué plus haut, *TextGrader* est un module qui cherche à renforcer la pratique de la lecture au sein d'un système plus complexe d'apprentissage électronique qui comprend d'autres compétences linguistiques (l'écriture, la traduction, la compréhension de l'orale et ainsi de suite). *TextLadder*, en revanche vise spécifiquement l'acquisition du nouveau vocabulaire. Dans son descriptif du logiciel Ghadirian (2002) envisage son utilisation pour des étudiants de cours spécifiques, comme l'anglais commercial, par exemple. Cette démarche nous intéresse particulièrement, en premier lieu parce que l'acquisition d'un vocabulaire spécifique est plus abordable que l'objectif plus large d'expansion du vocabulaire général. Nous discuterons de ce problème plus en détail dans la partie suivante. En second lieu, il nous semble que le classement automatique de textes avec un vocabulaire similaire pourrait s'avérer efficace pour la lecture de textes authentiques, dans la mesure où le vocabulaire se répétera plus souvent dans un corpus plus homogène. Par conséquent, nous avons privilégié des corpus homogènes pour nos tests, sans négliger d'examiner le résultat d'un traitement sur un corpus moins homogène à titre comparatif.

Le fonctionnement de notre version thaïe de *TextLadder* est détaillé dans la section 4.3.

## 3.3. L'élaboration et l'utilisation des listes de fréquence lexicale

L'élaboration des listes de fréquence lexicale permet de répondre (au moins partiellement) aux questions que se pose tout apprenant d'une langue étrangère sur le vocabulaire, « Combien de mots faut-il apprendre ? » et « Quels mots ? » Ces questions sont traitées en profondeur par Nation (2004), qui distingue vocabulaire de haute fréquence, qui couvre un pourcentage important des textes, et les différents types de vocabulaire qui couvrent le pourcentage restant. Le tableau ci-dessous illustre la taille de vocabulaire (en lemmes) et la couverture textuelle des textes du corpus anglais *Brown*.

*Tableau 1 Couverture textuelle des tranches successives de 1000 lemmes dans le corpus Brown (reproduit de Nation (2004, p15))*

| Taille vocabulaire (lemmes) | Couverture textuelle |
|---|---|
| 1000 | 72.0% |
| 2000 | 79.7% |
| 3000 | 84.0% |
| 4000 | 86.8% |
| 5000 | 88.7% |
| 6000 | 89.9% |

On constate que la première tranche de mille lemmes représentent 72% des mots du corpus, et plus le vocabulaire augmente, moins il rajoute à la couverture textuelle. Le seuil généralement retenu pour la création d'une liste lexicale de haute fréquence entre le vocabulaire de haute et de faible fréquence est 3000 lemmes, ou, comme nous avons évoqué, 2000 familles de mots, des mots avec la même racine et une association sémantique. Par exemple, les mots associés à *laver* et qui ont la même racine *lav-* appartiennent à la même famille de mots *: délavé, lavable, lavage, laverie, lavoir, prélavage*. La taille de ce vocabulaire de haute fréquence varie d'une langue à l'autre. Par exemple, une liste de fréquence lexicale des 1500 mots les plus fréquents de la langue française, constituée par le lexicologue Étienne Brunet est proposée aux enseignants de français sur le site *éduscol*[68].

Le pourcentage restant inclut des entités nommées et du vocabulaire de faible fréquence, mais pas uniquement. Les listes lexicales utilisées dans l'enseignement tentent de représenter la fréquence lexicale dans la totalité de la langue, mais restent limitées et ne couvrent pas toute la langue. Pour un texte donné il peut y avoir un pourcentage significatif de vocabulaire spécialisé qui ne figure pas dans la liste de haute fréquence. Par exemple, la liste lexicale de vocabulaire académique en anglais élaborée par Coxhead (2000)[69] couvre 8,5% du vocabulaire d'un corpus académique (Nation, 2001, p17). Certains ont tenté d'élaborer une telle liste pour le thaï (voir la section suivante 3.3.1), mais il n'est pas certain que les textes académiques de toutes les langues ont le même profil de fréquence lexicale. Cobb et Horst (2004) plaident contre une liste académique pour le français. Des listes analogues ont été développées pour le suédois, le norvégien et le danois (Johansson Kokkinakis et coll., 2012) et le portugais (Baptista et coll., 2010), souvent avec référence à la liste pour l'anglais.

Quel pourcentage des mots d'un texte est-il nécessaire de connaître pour qu'un lecteur d'une langue étrangère puisse le comprendre ? Hu & Nation (2000) suggèrent que pour l'anglais, une connaissance de 80% est insuffisante pour la compréhension. À 90%, une minorité de lecteurs pourront démontrer une compréhension suffisante, un peu plus pour 95%.

[68] http://eduscol.education.fr/cid50486/liste-de-frequence-lexicale.html

[69] Il s'agit du AWL = Academic Word List, qui est composée de 570 familles de mots.

Ils calculent qu'une connaissance de 98% (1 mot sur 50) des mots d'un texte est nécessaire pour une compréhension adéquate. Un taux de 95% de mots connus constitue donc un minimum pour comprendre un texte en anglais. Une liste de fréquence lexicale générale est insuffisante pour couvrir un texte à 95%, d'où la nécessité de recourir à des listes spécifiques.

Les programmes linguistiques modernes sont basés sur les compétences linguistiques, organisés de façon thématique. Alors qu'il y a bien une corrélation entre la taille du lexique connue d'un apprenant et son niveau de compétence linguistique, le vocabulaire n'est pas spécifié de manière précise dans les descriptifs officiels des niveaux, bien que certains aient tenté de créer des listes lexicales par niveau (Milton, 2010). Les estimations de taille de vocabulaire nécessaire pour atteindre les différents niveaux du CECR par exemple, se sont faites sur les connaissances réelles des étudiants pour des langues différentes. Les résultats diffèrent selon la langue maternelle des étudiants et la langue étudiée. Ainsi, la taille du vocabulaire nécessaire pour atteindre le même niveau en grec serait beaucoup plus importante que pour le français, environ 1400 vocables supplémentaires pour le niveau B2 par exemple (Milton, 2010 ; Milton & Alexiou, 2009). Un corpus de grande envergure, le *Cambridge Learner Corpus* (CLC) tient compte des compétences réelles des apprenants d'anglais langue étrangère. Le site *English Vocabulary Profile*[70] utilise ce corpus pour fournir un dictionnaire qui identifie le niveau CECR des entrées à côté des définitions et exemples. L'innovation de ce dictionnaire est de diviser les différents sens des entrées par niveau CECR. Par exemple, pour l'entrée *programme*, le sens *émission de télévision* est identifié niveau A2, alors que le sens *plan, projet* est au niveau B2.

Les listes de fréquence lexicale sont élaborées à partir de corpus, qui peuvent être rassemblés de manière différente selon les objectifs des listes (Browne, 2014), et doivent représenter l'état actuel de la langue pour des listes lexicales destinées à l'usage dans l'enseignement de langues étrangères. Il doit aussi prendre en considération l'équilibre des genres textuels dans le corpus, si le corpus est censé représenter la langue en général[71].

Une méthode pour créer une telle liste de fréquence lexicale est décrite dans Nation (2012), qui recommande l'utilisation du logiciel *AntWordProfiler*[72] qui permet d'établir une liste de fréquence lexicale à partir d'un corpus de textes, tenant compte de la dispersion et la répartition de vocabulaire dans les textes.

La maison d'édition *Routledge* publie une série de *Frequency Dictionaries* destinée aux apprenants de langues étrangères.[73] À l'exception du dictionnaire allemand (4034 mots), chacun présente les 5000 premiers lemmes de la langue en ordre de fréquence. Chaque liste est élaborée à partir d'un corpus de grande taille, chaque chercheur étant soucieux d'utiliser un corpus équilibré qui comporte des textes oraux et écrits des sources variées et représentatives de la langue contemporaine.

[70] http://vocabulary.englishprofile.org/

[71] Voir Browne (2014) et le site http://www.newgeneralservicelist.org/ pour une discussion sur l'élaboration d'une nouvelle GSL (NGSL).

[72] http://www.laurenceanthony.net/software/antwordprofiler/

[73] La liste de ces ouvrages figure en fin de mémoire.

*Tableau 2 Corpus utilisés dans l'élaboration des dictionnaires de fréquence lexicale Routledge*

| **Langue** | **Corpus** | **Taille (mots)** |
|---|---|---|
| anglais américain contemporain | Corpus of Contemporary American English[74] | 385 millions |
| allemand | Leipzig/BYU Corpus of Contemporary German | 4,2 millions |
| arabe | | 30 millions |
| chinois | Lancaster Corpus of Mandarin Chinese[75] | 50 millions (73 millions de caractères) |
| espagnol | textes du 20e siècle du Corpus del Español[76] | 20 millions |
| français | | 23 millions |
| japonais | Corpus of Spontaneous Japanese (CSJ) & Balanced Corpus of Contemporary Written Japanese (BCCWJ) | 100 millions |
| néerlandais | SoNaR STEVIN Nederlandstalig Referentiecorpus [77] CGN Corpus Gesproken Nederlands[78] | 290 millions |
| portugais | Corpus do Português[79] | 20 millions |
| russe | I-RU Russian Internet Corpus[80] | 150 millions |
| tchèque | SYN2005, ORAL2006 & ORAL2008 du Corpus National Tchèque[81] | 102 millions |

Comme nous voyons à partir du tableau qui résume les différents corpus utilisés, l'élaboration d'un tel corpus nécessite un travail conséquent de choix de textes représentatifs, d'étiquetage/lemmatisation et de correction. Nous avons donc utilisé un corpus existant pour élaborer nos listes de fréquence lexicale thaïes.

### 3.3.1. Listes de fréquence lexicale en thaï

Malgré l'existence d'un corpus national de thaï, les publications concernant les *listes* de fréquence lexicale en thaï sont rares. Le site du *Thai National Corpus*, le *TNC*[82], คลังข้อมูลภาษาไทยแห่งชาติ /kʰlaŋkʰɔ̂ːmuːnpʰaːsăːtʰajhèŋtɕʰaːt/ met à disposition une liste des 5000 unités lexicales les plus fréquentes dans le corpus.

[74] http://corpus.byu.edu/coca/
[75] http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/
[76] http://www.corpusdelespanol.org/
[77] http://lands.let.ru.nl/projects/SoNaR/
[78] http://lands.let.ru.nl/cgn/
[79] http://www.corpusdoportugues.org/
[80] disponible à http://corpus.leeds.ac.uk/ruscorpora.html
[81] http://korpus.cz/
[82] Il s'agit du TNC (Thai National Corpus) คลังข้อมูลภาษาไทยแห่งชาติ hébergé sur le site de l'Université de Chulalongkorn http://www.arts.chula.ac.th/~ling/TNCII/corp.php Voir aussi Aroonmanakoon (2007a).

Le concordancier[83] mis en ligne par le département de linguistique à l'Université de Chulalongkorn permet d'y trouver le rang d'un lexème dans le corpus, essentiellement un corpus de presse. Le concepteur, Wirote Aroonmanakun, propose un document de statistiques sur le thaï (Aroonmanakun, 2006), qui comprend une liste des 200 premiers mots de la langue calculée à partir d'un corpus de presse.

La recherche la plus étendue semble celle exposée sur la partie du site *SEAlang* dédiée au vocabulaire thaï[84]. Des listes de vocabulaire, thématiques, ou spécifique à un manuel de langue (Brown (1986) mentionné ci-dessus) donné sont proposées. Une liste de vocabulaire spécifique aux textes académiques (appelée *Thai AWL*) est aussi proposée. À chaque élément des listes de vocabulaire est attribué un *WebRank* de 1 à 7+ (du plus fréquent au moins fréquent) selon sa fréquence relative dans un corpus web.

La première liste comparable au *GSL* anglais, c'est-à-dire destinée à l'usage des apprenants et des enseignants de thaï langue étrangère, est la *Thai Frequency Dictionary* de Nilsen (2014), qui comporte 4000 entrées, liste élaborée à partir de la liste des 5000 unités lexicales les plus fréquentes publiée sur le site du TNC[85]. Les 4000 premiers éléments de cette liste sont mis à disposition sur son site *thaifrequency*[86], chaque élément accompagné de son rang dans le TNC, la prononciation, la traduction en anglais et une phrase type. Il faut noter que le contenu du corpus TNC fait que cette liste n'est pas forcément adaptée pour le thaï langue étrangère. Selon Nilsen, les catégories principales du corpus TNC sont : textes académiques (28.6%)[87], textes non académiques (16.7%), presse (15.2%), fiction (22.8%), droit (3.9%) et divers (12.6%).

Nilsen (sans date) a aussi élaboré une autre liste de vocabulaire spécifique aux textes académiques. Cette liste est élaborée à partir de la liste 5000 du *TNC*, qui est composé dans une proportion importante de textes académiques (28.6%), alors que la *Thai AWC* de *SEAlang* est une traduction d'une liste anglaise[88].

L'article de Bofman et Prez (2008) est la seule publication académique qui traite de la fréquence lexicale en thaï appliquée à l'apprentissage du thaï comme langue étrangère. L'étude porte sur un corpus d'un genre spécifique, les chansons thaïes, très populaires auprès des étudiants et professeurs de thaï, car la pratique du karaoké est très répandue en Thaïlande. Les auteurs ont créé un corpus à partir des paroles de 400 chansons, segmentées à la main, qu'ils ont examiné à l'aide du logiciel *Oxford WordSmith Tools 4.0*. Parmi les 103 000+ mots du corpus, ils ont identifié 4931 vocables différents. Ils ont trouvé que les 39 vocables les plus fréquents représentent 50,02% des mots du corpus, déduisant que ce corpus serait particulièrement utile pour les débutants. Une liste du vocabulaire le plus fréquent de 400 éléments est fournie en fin d'article. Une caractéristique notée par les auteurs est la présence de variantes orthographiques de certains mots, des translittérations (de thaï en thaï), par exemple ร๊

[83] http://www.arts.chula.ac.th/~ling/ThaiConc/

[84] http://sealang.net/thai/vocabulary/

[85] http://www.arts.chula.ac.th/~ling/TNC/category.php?id=58&

[86] http://www.thaifrequency.com/

[87] Ces informations sont de mai 2013, période durant laquelle Nilsen a téléchargé la liste du site de la TNC.

[88] L'AWL (Academic Word List) de Coxhead (2000) qui est censée remplace l'UWL de Xue & Nation (1984). Voir aussi http://www.victoria.ac.nz/lals/resources/academicwordlist/

/rɯ́ʔ/ au lieu de หรือ /rɯ̌ː/ *ou*, เค้า /kʰáːw/ au lieu de เขา / kʰăw/ *il(s), elle(s)*, et ไม๊ /máj/ au lieu de ไหม /măj/ (particule d'interrogation).

*Tableau 3 Les premiers éléments de la liste de fréquence du corpus ThaiWaC*

| Vocable | Fréquence |
|---|---|
| ที่ | 2048685 |
| เป็น | 1332566 |
| ใน | 1265380 |
| การ | 1237354 |
| มี | 1232348 |
| ได้ | 1165984 |
| ของ | 1163969 |
| ไม่ | 1147416 |
| ... | ... |

Par souci de cohérence avec les textes de nos corpus, nous avons choisi d'utiliser une liste de fréquence basée sur un corpus généré avec la méthode *Corpus Factory* (Kilgarriff et coll., 2010) sur le site de *Sketch Engine*[89] pour notre projet, car l'outil de segmentation utilisé (*SWATH*, voir ci-dessous) est le même que nous avons utilisé pour segmenter les textes de nos corpus. *Sketch Engine* est un outil de création de corpus à partir du web, disponible pour un grand nombre de langues, dont 36 possèdent déjà un corpus de référence. Le corpus thaï *ThaiWaC* fournit par *Sketch Engine*, contient 108 013 897 tokens. De cela nous avons généré une liste de fréquence, de 319 756 éléments, triés par fréquence lexicale, chaque élément accompagné du nombre de ses occurrences dans le corpus de référence. La méthode *Corpus Factory* détaillée dans Kilgarriff et coll. (2010) consiste à télécharger le vidage-mémoire Wikimédia pour une langue donnée afin de créer un corpus à partir de Wikipédia et générer une liste de fréquence lexicale. Les éléments de moyenne fréquence sont utilisés pour interroger des moteurs de recherche et récupérer des pages web, qui sont nettoyées de tout balisage et publicité, avant d'être filtrées pour créer un corpus « propre ». Pour le corpus thaï, les textes sont segmentés au moyen de l'outil de segmentation *SWATH* (Meknavin et coll., 1997).

## 3.4. La segmentation lexicale en thaï

Tout traitement automatique de textes doit passer par une étape préalable de segmentation du texte en *mots*. Comment identifier ces unités ?

### 3.4.1. Le mot thaï

Avant de parler de la segmentation automatique, nous voulons expliciter l'identification des *mots* et la notion de *lexème* en thaï.

Nous avons déjà vu certains aspects de la langue et l'écriture thaïes qu'il faut prendre en compte dans une discussion sur la définition du mot en thaï. D'abord, le *mot orthographique* est quasi inexistant en thaï en vertu du fait que son système d'écriture n'utilise pas de délimiteur explicite de ces éléments (voir la liste en annexe 12.2 de règles d'utilisation de l'espace typographique en thaï pour de rares exemples de mots séparés du texte environnant par des

[89] https://www.sketchengine.co.uk/

espaces). Dans d'autres langues, où l'espace typographique s'utilise comme délimiteur de mots, le mot orthographique ne représente pas forcément le lexème. En français, par exemple, *pomme de terre,* un seul lexème, est composé de trois *mots orthographiques*. Le thaï ne connaît pas cet inconvénient. Compte tenu de ceci, et l'absence de flexion en thaï, il serait tentant de dire qu'on pourrait rapprocher le *mot*, concret, du *lexème,* notion abstraite. En effet, pour les besoins de segmentation de textes pour les apprenants d'une langue, une correspondance mot-lexème serait idéale pour l'élaboration des listes de vocabulaire. Mais comme nous allons voir, identifier un *mot* en thaï pour les besoins du traitement automatique n'est pas si simple.

Un aspect difficile à résoudre est celui de l'ambiguïté dans la segmentation de mots composés (voir la section 2.3.2.1 sur les mots composés). Prenons l'exemple de รถไฟฟ้า /rótfajfáː/ qui est composé de รถ /rót/ *véhicule* + ไฟ /faj/ *feu* + ฟ้า /fáː/ *ciel*. Selon le choix de la segmentation, il peut s'agir de

(31) รถ /rót/ *véhicule* + ไฟฟ้า /fajfáː/ *électricité* ⇒ *véhicule électrique*

(32) รถไฟ /rótfaj/ *train* + ฟ้า /fáː/ *ciel* ⇒ *métro aérien*

Dans ce cas, seul le contexte permet de décider définitivement.

Aroonmanakun (2007b) explique qu'une correspondance concept-mot n'est pas un critère suffisant pour déterminer les frontières des mots, un concept dans une langue pouvant s'exprimer avec plusieurs mots dans une autre. Par ailleurs, la segmentation qui précise ce qui constitue les frontières d'un lexème peut s'avérer subjective dans certains cas à cause des procédés de création des mots composés en thaï, car certains composés sont plus étroitement liés que d'autres. Aroonmanakun donne l'exemple de *garde-robe* ตู้เสื้อผ้า /tûːsûːapʰâː/ comparé à ตู้เย็น /tûːjen/ *frigo*. Le dernier, composé de ตู้ /tûː/ *armoire* + เย็น /jen/ *froid* ne désigne pas simplement un armoire qui est froid et ne peut pas s'analyser comme deux mots, alors que *garde-robe* ตู้เสื้อผ้า /tûːsûːapʰâː/, qui est composé de ตู้ /tûː/ *armoire* + เสื้อผ้า /sûːapʰâː/ *vêtements*, peut s'analyser comme étant composé de deux mots, d'autant plus que si on rajoute un adjectif, comme สีขาว /sǐːkʰǎːw/ *blanc*, il y a ambiguïté dans la structure. ตู้เสื้อผ้าสีขาว /tûːsûːapʰâːsǐːkʰǎːw/ peut s'interpréter comme

(33) [[ตู้[เสื้อผ้า]]สีขาว] *armoire blanche à vêtements* ou bien

(34) [ตู้[เสื้อผ้า[สีขาว]]] *armoire à vêtements blancs*.

Aroonmanakun pense que les outils de segmentation devraient se contenter de ne segmenter que les mots simples et les « vrais composés », les mots dont le sens diffère d'une manière signifiante de ses composants. Il est clair que les choix de segmentation vont différer selon les besoins du traitement. Pour un système de traduction automatique par exemple, une segmentation qui fournirait des mots aussi proches que possible des lexèmes serait souhaitable. Ces décisions sont aussi importantes pour les besoins de systèmes d'apprentissage de langues, car il faudrait fournir à l'apprenant du vocabulaire utile et pertinent, mais en même temps permettre à l'apprenti lecteur d'acquérir des stratégies de lecture, utilisant ses connaissances lexicales existantes pour les extrapoler au vocabulaire inconnu. Il faudrait donc en théorie jongler avec ces deux exigences, mais en réalité notre définition des mots est, au moins dans un premier temps, limitée aux résultats fournis par l'outil de segmentation choisi.

### 3.4.2. Les outils de segmentation automatique du thaï

Il existe un certain nombre d'outils de segmentation des textes en thaï. Haruechaiyasak et coll. (2008) identifient deux approches distinctes, celles à base de dictionnaires et celles qui utilisent la modélisation statistique, pour conclure qu'une approche qui incorpore une combinaison des deux serait l'idéale. Les outils qui utilisent comme base un dictionnaire segmentent les textes utilisant des stratégies divers[90] pour contourner les problèmes d'ambiguïté et de mots inconnus, choisissant les mots du dictionnaire selon des critères spécifiques, tels que le mot le plus long (*greedy matching* Poowarawan, 1986) ou les mots qui donnent une phrase avec un minimum de mots (*maximal matching* Sornlertlamvanich, 1993). D'autres ont employé des approches basées sur corpus, avec apprentissage statistique pour améliorer le dictionnaire, comme le modèle à trigramme (Kawtrakul et coll., 1997), ou la segmentation basée sur les caractéristiques (Meknavin et coll., 1997). Certains n'utilisent pas de dictionnaire pour contourner le problème des mots inconnus (Theeramunkong et coll., 2000). Aroonmanakun (2002) préconise la segmentation en syllabes, puis la reconnaissance de collocations. L'outil que nous utilisons, *SWATH* (*Smart Word Analysis for THai*)[91] de Meknavin et coll. (1997) améliore la segmentation des mots inconnus avec une analyse linguistique qui regarde l'environnement des mots, cherchant des mots de contexte et des collocations pour déterminer la segmentation la plus probable. Voir Berment (2004, p129) et Kosawat (2003) pour des discussions plus approfondies sur le développement des outils de segmentation en thaï.

En premier lieu, nous avons créé des corpus de textes segmentés au moyen du service REST de segmentation par *SWATH*, développé par Poltree et Saikaew (2011)[92]. À terme, *SWATH* sera intégré dans notre dispositif, car nous ne pouvions choisir le dictionnaire de l'outil et, en outre, le temps de réponse n'est pas très satisfaisant par rapport à d'autres méthodes (Noyunsan et coll., 2014).

วัน|หนึ่ง|_|เขา|ตื่น|ลืมตา|ขึ้น|มา|จาก|ความ|ฝัน|อัน|สับสน|_|รู้สึก|อ่อน|_|เพลีย|จาก|การ|พักผ่อน|ไม่|เพียงพอ|_|เขา|เดิน|โซเซ|ไป|ชน|ก|อง|หนังสือ|ใน|ห้องรับแขก|ที่|ตั้ง|ไว้|สูง|จนเป็น|กำแพง|กระดาษ|_|หนังสือ|เล่ม|ใหญ่|เกือบ|สิบ|เล่ม|หล่น|ลง|มา|ทับ|เท้า|ทำเอา|เขา|ร้อง|โอดโอย|_|แล้ว|เซ|ไป|เตะ|ซีดี|ที่|กอง|ไว้|อีก|มุม|ห้อง|จน|กระจัดกระจาย|_|ต้อง|เดิน|กะเผลก|_|ๆ|_|ผ่าน|ห้องหับ|อัน|มัวซัว|ไป|เข้า|ห้องน้ำ|ล้างหน้าล้างตา

*Figure 8 Exemple de segmentation résultant d'un traitement avec SWATH*

[90] Notons au passage des sites de projets d'outils de segmentation :
*SWATH* http://www.cs.cmu.edu/~paisarn/software.html
*ThaiWordseg* http://thaiwordseg.sourceforge.net/
*Wordcut* http://sourceforge.net/projects/thaiwordseg/
*thaiseg* http://pioneer.chula.ac.th/~awirote/resources/thai-word-segmentation.html
[91] http://www.cs.cmu.edu/~paisarn/software.html
[92] http://www.thaisemantics.org/service/swath/index (lien mort)

# 4. Méthodologie

## 4.1. Corpus

Nous avons utilisé des corpus de nature différente. Il s'agit de deux corpus de textes didactiques, et de corpus de textes authentiques. Tous les textes de ces corpus ont été segmentés avec l'outil de segmentation *SWATH* que nous venons de présenter.

- Un corpus de 64 textes didactiques du site *GLOSS* mentionné précédemment (voir section 3.1.2) : 427 mots par texte en moyenne, 27 362 tokens au total.
- Un corpus composé des 76 chapitres du *Thai Reader Project* (voir section 3.1.2), 47 pour le premier niveau et 29 pour le deuxième niveau. La conversion en texte du format PDF avec pdftotext[93] a provoqué des erreurs dont la plupart ont été rectifiées manuellement.
- Une collection de 80 nouvelles provenant du site *thaifiction* déjà cité. Chaque nouvelle est d'un auteur thaï différent, 2480 tokens par nouvelle en moyenne, 198 401 tokens au total.
- Un deuxième corpus littéraire composé d'un texte long (42 066 tokens), le troisième tome du roman *SiPhanDin* สี่แผ่นดิน /sì: pʰèndin/ (*Quatre Règnes*) de l'auteur Kukrit Pramoj คึกฤทธิ์ ปราโมช /kʰúʔkrìt pra:mô:t/.
- Un corpus de presse composé d'articles publiés en 2013[94] par un journal quotidien populaire à grand tirage, *Thai Rath* ไทยรัฐ /tʰajrát/, divisé en rubriques :

*Tableau 4 Composition du corpus de presse ThaiRath2013*

| Rubrique | Nombre d'articles[95] | Taille moyenne des articles (tokens) | Taille totale (tokens) |
|---|---|---|---|
| **International** | 2300 | 398 | 915 434 |
| **Société** | 2546 | 676 | 1 720 007 |
| **Divertissement** | 1829 | 543 | 993 321 |
| **Business** | 2464 | 495 | 1 215 675 |
| **Lifestyle** | 2418 | 685 | 1 655 514 |
| **National** | 3589 | 518 | 1 860 088 |
| **Sport** | 4260 | 430 | 1 832 586 |
| **Politique** | 4210 | 562 | 2 366 964 |

[93] http://linux.die.net/man/1/pdftotext
[94] L'année 2013 correspond à l'année 2556 du calendrier bouddhiste utilisé en Thaïlande.
[95] Nous n'avons gardé que les textes d'une longueur de plus de 200 mots.

## 4.2. Élaboration des listes de fréquence lexicale pour le thaï

### 4.2.1. Évaluation des listes de fréquence

Avant de déterminer le contenu de nos listes de vocabulaire connu et vocabulaire cible pour notre *Thai TextLadder*, nous devons trouver une liste de vocabulaire de haute fréquence. Pour ce faire, nous utilisons une grande liste de fréquence lexicale créée à partir du corpus de référence de thaï de *Sketch Engine,* (détaillée dans la section 3.3.1) *ThaiWaC*, qui comporte 319 756 éléments.

Le fait que nous traitons une langue isolante signifie qu'une étape de lemmatisation est inutile. La question de familles de mots est plus délicate, compte tenu des caractéristiques de morphologie lexicale que nous avons vues (voir sections 2.3.2.1 et 3.4.1). Les procédés de composition des mots composés, une fois identifiés par l'apprenant, peuvent l'aider à tirer du sens d'un vocable inconnu, mais étant donné qu'il ne s'agit pas à chaque fois de racines et d'affixes, mais de vocables à sens plein, un vocable appartiendrait à plusieurs familles de mots. Nous avons remarqué que l'outil de segmentation sépare parfois certains préfixes comme ความ /kʰwaːm/ et การ /kaːn/ du verbe ou substantif qu'ils modifient. Cela a pour l'effet de réduire le nombre de lexèmes, assimilant certains lexèmes à préfixe au lexème modifié. Ainsi, un lexème comme ความรัก /kʰwaːmrák/ *l'amour* serait segmenté en ความ /kʰwaːm/ (*préfixe*) et รัก /rák/ *aimer*. Il faut néanmoins souligner que le résultat du traitement de ces préfixes par l'outil de segmentation n'est pas uniforme, ni toujours souhaitable, car le sens de certains mots à préfixes n'est pas identifiable par ses composants. Le mot ความเป็นมา (*histoire, origine, contexte*) par exemple, est composé de ความ /kʰwaːm/, เป็น /pen/ (*être*) et มา /maː/ (*venir*).

Afin de s'assurer de la capacité de couverture des listes issues des données de *ThaiWaC,* nous comparons ces données avec d'autres listes que nous avons trouvées en les testant sur certains de nos corpus. Nous comparons chaque liste de fréquence lexicale avec une liste créée à partir de *ThaiWaC* de la même taille (ou d'une taille similaire). Le tableau 4.2 résume les résultats. Chaque nom de liste est suivi d'un _ et sa taille en mots. La couverture d'un corpus par une liste donnée est exprimée en pourcentages.

Plusieurs facteurs entrent en jeu dans l'explication des différences constatées. Tout d'abord la méthode de segmentation. L'utilisation d'un outil de segmentation autre que celui utilisé pour segmenter nos corpus réduit le nombre de mots reconnu d'une liste. N'oublions pas non plus que les listes *AUA_1275* et *chansons_400* ont été faites à la main, ce qui peut entraîner des erreurs de cohérence. L'année de l'élaboration de la liste peut expliquer la différence de couverture de listes de même taille surtout en ce qui concerne des textes de presse. Certaines listes ont été nettoyées de leur ponctuation et des mots écrits en lettres latines, comme celle de Nilsen (*TNC_JN_4000*), une version nettoyée de celle de la *TNC*.

Nous ne perdons pas de vue que la liste AUA_1275, tirée des manuels de langue, a été faite dans le but de fournir du vocabulaire utile à l'apprenant, sans prendre en compte la fréquence lexicale explicitement. Toutefois, il est intéressant de constater que la liste de vocabulaire que nous avons extrait du *Thai Reader Project*, un programme de cours destinés exclusivement à la lecture en thaï, démontre un très bon pourcentage de couverture, dépassant *ThaiWaC* pour les textes didactiques et un des corpus littéraires.

En général, on constate que la couverture des listes élaborées à partir de *ThaiWaC* surpasse toutes les autres. Les seules instances où le résultat est meilleur que celui d'une liste

de *ThaiWaC* apparaissent en rouge. Les résultats pour les données de *ThaiWaC* ne diffèrent pas sensiblement de ceux pour les données dont nous disposons pour le *TNC*, le plus grand corpus de thaï connu en date. Nous en déduisons que les données de *ThaiWaC* constituent de bonnes bases pour une analyse de la fréquence lexicale générale en thaï.

*Tableau 5 Évaluation de la couverture de listes de fréquence lexicale ThaiWaC*

| **liste_taille** | **Corpus** | | | |
|---|---|---|---|---|
| | **GLOSS** (textes didactiques) *27 362 mots* | **SiPhanDin** (texte littéraire) *42 066 mots* | **Nouvelles** (textes littéraires) *198 401 mots* | **ThaiRath2013** (presse - politique) *2 366 964 mots* |
| **newspaper_200**[96] | 37.22 % | 36.70 % | 37.22 % | 36.37 % |
| **ThaiWaC_200** | 48.27 % | 52.32 % | 49.51 % | 43.78 % |
| **chansons_400**[97] | 39.17 % | 49.77 % | 47.42 % | 35.04 % |
| **ThaiWaC_400** | 55.92 % | 62.74 % | 57.82 % | 51.89 % |
| **AUA_1275**[98] | 59.10 % | 63.42 % | 62.31 % | 46.96 % |
| **ThaiWaC_1000** | 68.55 % | 70.13 % | 69.00 % | 64.17 % |
| **TNC_4000**[99] | 81.92 % | 81.23 % | 83.23 % | 77.07 % |
| **TNC_JN_4000**[100] | 75.73 % | 80.57 % | 81.95 % | 75.03 % |
| **ThaiWaC_4000** | 83.00 % | 82.11 % | 83.76 % | 77.29 % |
| **TRW1&2_4891**[101] | 85.26 % | 83.63 % | 83.89 % | 74.31 % |
| **TNC_5000** | 83.45 % | 82.87 % | 84.72 % | 78.43 % |
| **ThaiWaC_5000** | 84.56 % | 83.26 % | 85.63 % | 78.52 % |

[96] Aroonmanakoon (2006)
[97] Bofman et Prez (2008)
[98] Le manuel de Brown (1986) sur la lecture est un complément des autres manuels de la même collection pour l'apprentissage du thaï publiés par *AUA Language Center*. La liste AUA_1275 est extraite des listes de vocabulaire de ces livres. Source : http://sealang.net/thai/vocabulary/
[99] Thai National Corpus http://www.arts.chula.ac.th/~ling/TNC/category.php?id=58&
[100] *Thai Frequency Dictionary* de Nilsen (2014), basée sur la liste 5000 du TNC.
[101] Liste de vocabulaire utilisé dans *Thai Reader Project* (détaillé dans la section 3.1.2) textes, exercices et vocabulaire cible confondus.

### 4.2.2. Détermination du seuil du lexique de haute fréquence

Afin de pouvoir départager le vocabulaire de haute fréquence du vocabulaire spécialisé, nous avons évalué la couverture des listes de taille croissante créées à partir de la liste de fréquence lexicale provenant du *ThaiWaC*. Les résultats (dont une partie est fournie en forme de tableau, voir tableau 6) sont transformés en courbe (figure 9). Les premiers 500 mots couvrent 51 à 60% des corpus, les premiers 1000 mots permettent de couvrir entre 60% et 68% des corpus. La deuxième tranche de 1000 mots ajoute 7 à 8% de couverture, la troisième environ 4%, et la quatrième seulement 2% environ et la cinquième entre 1,5 et 2%. On voit que la pente de la courbe diminue nettement à partir de 3000 mots, pour s'aplatir autour de 5000 mots. Nous avons donc décidé de prendre le seuil de 5000 mots pour définir le vocabulaire de haute fréquence. Il faut garder à l'esprit que le seuil est assez arbitraire, car la fréquence des mots de part et d'autre de ce seuil ne diffèrent pas de manière significative.

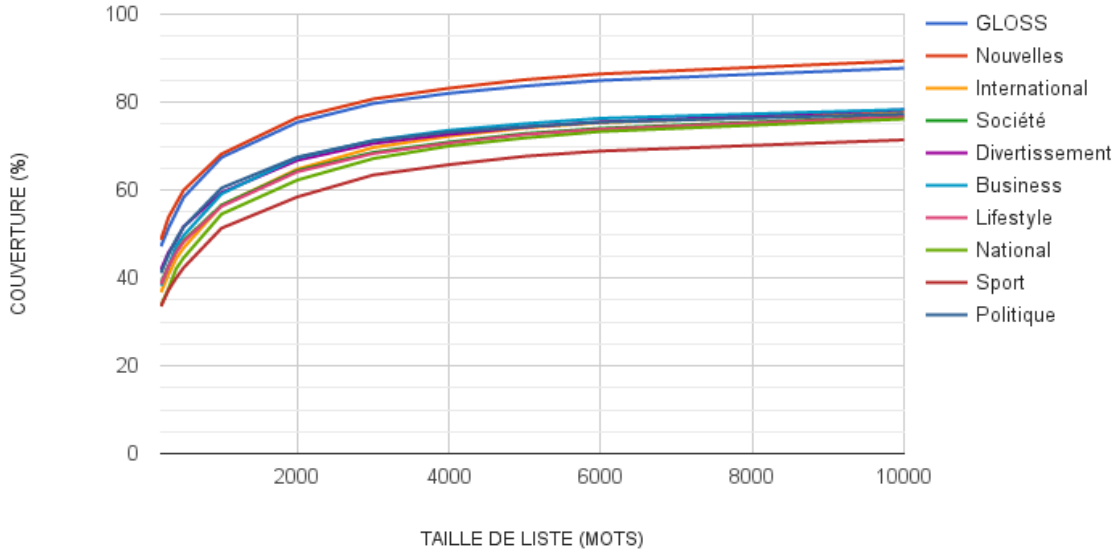



*Figure 9 Courbe de couverture textuelle par taille de liste ThaiWaC pour chaque corpus*

*Tableau 6 Couverture de corpus et fréquence lexicale (détail)*

| **Taille de liste *ThaiWaC* (mots)** | **GLOSS** (textes didactiques) *27 362 mots* | **Nouvelles** (textes littéraires) *198 401 mots* | **ThaiRath** (presse - internationale) *2 366 964 mots* |
|---|---|---|---|
| **200** | 47.18 % | 48.63 % | 41.09 % |
| **300** | 51.32 % | 53.75 % | 45.22 % |
| **400** | 54.83 % | 56.88 % | 48.77 % |
| **500** | 58.32 % | 59.96 % | 51.58 % |
| **1000** | 67.42 % | 68.21 % | 60.45 % |
| **2000** | 75.40 % | 76.45 % | 67.49 % |
| **3000** | 79.68 % | 80.71 % | 71.19 % |
| **4000** | 82.00 % | 83.16 % | 73.18 % |
| **5000** | 83.65 % | 85.10 % | 74.38 % |
| **6000** | 84.93 % | 86.41 % | 75.43 % |
| **10000** | 87.72 % | 89.37 % | 77.28 % |

Rappelons que *TextLadder* utilise deux listes de vocabulaire, une de vocabulaire connu, l'autre de vocabulaire cible. Il faudrait déterminer le vocabulaire que l'utilisateur connaît déjà pour pouvoir l'attribuer à la liste de vocabulaire connu, le vocabulaire de haute fréquence restant étant attribué à la liste de vocabulaire cible.

Un examen rapide de certains manuels de thaï langue étrangère nous apprend que le vocabulaire du premier niveau dépasse rarement 1000 vocables[102] (~800 pour Delouche, 2009a ; ~900 pour Hoonchamlong, 2007a ; ~950 pour Kesavatana-Dohrs 2007). Le vocabulaire des manuels qui présentent des matériels pédagogiques permettant d'acquérir un niveau de compétence linguistique intermédiaire ne dépasse pas 2000 vocables (~1700 pour Conjeaud & Pooput, 2010a ; ~1400 pour Hoonchamlong, 2007b)[103]. À ce niveau, seuls environ trois mots sur quatre des textes de nos corpus constituent du vocabulaire connu (voir tableau 4.3). Il est clair qu'il faudrait atteindre au moins un niveau de vocabulaire intermédiaire avant de prétendre à lire des textes authentiques.

La taille du vocabulaire d'un étudiant intermédiaire est plus difficile à cerner. Nous avons pu examiner le vocabulaire du *Thai Reader Project* plus en détail. Rappelons que *Thai Reader Project* est divisé en deux niveaux, intermédiaire et avancé, chaque chapitre étant composé d'un texte authentique, des exercices et un glossaire de vocabulaire cible présent dans le texte. Nous supposons que le niveau de vocabulaire nécessaire pour entamer confortablement le premier niveau serait le vocabulaire entier de tous les textes de ce niveau (3011 vocables) moins le vocabulaire cible à acquérir (482 vocables), soit un vocabulaire de 2529 mots. De même, la taille du vocabulaire prérequise pour entamer le niveau avancé serait de 3011 vocables. À la fin du niveau avancé l'apprenant aurait vu 3462 vocables, combiné avec le niveau intermédiaire, 4895 vocables. Notons qu'on est très proche du seuil de 5000 vocables que nous avons fixés pour le vocabulaire de haute fréquence.

*Tableau 7 Seuils de vocabulaire du Thai Reader Project*

| Niveau(x) | Taille vocabulaire prérequis | Taille vocabulaire total |
|---|---|---|
| intermédiaire | 2529 | 3011 |
| avancé | 3118 | 3462 |
| intermédiaire + avancé | | 4895 |

Le seuil entre niveau intermédiaire et niveau avancé semble se situer à 3000 vocables. C'est ce seuil que nous avons retenu, mais il serait certes souhaitable de tester le niveau de vocabulaire de l'utilisateur avant de déterminer la répartition du vocabulaire de haute fréquence.

[102] Ces chiffres ne sont pas exacts, mais sont basés sur les listes de vocabulaire fournies à la fin de chaque ouvrage, parfois il s'agit de lexiques L1-thaï au lieu de thaï-L1.

[103] Seule Hoonchamlong (2007a ; 2007b) précise un niveau de compétence linguistique de référence pour ses manuels, en l'occurrence le niveau intermédiaire moyen selon l'échelle ACTFL. Voir annexe 11.1 pour les équivalents CECR.

### 4.2.3. Lexique de basse fréquence

Au-delà du seuil des mots de haute fréquence, la figure 9 montre que la courbe de couverture s'aplatit nettement. Le nombre de mots qu'il faut apprendre pour atteindre les 95% de connaissance du vocabulaire semble insurmontable -- l'apprenant ne peut plus s'appuyer sur la stratégie d'apprentissage des mots de la langue en ordre de fréquence. En effet, l'acquisition de tous les mots d'une langue est non seulement un objectif irréaliste, mais irréalisable. Pour illustrer l'ampleur de la tâche, prenons l'exemple du corpus de presse, rubrique internationale. Après l'enlèvement du vocabulaire de haute fréquence, il reste 23 743 vocables, dont 10 669 (44,9%) sont des hapax. À partir de la recherche récente, Nation (2004) estime qu'un locuteur natif ajoute environ 1000 familles de mots à son vocabulaire chaque année de sa vie jusqu'à l'âge adulte. Ces objectifs ne sont pas impossibles pour l'apprenant d'une langue étrangère, surtout s'il s'agit d'une deuxième langue apprise *in situ*, mais dépassent les capacités de bon nombre des apprenants. Toutefois, il est possible de créer des listes de vocabulaire spécifique à un type de texte, stratégie qui permet de ramener le vocabulaire à apprendre à une taille plus raisonnable. La liste de vocabulaire cible de *TextLadder* et *TextGrader* s'appuie sur de telles listes de vocabulaire spécifique pour atteindre une couverture lexicale de 95%.

Une liste de vocabulaire spécialisé est constituée d'une fraction du vocabulaire de basse fréquence qui, ajoutée au vocabulaire de haute fréquence, permet d'augmenter de façon significative la couverture d'un corpus composé d'un genre textuel ou thème uniforme. Nous avons déjà vu qu'en anglais les listes de vocabulaire académiques ont été mises au point pour augmenter la couverture de textes académiques. La liste de 570 familles de mots de l'AWL (Coxhead, 2000) rajoute jusqu'à 8,5% de couverture d'un texte, alors que l'apprentissage de 1000 familles de mots supplémentaires n'aurait rajouté de 4,3% Nation (2001, p17). L'élaboration d'une telle liste doit prendre en compte non seulement la fréquence brute des vocables dans un corpus représentatif des textes que la liste est censée représenter, mais aussi leur répartition dans le corpus.

L'élaboration simpliste d'une telle liste de vocabulaire spécifique à la presse à partir de notre corpus *ThaiRath* permet d'ajouter un pourcentage significatif à la couverture de textes de presse par rapport à la couverture d'une liste de fréquence générale.

Notre liste est composée des 3404 vocables qui apparaissent dans toutes les rubriques de 80% de notre corpus presse et qui apparaissent au moins cinq fois à l'intérieur de chaque rubrique. Le tableau ci-dessous montre que la liste de haute fréquence en combinaison avec la liste de vocabulaire spécifique surpasse la liste de vocabulaire de fréquence générale de taille analogue en matière de couverture de textes de presse (testé sur les 20% restant du corpus). Cette liste de vocabulaire spécifique au corpus rajoute 13,86% de couverture en moyenne au 72,86% de la liste de vocabulaire de haute fréquence, une couverture totale de 86,72%, environ 10% de couverture de plus que la liste de vocabulaire de fréquence générale de taille similaire (76,58% pour la liste de 10 000 tokens).

*Tableau 8 Couverture d'une liste de vocabulaire spécifique au corpus de presse*

| Rubrique | Vocabulaire de haute fréquence 5000 tokens | Liste de vocabulaire spécifique presse 3404 tokens | Liste de vocabulaire de fréquence générale de 10 000 tokens |
|---|---|---|---|
| **Business** | 75,08 % | 9,67 % | 78,31 % |
| **Politique** | 74,38 % | 13,96 % | 77,28 % |
| **Divertissement** | 74,31 % | 14,98 % | 78,23 % |
| **International** | 74,11 % | 12,15% | 77,72 % |
| **société** | 72,85 % | 14,48 % | 76,84 % |
| **Lifestyle** | 72,64 % | 14.55 % | 76,68 % |
| **National** | 71,84 % | 15,75 % | 76,14 % |
| **Sport** | 67,67 % | 15,34 % | 71,40 % |
| **MOYENNE** | 72,86 % | 13,86 % | 76,58 % |

Bien que la couverture de ces listes spécifiques puisse s'avérer significative, le pourcentage restant n'est toujours pas pris en compte. Les deux systèmes *TextLadder* et *TextGrader* ont eu recours à des listes spécifiques non seulement au type de texte, mais spécifiques au corpus, afin de garantir qu'un nombre suffisant de textes sera couvert à 95%. L'autre inconvénient de ces listes de vocabulaire spécifique est que leur élaboration exige un corpus de textes représentatif et nous ne pouvons prédire quels textes un apprenant voudrait lire. Nous avons donc décidé de nous passer de listes spécifiques préfabriquées. Notre système créera plutôt une liste de vocabulaire spécifique au corpus de textes à classifier qui, en combinaison avec la liste de haute fréquence, pourra plus aisément couvrir 95% des textes.

## 4.3. Présentation de *ThaiTextLadder*

Cette partie détaille le fonctionnement de base de notre version de *TextLadder* adaptée pour le thaï. Les modifications mineures que nous y avons apportées lors de nos tests sur corpus sont détaillées dans la partie suivante.

À l'instar du logiciel *TextLadder* de Ghadirian (2002), notre dispositif dispose d'une liste de vocabulaire connu et une liste de vocabulaire cible. La liste de vocabulaire connu comprend les 3000 premiers éléments de la liste de fréquence lexicale tirée de *ThaiWaC*, et la liste de vocabulaire cible est composée des 2000 éléments suivants de la même liste *ThaiWaC*. Autrement dit, la liste de vocabulaire cible est composée des tous les éléments à partir du 3001e élément jusqu'au 5000e élément. Les textes du corpus soumis à la classification par le logiciel sont déjà segmentés par l'outil de segmentation *SWATH*. Le classement automatique des textes suit les étapes suivantes :

**Étape 1** Sélection des textes du corpus par la longueur (par défaut entre 300 et 1500 mots).

**Étape 2** Création d'une liste de vocabulaire spécifique au corpus. Cette liste comprend tous les éléments thaïs qui ne figurent pas dans les listes initiales, qui ont une fréquence minimum de huit occurrences et qui apparaissent dans au moins cinq textes.

**Étape 3** Ajout de la liste de vocabulaire spécifique au corpus à la liste de vocabulaire cible.

**Étape 4** Sélection des textes qui sont couverts à 95% par l'ensemble des listes. Sont considérés comme appartenant au vocabulaire connu tous les mots en lettres latines, les logogrammes (tels que les chiffres arabes, les symboles monétaires, etc.) et la ponctuation.

**Étape 5** Choix du texte le plus facile, celui qui contient le plus de vocabulaire connu.

**Étape 6** Ajout du texte à la séquence de lecture.

**Étape 7** Ajout du vocabulaire cible du texte sélectionné à la liste de vocabulaire connu.

Répétition des étapes 5 à 7, choisissant comme texte suivant le texte le plus facile qui n'a pas encore été sélectionné.

Le schéma de la page suivante (figure 11) résume les principales étapes du traitement.

Le vocabulaire cible des textes est balisé pour informer le lecteur s'il s'agit de la première occurrence du vocable, ou d'une répétition. La première occurrence du vocable dans la séquence de textes est soulignée en vert, les occurrences suivantes sont soulignées avec de différentes nuances de bleu, qui s'éclaircissent progressivement au fur et à mesure des répétitions. La figure 10 donne un exemple d'un texte balisé de cette manière.

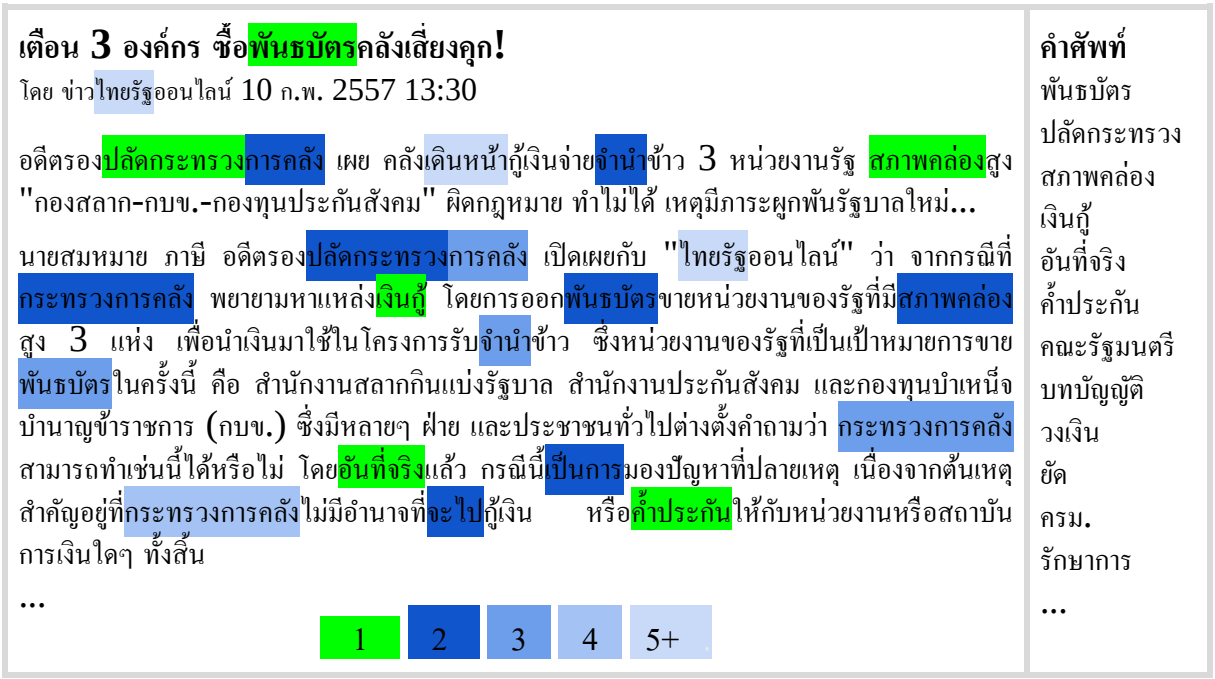



*Figure 10 Balisage du vocabulaire cible par ThaiTextLadder - échantillon d'un texte du corpus ThaiRath2013 rubrique Business en neuvième position sur une séquence de lecture générée par ThaiTextLadder. La liste du nouveau vocabulaire cible du texte est générée automatiquement (ici, à droite de texte).*

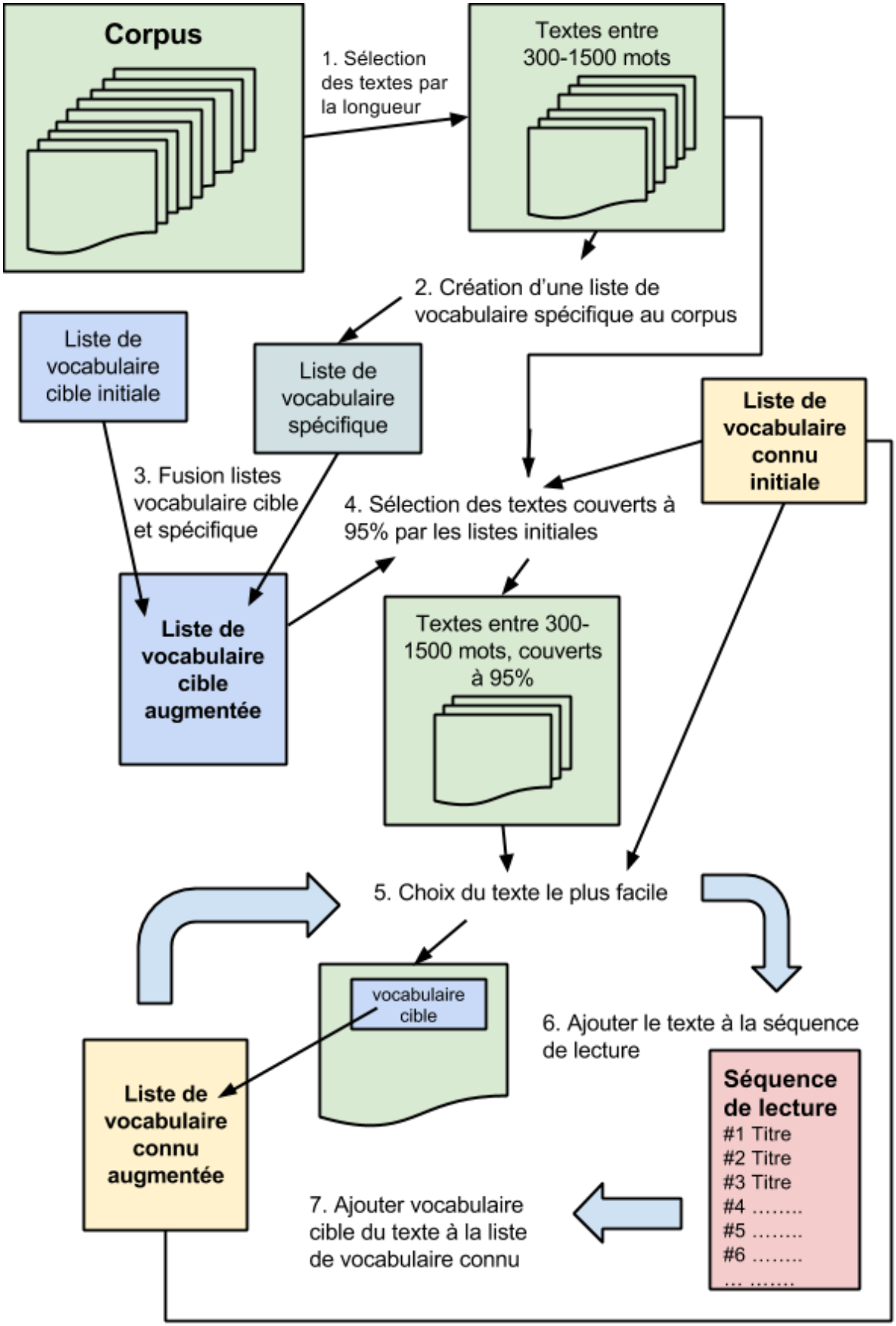



*Figure 11 Fonctionnement de ThaiTextLadder*

# 5. Tests sur corpus et résultats

## 5.1. Les maux de la segmentation

Tout comme Jean (2009), nous avons trouvé beaucoup d'instances de sursegmentation d'entités nommées dans les textes de nos corpus. Les textes de la rubrique *Sport* du corpus *ThaiRath2013* fournissent une multitude d'exemples de ce phénomène et ses conséquences, surtout en ce qui concerne les entités nommées étrangères. Prenons par exemple trois noms propres de notre corpus de presse *ThaiRath2013*, rubrique *Sport,* qui sont écrits en thaï

(35a) โรแบร์โต มันชินี /roːbɛːtoː mantɕʰíʔniː/ *Roberto Mancini*
(36a) โรนัลดินโญ /roːnandinjoː/ *Ronaldinho*
(37a) โรลังด์ การ์รอส /roːlaŋ kaːrɔ̂ːt/ *Roland-Garros*

L'outil de segmentation, qui n'a pas ces mots dans son dictionnaire, les divise en syllabes comme ceci :

(35b) โร แบร์ โต มัน ชิ นี
(36b) โร นัล ดิน โญ
(37b) โร ลังด์ การ์ รอ ส

Or, ces trois noms commencent par la même syllabe, โร /roː/, qui, en vertu de la renommée internationale de ces noms, devient un token très fréquent dans un corpus d'articles sur le sport. Notre système de création de liste de vocabulaire spécifique au corpus destinée à être ajoutée à la liste de vocabulaire cible va donc introduire soit des mots inexistants dans la langue, soit risquer d'y introduire des mots de très faible fréquence dans des cas d'homographie entre d'autres mots et ces syllabes.

Ce problème de sous-segmentation ne concerne pas uniquement les noms propres, mais aussi les mots anglais écrits en thaï. Nous avons constaté un grand nombre de ces mots dans les textes de presse, certains fréquemment utilisés en thaï peuvent se considérer comme des emprunts intégrés et d'autres, clairement considérés comme des mots étrangers à la langue, n'existant dans aucun dictionnaire de thaï. Parfois, ces mots sont utilisés comme des marques ou dans des slogans publicitaires, mais ils sont souvent introduits comme s'il s'agissait de mots thaïs. Parmi la multitude d'exemples que nous avons trouvée :

(38) วิคตอรี /wíktɔːriː/ (*victoire*)
(39) แชมเปียน /tɕʰɛːmpiːan/ (*champion*)
(40) แฮปปี้ไลฟ์ /hɛ̂ːppîːlaj/ (*vie heureuse*)
(41) ซัมเมอร์ /sammɤː/ (*l'été*)
(42) แฮตทริก /hɛːttrík/ (*coup du chapeau*)

L'outil de segmentation a un problème particulier avec des mots composés thaïs-anglais, comme การขายชอร์ต /kaːnkʰăːjtɕʰɔ̂ːd/ (*la vente à découvert*), composé d'un préfixe thaï การ /kaːn/, un verbe thaï ขาย /kʰăːj/ (*vendre*) et un mot anglais ชอร์ต /tɕʰɔ̂ːd/ (*court* de l'anglais *short*), tantôt segmenté การ ขายชอร์ต, tantôt การขาย ชอร์ต.

Nous avons aussi trouvé que notre outil de segmentation sursegmente aussi les noms propres thaïs, notamment les noms d'auteurs de notre corpus de nouvelles.

Prenant en compte nos connaissances de l'utilisation de l'espace en thaï (voir annexe 12.2), nous avons introduit un dispositif d'amélioration de la segmentation pour les entités nommées avant l'étape (2) de la création de la liste de vocabulaire spécifique au corpus, à l'aide d'une liste d'entités nommées extraite d'un vidage mémoire des en-têtes des pages Wikipédia thaï[104]. Ceci a eu l'effet d'améliorer sensiblement la segmentation des noms de personnes et de lieux avec très peu de cas de sous-segmentation, au moins pour le corpus *Sport*. Toutefois, les en-têtes des articles de Wikipédia contiennent un nombre considérable de titres de chansons et de films, ce qui porte à croire que cette méthode ne serait pas aussi efficace pour le traitement d'autres textes sans vérification humaine qualitative du contenu du vidage mémoire. Cette méthode très conservatrice qui ne considère que les morceaux de texte entourés d'espaces typographiques ne relève pas toutes les instances de sursegmentation, et il serait plus efficace d'améliorer l'outil de segmentation même en ajoutant ces entrées à son dictionnaire interne.

## 5.2. Tri par quantité de vocabulaire connu

Nous avons pris un échantillon de 300 textes pris au hasard de notre corpus d'articles de presse *ThaiRath2013*, rubrique *Business* pour une première aperçue de la classification de textes en séquence de lecture par *ThaiTextLadder*. Après les étapes d'élimination de textes par la longueur de couverture par les listes de vocabulaire initiales, il reste 156 textes à classifier.

La figure 12 représente la quantité de nouveau vocabulaire par texte sur la séquence de lecture ; l'axe horizontal représente la séquence de lecture, par numéro de texte sur la séquence. On constate que le vocabulaire nouveau du premier texte est très élevé, 71 vocables. La taille du nouveau vocabulaire des textes numéro 98 (37 vocables) et 148 (63 vocables) s'expliquent par une présence importante de noms de lieux peu présents dans les autres textes. Nous remarquons aussi qu'il y a 11 textes sans nouveau vocabulaire et 19 textes avec un seul vocable nouveau. La préférence donnée aux textes avec une grande quantité de vocabulaire connu a l'effet de privilégier la sélection de textes longs (voir figure 13). Ceci n'est pas un inconvénient en soi, mais il n'est pas nécessaire. En effet, l'intuition voudrait que les textes les plus longs soient les plus difficiles, et il serait donc préférable de les situer vers la fin de la séquence. Comme nous avons déjà évoqué, la compréhension d'un texte dépend du pourcentage de mots connus dans un texte, avec un seuil de 95% nécessaire pour la compréhension. La figure 14 montre que cette version de *ThaiTextLadder* ne permet pas une couverture textuelle du vocabulaire connu approchant ce seuil.

[104] Les vidages mémoire (*dumps*) du contenu de Wikipédia en thaï ont la préfixe *th*. Celui que nous avons utilisé est intitulé *List of all page titles*. Les vidages mémoire de Wikimédia sont disponibles à cette adresse : https://dumps.wikimedia.org/backup-index.html

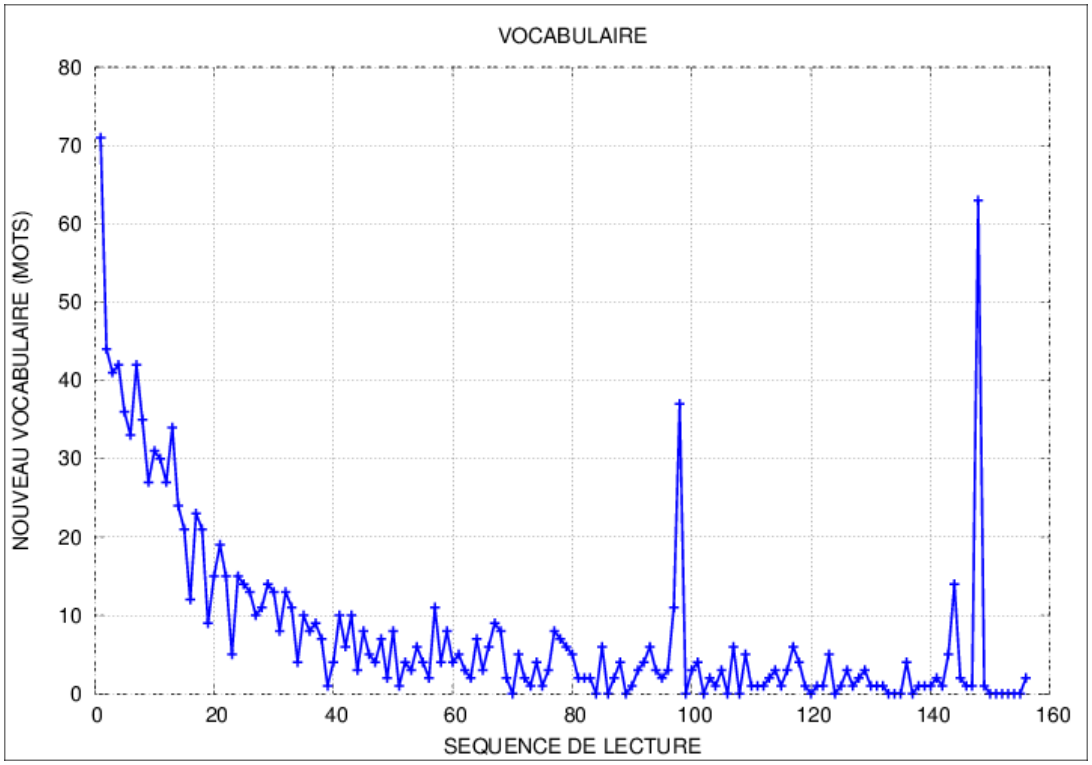



*Figure 12 Tri par quantité de vocabulaire connu - évolution du nouveau vocabulaire cible (corpus Business de 300 textes)*

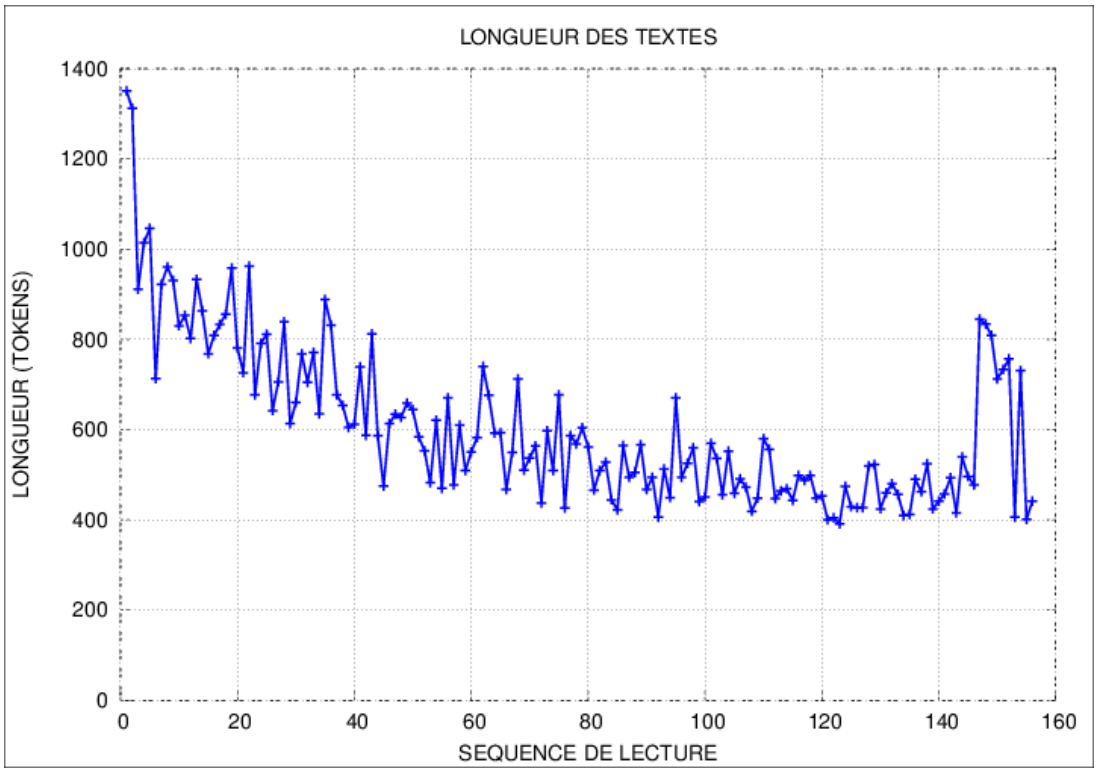



*Figure 13 Tri par quantité de vocabulaire connu - longueur de texte (corpus Business de 300 textes)*

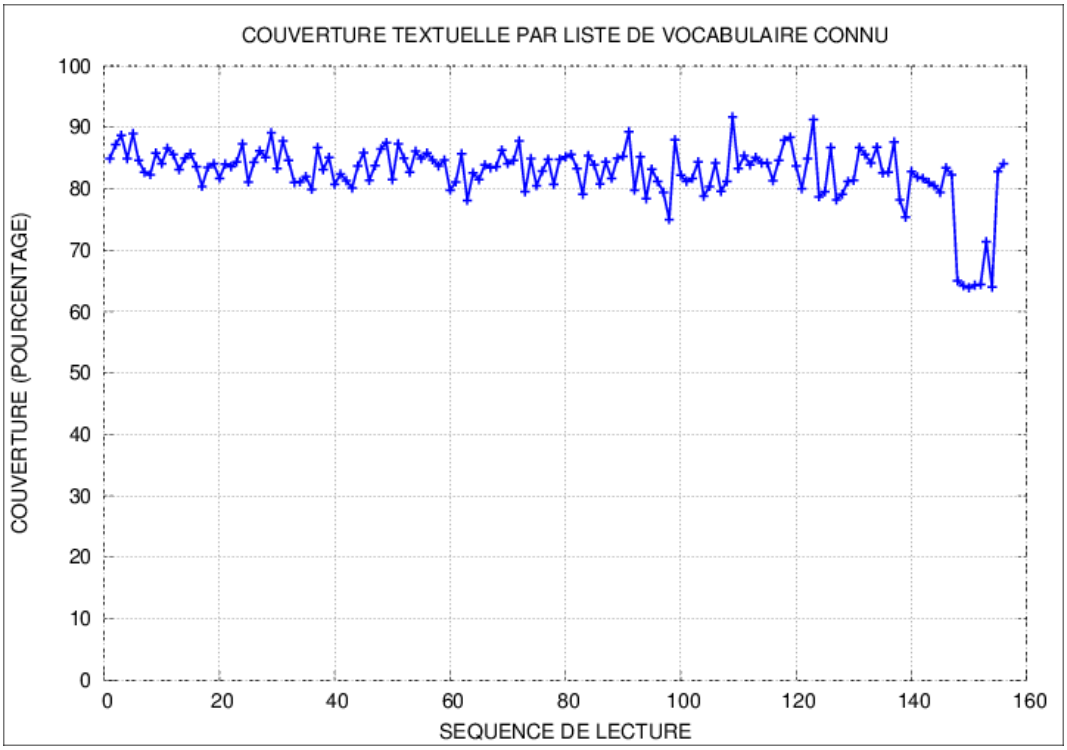



*Figure 14 Tri par quantité de vocabulaire connu - couverture textuelle (corpus Business de 300 textes)*

### 5.3. Tri par couverture textuelle

À la lumière des résultats que nous venons de présenter, nous avons modifié le processus de tri, pour favoriser le texte avec le plus grand pourcentage de couverture textuelle de la liste de vocabulaire connu. Pour mémoire, cette liste de vocabulaire connu est augmentée par le vocabulaire déjà vu au fur et à mesure que l'on ajoute les textes à la séquence de lecture. Si plusieurs textes ont le même pourcentage de couverture textuelle, *ThaiTextLadder* choisit le prochain texte à mettre sur la séquence de lecture selon les critères suivants : si le pourcentage de couverture est en dessous de 95%, il choisit celui qui a le moins de vocabulaire cible, mais si la couverture textuelle dépasse 95%, le texte avec le plus de vocabulaire cible est choisi. Ceci permet d'augmenter la couverture textuelle de vocabulaire connu et répartir le nouveau vocabulaire plus équitablement.

L'évolution de la taille du nouveau vocabulaire par texte est illustrée par la figure 15. Le premier texte a un vocabulaire de 38 éléments (alors que la version précédente mettait un texte d'un vocabulaire de 71 éléments). Le pic du troisième texte (vocabulaire de 48 éléments) s'explique par la longueur du texte et celui du texte numéro 151 par la présence de noms de lieux. Le nombre de textes sans nouveau vocabulaire a diminué légèrement (8 textes, contre 11 précédemment) et le nombre de textes avec un seul vocable nouveau a aussi diminué (le nombre de textes est passé de 19 à 15). Les courbes des figures 16 et 17 illustrent l'effet du changement de stratégie : la longueur de texte n'évolue plus en fonction de la séquence de lecture et la couverture textuelle est améliorée, pour dépasser le seuil de 95% à partir du texte numéro 13. 106 des 156 textes ont une couverture textuelle d'au moins 95% par la liste de vocabulaire connu.

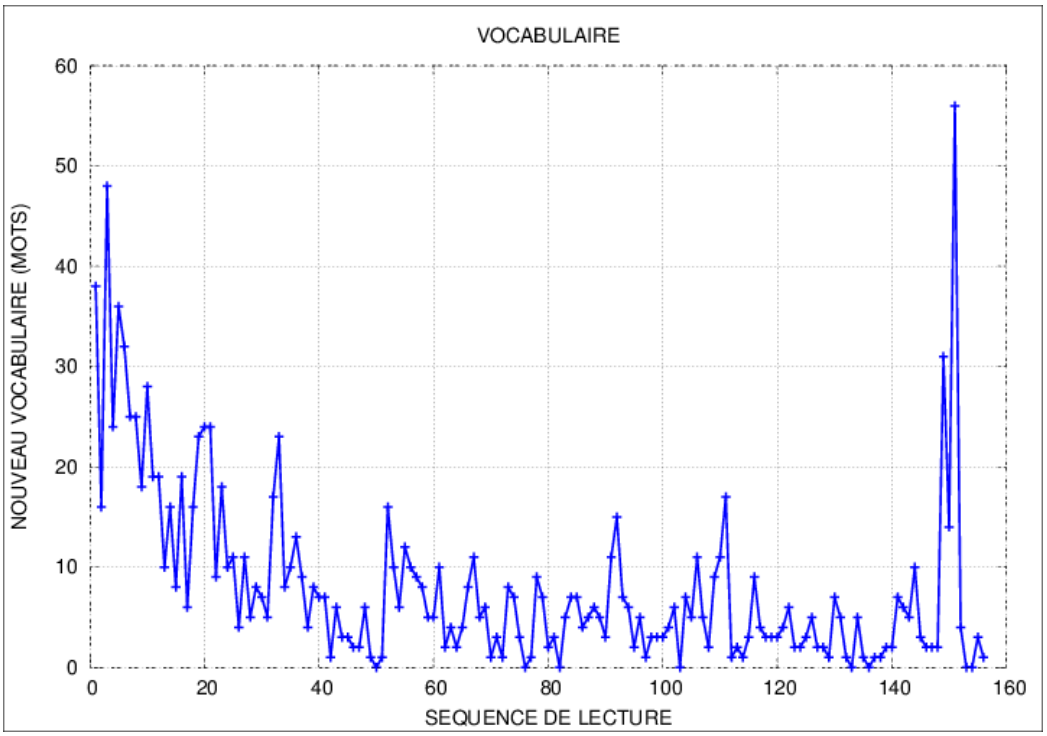



*Figure 15 Tri par couverture textuelle - évolution du nouveau vocabulaire cible (corpus Business de 300 textes)*

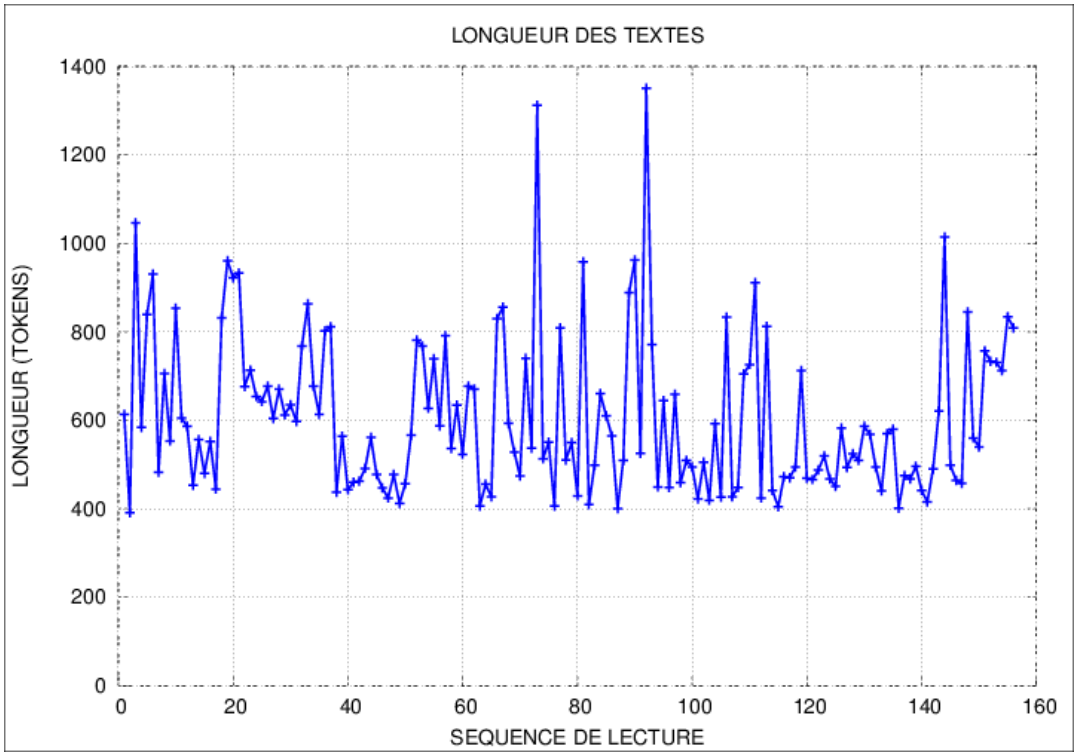



*Figure 16 Tri par couverture textuelle - longueur de texte (corpus Business de 300 textes)*

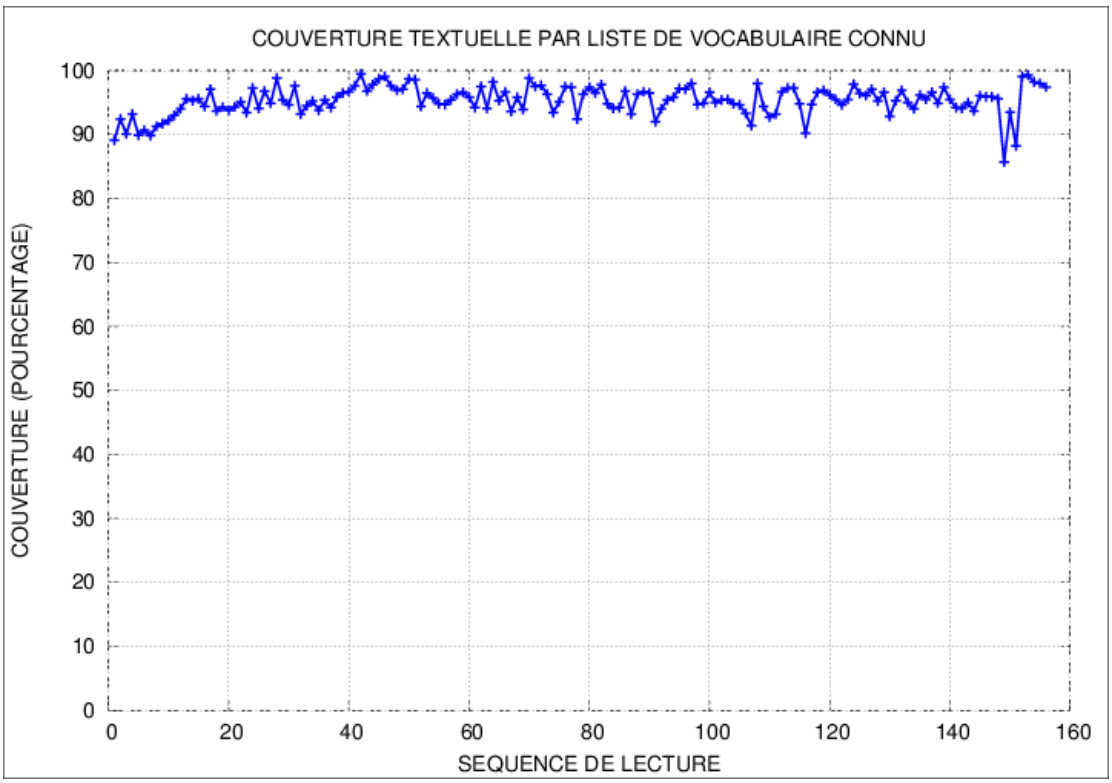



*Figure 17 Tri par couverture textuelle - couverture textuelle (corpus Business de 300 textes)*

## 5.4. Homogénéité de corpus

Pour illustrer un classement de textes plus hétérogènes, nous avons testé *ThaiTextLadder* sur un corpus de 300 articles de *ThaiRath2013* pris dans diverses rubriques (*Business*, *Divertissement*, *Sport*, *Lifestyle*, *National* et *International*).

Nous constatons d'abord que des 300 textes, seuls 20 sont couverts à 95% avec les listes de vocabulaire initiales. Ceci est dû au fait que la liste de vocabulaire spécifique créée automatiquement pour un corpus hétérogène est trop petite. La liste initiale ce vocabulaire spécifique à notre corpus de textes de sources mixtes (801 vocables) est réduite de moitié par les critères de fréquence minimum et nombre de textes minimum, alors que la liste initiale du corpus constitué de 300 articles tirés exclusivement de la rubrique *Business*, initialement de 1115 vocables, atteint toujours 723 vocables après application de ces critères.

Le nombre de vocables hapax par rapport au vocabulaire répété dans ces deux corpus est illustré par les figures 18 et 19. La figure 18 montre la distribution des répétitions de vocabulaire d'une séquence de lecture composée de 300 articles de la seule rubrique *Business* (le dernier groupe comprend neuf occurrences et plus) et la figure 19, une séquence de lecture composée de 300 articles de six rubriques différentes. Pour cette dernière, nous avons réduit l'exigence de couverture des listes initiales à 85% pour disposer de plus d'articles (191 articles au lieu de 20 pour la couverture de 95%). On voit une plus grande proportion d'hapax dans le corpus plus hétérogène, mais il faut noter qu'il s'agit en grande partie d'entités nommées. L'effet est encore plus marqué avec une séquence de lecture basée sur le corpus didactique *GLOSS* (figure 20) qui est composé de textes de thématiques très différents.

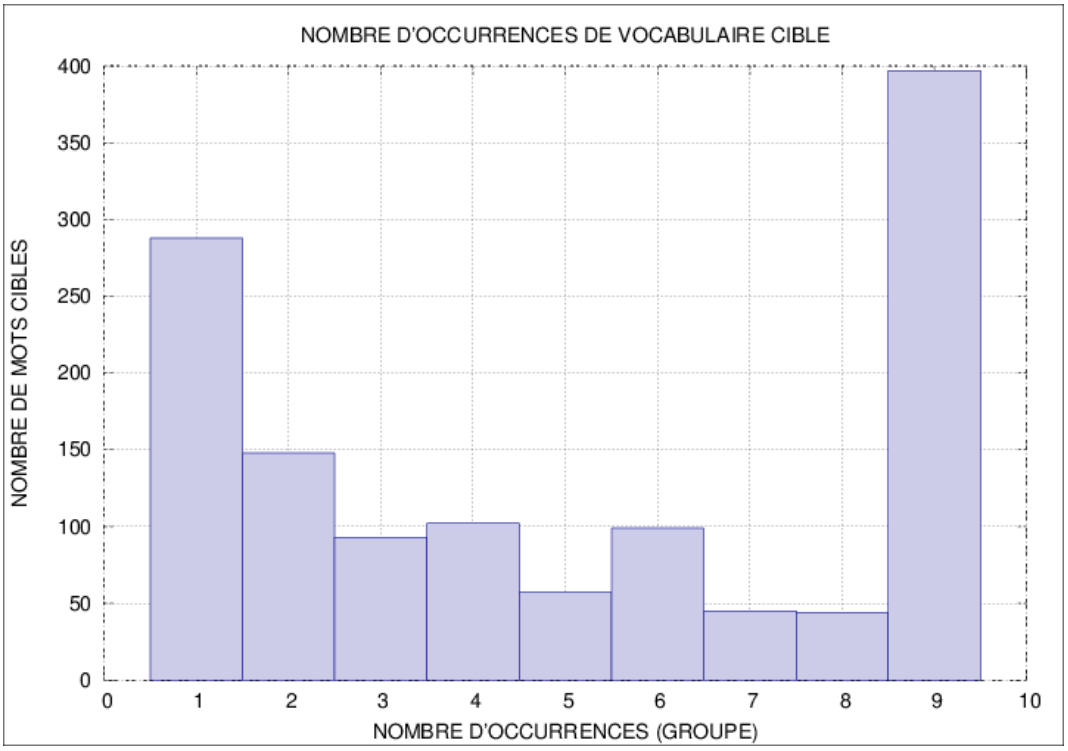



*Figure 18 Distribution des répétitions de vocabulaire (corpus Business de 300 textes)*

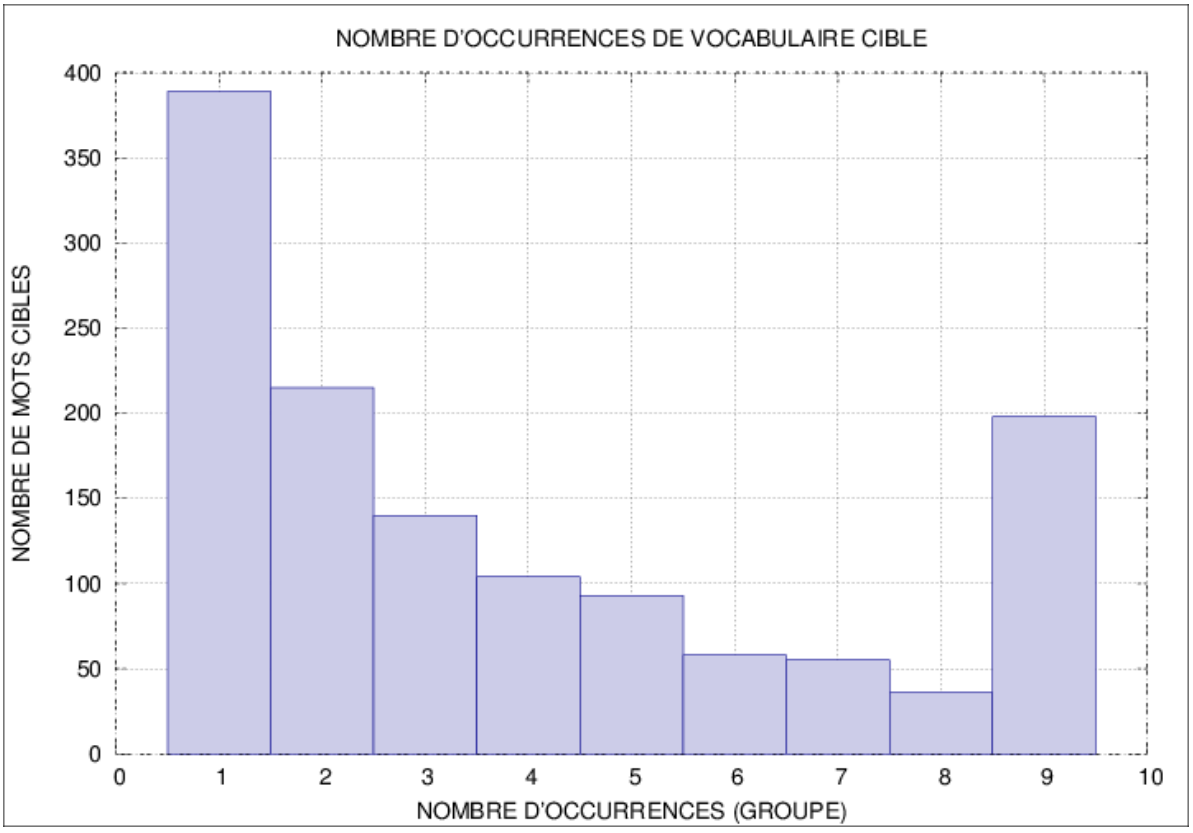



*Figure 19 Distribution des répétitions de vocabulaire (corpus de six rubriques différentes de ThaiRath2013 de 300 textes)*

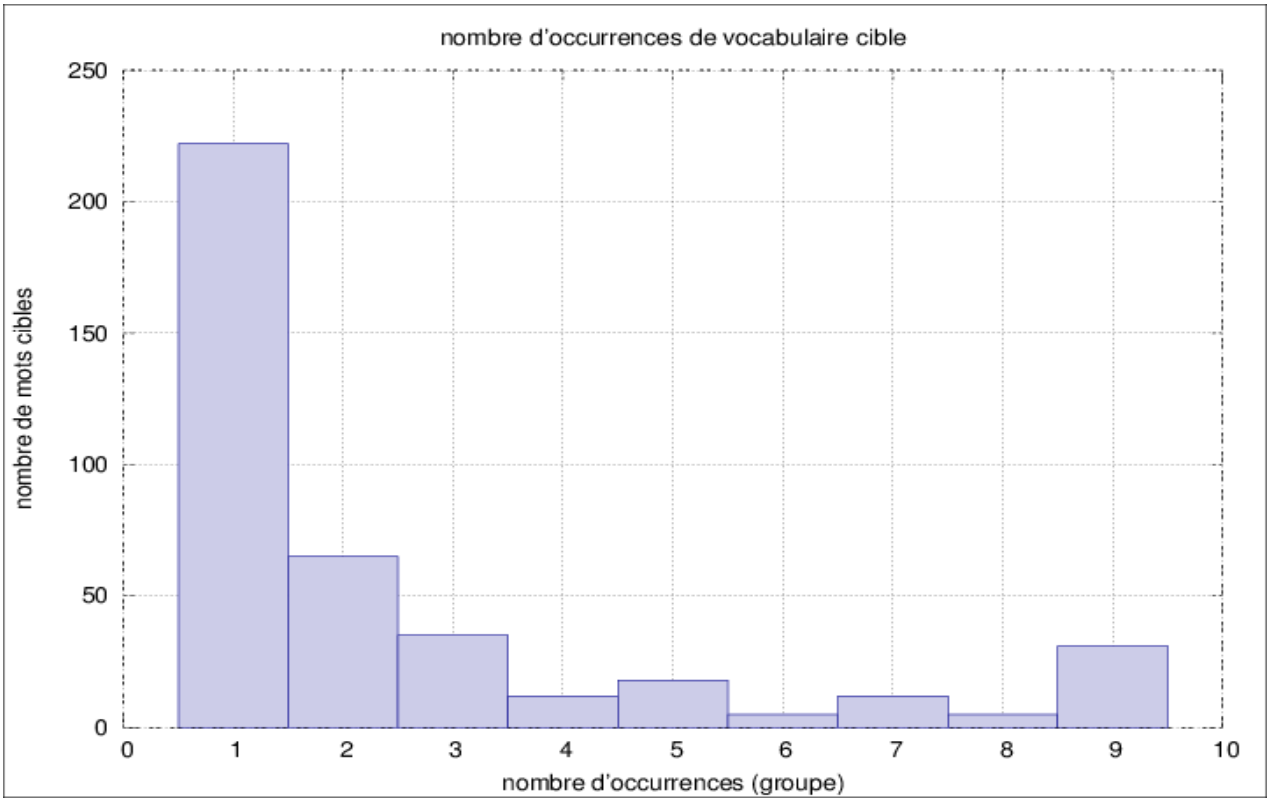



*Figure 20 Distribution des répétitions de vocabulaire (corpus GLOSS de 53 textes)*

## 5.5. Test sur corpus de nouvelles

La littérature étant souvent un composant important des études de langues étrangères, il est important de considérer des corpus littéraires. Notre corpus de nouvelles, bien que composé de textes écrits de mains différentes et comportant un nombre considérable de mots hapax, une fois classifié par *ThaiTextLadder* a révélé une proportion significative de vocabulaire répété. Si l'on regarde la figure 9 qui illustre la courbe de couverture textuelle par taille de liste *ThaiWaC* pour chaque corpus, nous constatons qu'une plus grande proportion du vocabulaire du corpus de nouvelles se situe parmi le vocabulaire de haute fréquence par rapport au corpus de presse. Il semble donc que même si la lecture de ces textes demande plus d'effort par texte (plus de nouveau vocabulaire, textes plus longs), leur lecture soit plus rentable pour l'étudiant dans son apprentissage de la langue en général.

Le programme a été modifié pour prendre en compte le fait que les noms de personnages de fiction ont une fréquence élevée à l'intérieur d'une nouvelle, mais ne sont presque jamais partagés par les textes du corpus (liste de vocabulaire spécifique plus petite et donc couverture textuelle minimum baissée à 90%). La plupart des nouvelles de notre corpus ayant une longueur entre 500 et 3000 mots, les exigences de longueur de *ThaiTextLadder* ont été aussi modifiées en conséquence pour les besoins de ce test. Ces précautions ont permis d'effectuer une classification de 50 des 80 nouvelles du corpus initial.

Nous constatons d'abord que la répétition de vocabulaire (figure 21) n'est pas aussi marquée que pour le corpus *Business* que nous venons de voir et que le nombre de répétitions n'est pas aussi élevé. Ceci s'explique en partie par le nombre de textes (156 textes classifiés pour le corpus *Business* contre 50 pour le corpus de nouvelles).

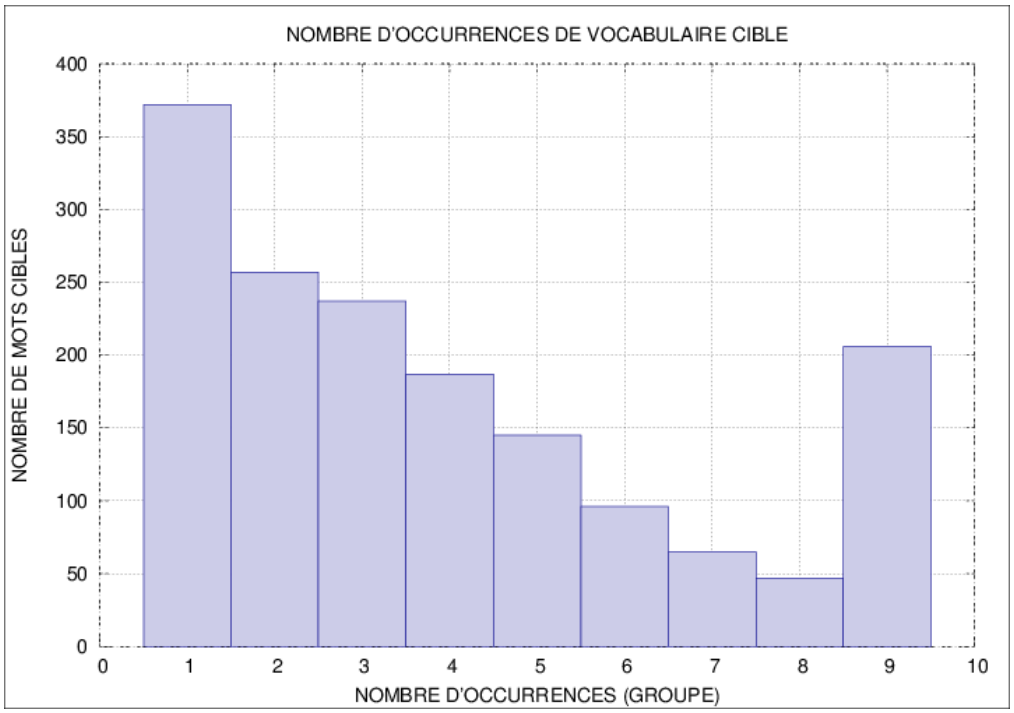



*Figure 21 Distribution de répétitions de vocabulaire (corpus de nouvelles)*

Nous voyons une réduction très marquée du nouveau vocabulaire cible au cours de la séquence de lecture (figure 22).

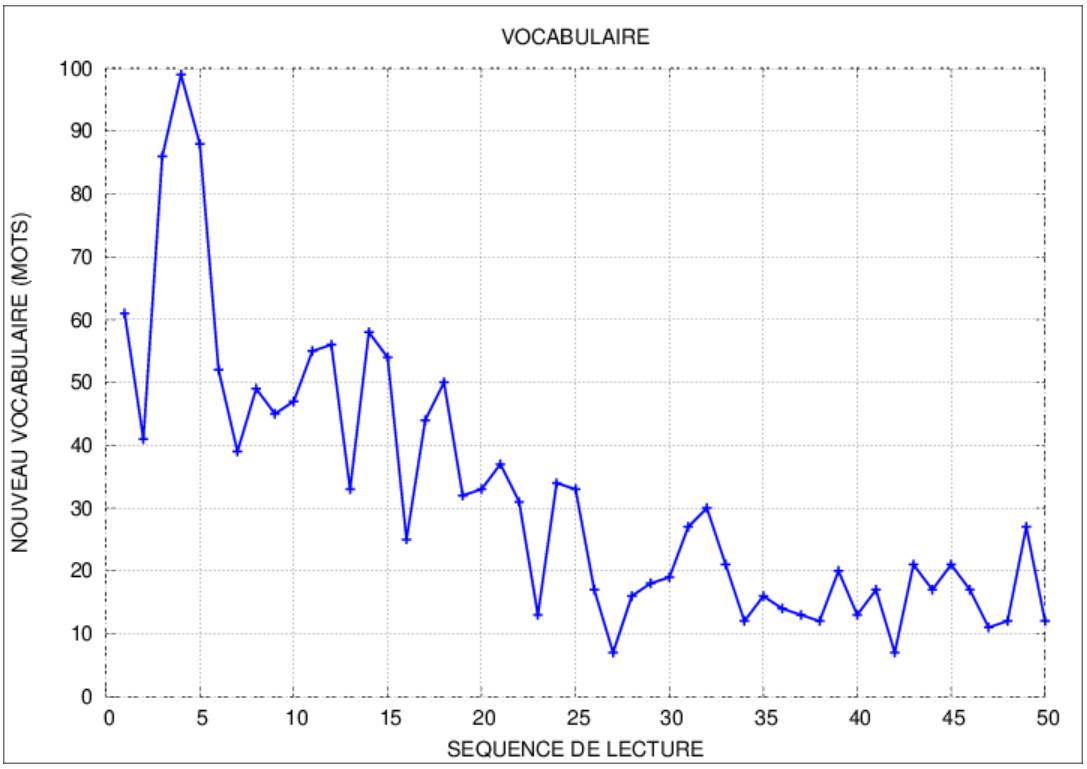



*Figure 22 Évolution du nouveau vocabulaire (corpus de nouvelles)*

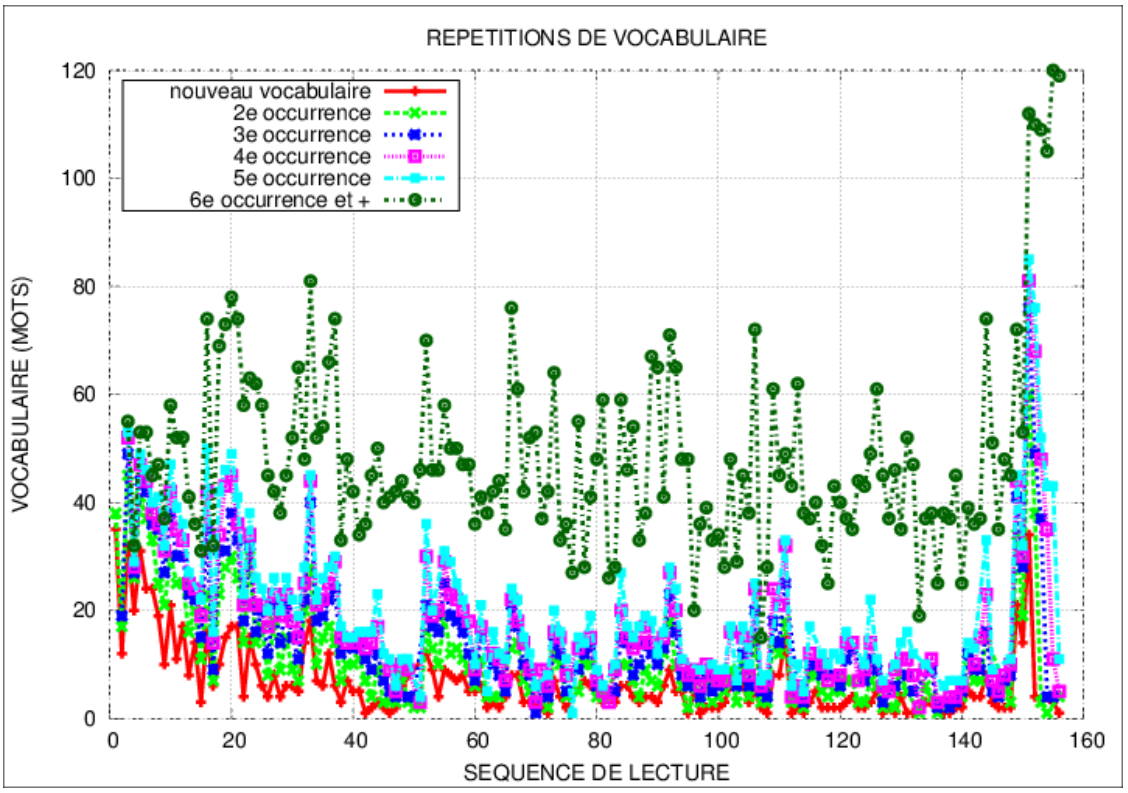



*Figure 23 Répétitions de vocabulaire nouveau (corpus de nouvelles)*

L'examen du contenu de ces textes révèle un nombre important de répétitions de vocabulaire, à la fois de vocabulaire répété à l'intérieur des textes qu'entre les textes. La figure 23 permet d'apprécier la quantité de mots cibles nouveaux (en rouge) par rapport à la quantité de mots cibles répétés (les courbes représentent des fréquences cumulatives). Il nous semble que l'on pourrait améliorer le classement en défavorisant un texte avec beaucoup de mots nouveaux et peu de mots répétés, comme le texte numéro 7, par exemple.

Nous constatons à partir des courbes de répétitions de vocabulaire (figure 23) et la courbe qui représente la longueur de textes (figure 24) que les pics et les dépressions de la courbe du nombre total de mots répétés d'un texte (en vert foncé) ressemblent à ceux de la courbe qui représente la longueur de textes. Cet effet semble moins marqué dans le corpus *Business* (figures 25 et 26), dont les textes ont des longueurs plus similaires. Un examen du contenu des articles du corpus *Business* nous révèle plus de formes répétées entre les textes qu'à l'intérieur des textes.

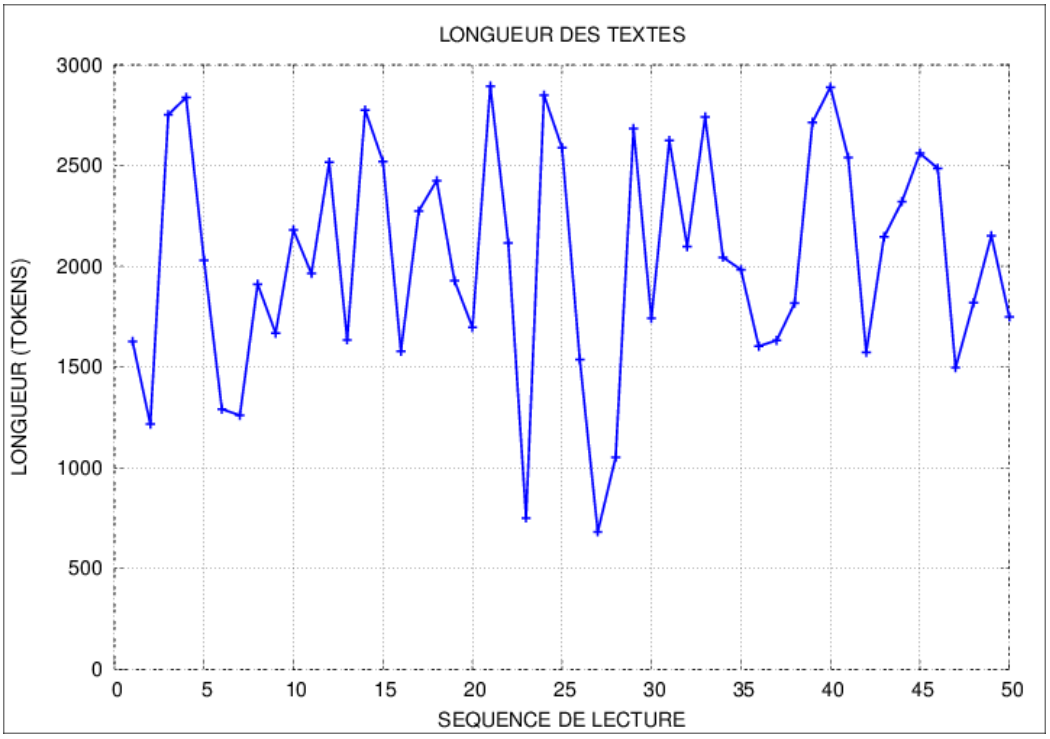



*Figure 25 Longueur de texte (corpus de nouvelles)*

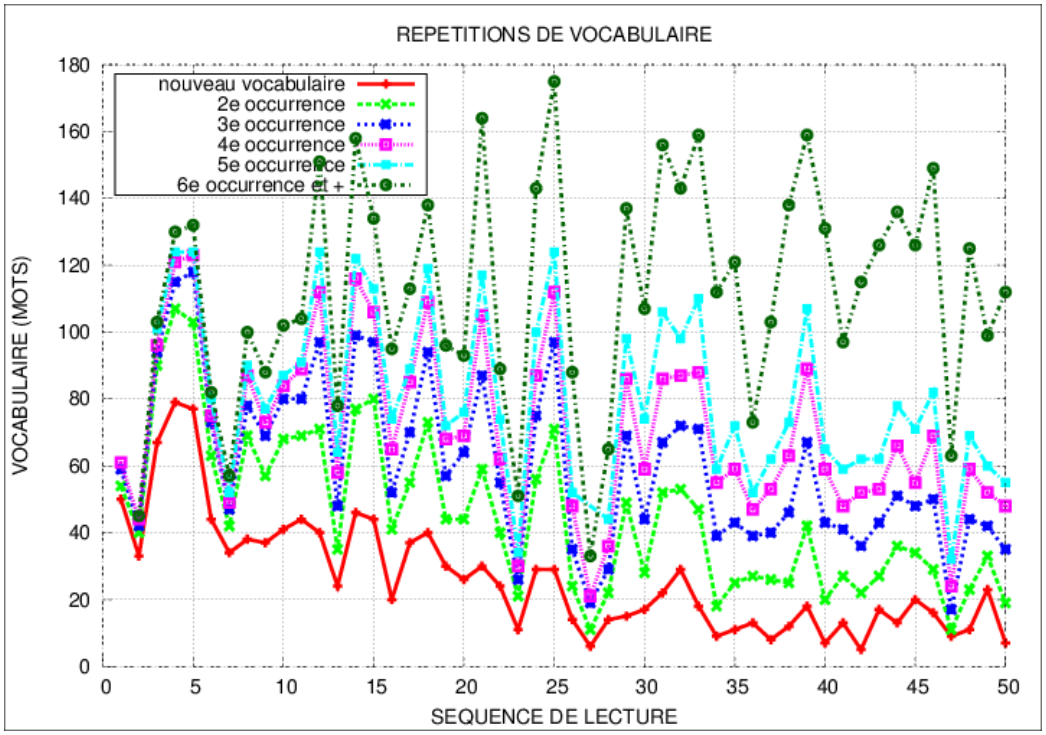



*Figure 24 Répétitions de vocabulaire nouveau (corpus Business de 300 textes)*

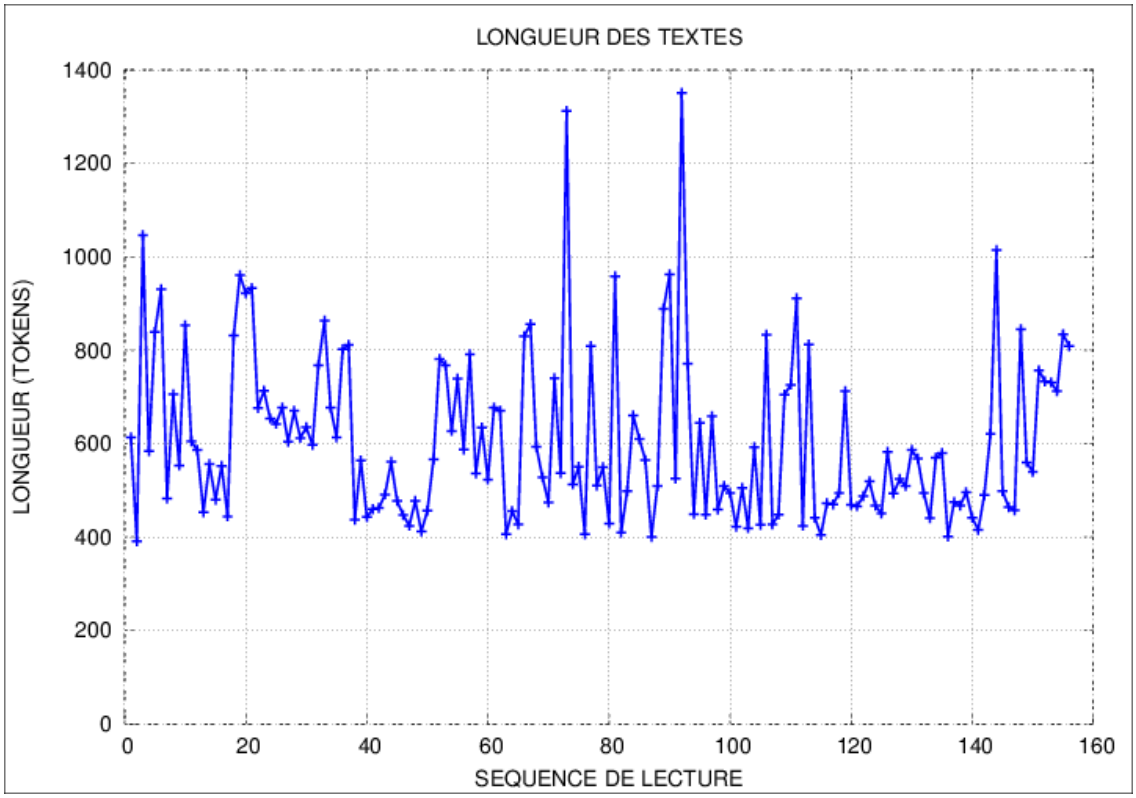
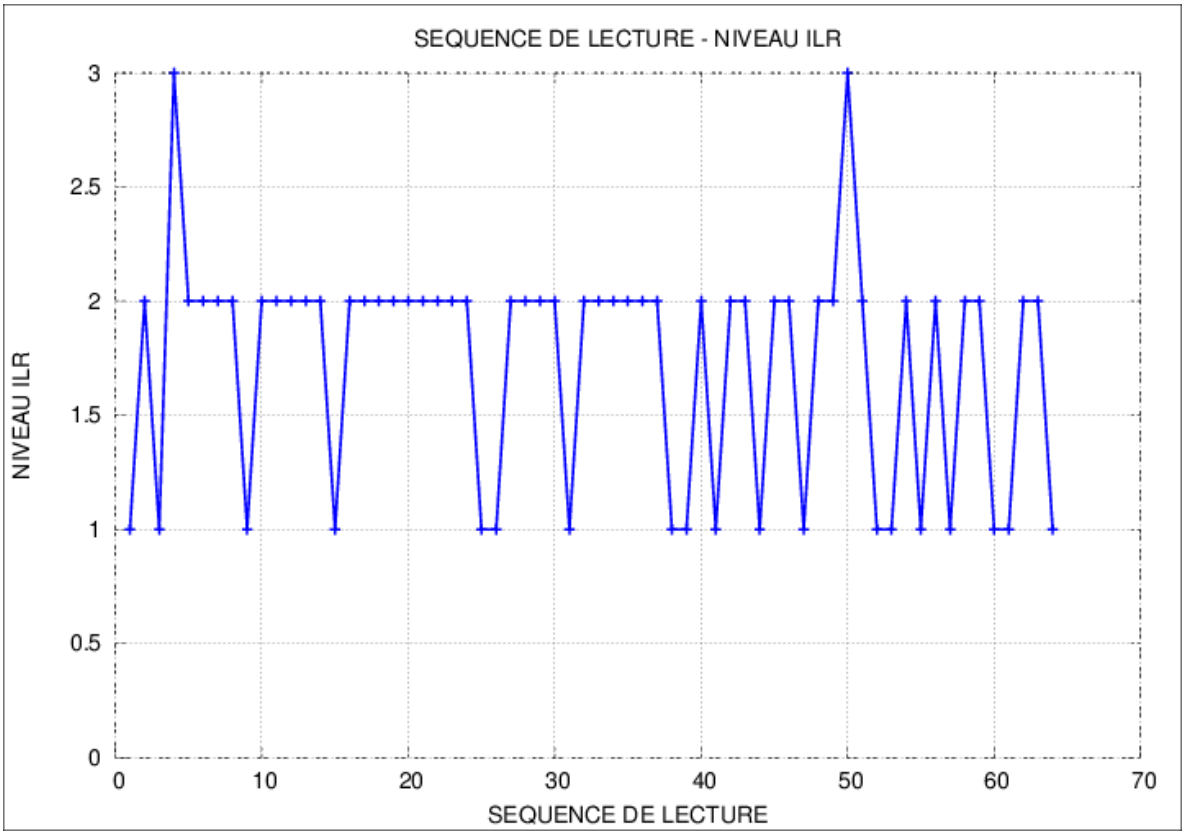



*Figure 26 Longueur de texte (corpus Business de 300 textes)*

## 5.6. Séquence de lecture et niveau de compétence standardisé

Nous savons déjà que les échelles de niveaux de compétence standardisées sont basées sur les compétences communicatives, qui ne tiennent pas compte de la fréquence lexicale de manière systématique. Cette disparité est illustrée par la figure 27 qui montre les niveaux ILR des textes d'une séquence de lecture des textes du corpus *GLOSS* générée par *ThaiTextLadder* (1.5 correspond au niveau 1+, 2.5 au niveau 2+). Ce résultat ne s'explique pas par la diversité de sujets des leçons, car si l'on ne prend que les 25 leçons se portant sur les sujets de la société, la tendance est la même (figure 28).




,

*Figure 27 Séquence de lecture et niveau ILR (corpus GLOSS entier)*

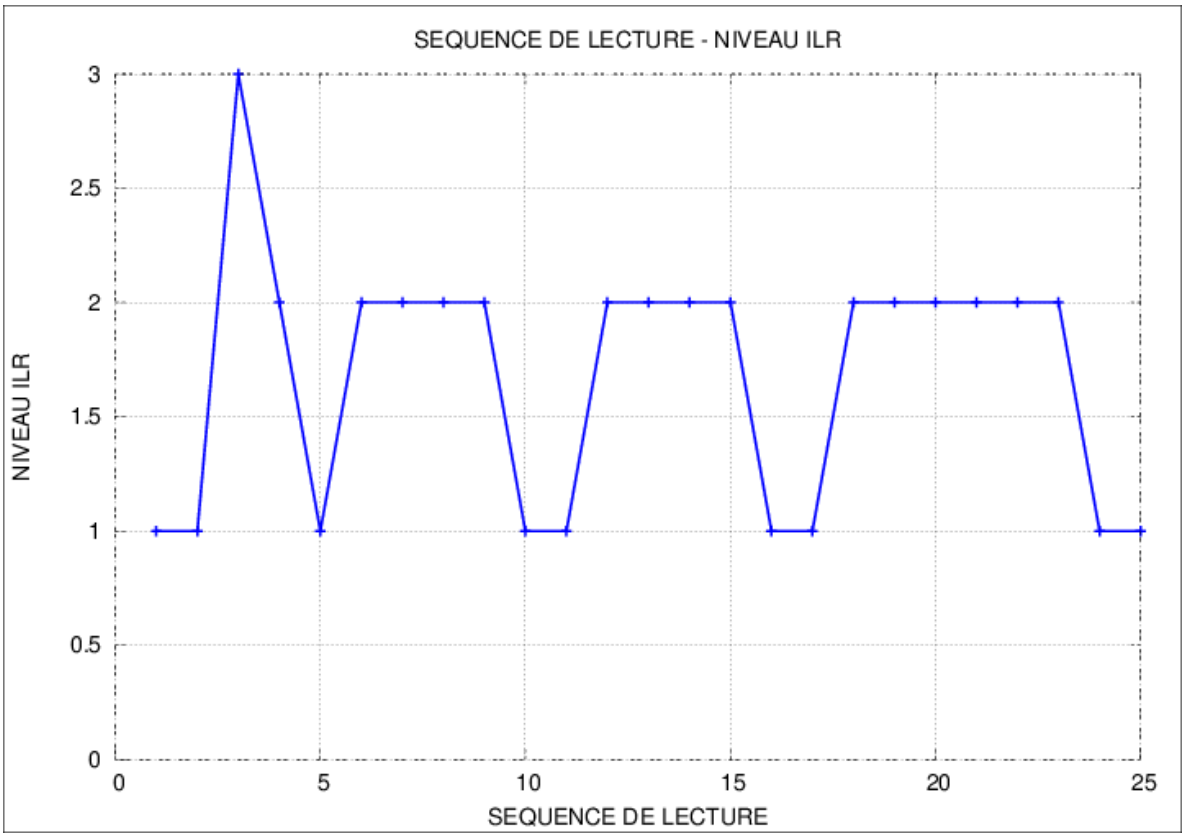



*Figure 28 Séquence de lecture et niveau ILR (textes société du corpus GLOSS)*

# 6. Discussion

La présente étude possède un certain nombre de limites méthodologiques qu'il convient de prendre en compte dans une discussion sur les résultats que nous avons obtenus.

Tout d'abord, la connaissance lexicale ne suffit pas comme critère d'évaluation de la difficulté d'un texte, d'autres facteurs tels que la complexité syntaxique, les références culturelles ou la cohérence discursive contribue à la difficulté d'un texte (Cembreros Castaño, 2010). Heilman et coll. (2007) indiquent que les aspects grammaticaux contribueraient plus à la difficulté d'un texte pour le lecteur en L2 que le lecteur en L1. En ce qui concerne la connaissance lexicale, notre utilisation de listes de vocabulaire connu et cible ne tient pas compte de la complexité de la connaissance lexicale en L2. Les chercheurs s'accordent généralement à dire que celle-ci se situe dans un continuum partant de compétences en réception aux compétences en production (Nation, 2001). En effet, notre système reflète une vision assez simpliste du vocabulaire, car il ne prend pas en compte la polysémie des lexèmes, exagérée pour des lexèmes d'une langue isolante dépourvus de repères que donne la syntaxe. À titre d'exemple, Conjeaud (2009) fournit un tableau des 24 hypothèses avancées sur la valeur de la seule particule จะ /cà/ en thaï. Nous ne tenons pas compte non plus du fait que l'apprenant peut acquérir les différents sens d'un lexème à des niveaux différents. Ghadirian (2002) signale lui-même que sa méthode n'analyse pas les collocations, ou les verbes à particule (*phrasal verbs*). Par ailleurs, Schmitt (2010) souligne l'importance, en L2 comme en L1, de l'utilisation intensive d'expressions figées qui constituent une grande partie du discours, car c'est la maîtrise de ces expressions qui facilite la production du langage et permet au locuteur de gagner en aisance et en fluidité. Dans les manuels de langue, les expressions idiomatiques dont le sens n'est pas déchiffrable à partir de ses composants figurent souvent dans les listes de vocabulaire, mais les expressions figées déchiffrables ne sont jamais indiquées.

Avant de réfléchir aux modifications possibles à *TextLadder*, il faudrait définir l'objectif de la lecture. S'il s'agit d'une lecture *intensive* destinée à la « capitalisation » lexicale ou syntaxique ou une pratique *extensive* de la lecture, pour la compréhension du contenu et le plaisir de découvrir du texte authentique. En effet, Ghadirian (2002) prévoit un usage de *TextLadder* pour la lecture intensive des textes avec une quantité importante de nouveau vocabulaire, alors que Huang & Liou (2007) ont utilisé cette méthode de classification de textes pour encourager la lecture extensive. Dans le cas de la lecture extensive, est-il nécessaire d'augmenter et cibler les expressions idiomatiques et les collocations ? Ne serait-il pas préférable de laisser le lecteur identifier les cas de polysémie ?

L'utilisation de la méthode *TextLadder* pour créer du matériel pédagogique de lecture intensive pourrait bénéficier d'éventuelles modifications au logiciel. Il est possible de différencier entre certains cas de polysémie par un traitement d'étiquetage morphosyntaxique, mais le traitement de collocations et expressions idiomatiques est plus fastidieux. En thaï, il serait possible de les rajouter tout simplement au dictionnaire de segmentation, qui aurait l'effet de ne pas prendre en compte les composants de l'expression dans le calcul de répétitions. Un calcul de similarité tenant compte de répétitions d'expressions figées décomposables ne serait pas utile pour une séquence limitée à un certain nombre de textes, mais l'indication de ces expressions nous semble très utile. Leur identification demanderait un travail d'analyse de corpus considérable.

Les entités nommées et mots anglais écrits en thaï nous ont donné beaucoup de matière à réflexion, car il est difficile de situer les connaissances de l'apprenant sur un continuum de compétence lexicale, car tout dépend des connaissances extralinguistiques du monde réel de l'apprenant, et cela peut varier considérablement. Il n'est pas sûr qu'un étudiant chinois de thaï, par exemple, va reconnaître le nom d'une société qui a un nom complètement différent en chinois, comme par exemple ไมโครซอฟท์ /majkʰroːsɔ̂ːf/ qui est connu sous le nom 微软 /wēi ruǎn/ en Chine, une traduction des éléments *Micro* + *soft*. La présence d'entités nommées en quantité, comme dans notre corpus de presse d'articles de la rubrique *Sport* de *ThaiRath2013*, si ces unités sont considérées comme des mots ordinaires, a l'effet d'augmenter la quantité de vocabulaire de basse fréquence. Cet effet est illustré par la figure 4.1, où la courbe de couverture de textes de la rubrique sport est beaucoup plus basse que les autres en vertu de la présence d'entités nommées. En réalité, une fois ces mots identifiés comme étant des entités nommées, ils sont plus faciles à interpréter que d'autres mots inconnus.

Nous nous posons aussi des questions par rapport aux mots anglais écrits en thaï. Peut-on considérer qu'un apprenant non natif d'anglais possède un niveau suffisant d'anglais pour pouvoir déchiffrer les mots anglais qu'il rencontre en écriture thaïe ? Ce déchiffrage de mots anglais écrits en lettres thaï constitue une compétence en soi. Dans certaines moutures de *ThaiTextLadder*, nous avons rajouté des listes d'entités nommées internationales à la liste de vocabulaire connu et les entités nommées thaïes à la liste de vocabulaire cible. Tout comme Ghadirian (2002), nous pensons que les traitements différents d'entités nommées doivent être proposés à l'utilisateur du logiciel en option.

D'autres problèmes méthodologiques sont survenus à cause de nos outils et ressources.

Nous avons déjà évoqué la question des erreurs de segmentation. Malgré le fait que nous ayons pris la précaution d'utiliser le même outil de segmentation pour le traitement de nos corpus que celui utilisé dans le traitement du corpus *ThaiWaC* sur lequel nos listes de vocabulaire connu et cible sont basées, nous ne savons pas s'ils ont utilisé le même algorithme de segmentation. *SWATH* propose plusieurs algorithmes et malencontreusement, l'algorithme utilisé n'est précisé ni par le service REST de segmentation par *SWATH* (Poltree & Saikaew, 2011) que nous avons utilisé pour segmenter nos corpus ni sur le site qui héberge le corpus *ThaiWaC, Sketch Engine*. Ceci expliquerait certains cas de mots inconnus de nos listes de vocabulaire (et de vocables de nos listes inutilisés). L'inclusion d'un outil de segmentation dans *ThaiTextLadder* nous parait indispensable, car il permettrait non seulement de définir l'algorithme, mais aussi d'améliorer le dictionnaire de segmentation en temps réel. De plus, l'utilisation du même outil avec le même algorithme et le même dictionnaire pour élaborer une liste de fréquence lexicale basée sur un corpus général de la langue devrait garantir l'uniformité d'identification du vocabulaire. Nous avons trouvé que nos textes de presse contiennent un certain nombre de mots mal orthographiés ou des variantes graphiques de noms de personnes[105] qu'il faudrait corriger à la main et/ou ajouter au dictionnaire interne de l'outil de segmentation. L'amélioration du dictionnaire interne à l'outil de segmentation devrait réduire le nombre d'erreurs constatées dans la liste de vocabulaire spécifique au corpus créée automatiquement.

[105] Rappelons que le système d'écriture thaïe contient beaucoup plus de signes que de phonèmes. En outre, la plupart de noms de personnes étrangers n'ont pas de tons, ce qui augmente les possibilités. Par exemple la lettre latine « s » peut se transcrire ช ษ ศ ou ส.

En l'état, l'introduction d'erreurs dans la liste de vocabulaire cible via la création de la liste de vocabulaire spécifique constitue un sérieux défaut du système.

Nos listes lexicales générées à partir du corpus *ThaiWaC* ne prennent en compte que les fréquences simples de ce corpus web et ne tiennent compte ni de la distribution de lexèmes ni de leur répartition parmi les textes. Il nous semble que des listes élaborées à partir d'un corpus web ne peuvent représenter avec exactitude les connaissances d'un apprenant en L2, ne serait-ce que parce qu'il ne prend pas en compte la langue orale. L'inclusion de la langue orale dans un corpus représentatif du thaï destiné à l'élaboration des listes de fréquence lexicale serait très importante, car selon Jantharat (2012) les méthodes d'enseignement du thaï se focalisent particulièrement sur les compétences orales. D'une manière plus générale, l'élaboration d'un tel corpus devrait prendre en compte les différents niveaux de langue en thaï.

Les données du corpus *ThaiWaC* contiennent aussi des mots et entités nommées écrits en lettres latines. Pour les apprenants en L2 de thaï possédant un certain niveau d'anglais, il est certain que ces fréquences ne représentent pas ses compétences lexicales réelles. Le tableau ci-dessous démontre ce décalage entre les classements de la liste de fréquence lexicale *ThaiWaC* et la *New General Service List* (*NGSL*)[106], liste de fréquence lexicale pour l'anglais, pour trois mots en lettres latines trouvés dans la liste *ThaiWaC*. On voit qu'un mot de base anglais comme *how*, classé en 93e position dans la *NGSL* est classé en 9852e position dans la liste de fréquence générée du corpus *ThaiWaC*.

*Tableau 9 Comparaison de classement de mots anglais de la liste de fréquence ThaiWaC et de la NGSL*

| | **ThaiWaC** | **NGSL** |
|---|---|---|
| *information* | 9968 | 213 |
| *author* | 9946 | 909 |
| *how* | 9852 | 93 |

Pour nous simplifier la tâche, nous avons fait le choix de considérer tous les mots en lettres latines comme du vocabulaire connu, mais il est possible qu'une approche qui prend en compte les connaissances réelles de l'utilisateur soit plus appropriée.

Comme nous avons vu, notre classement ne correspond en aucune manière aux niveaux de compétence définis par l'échelle standardisée ILR. Nous avons aussi constaté que certains éléments de vocabulaire de base comme les mois et les couleurs manquent à notre liste de vocabulaire cible. Il serait souhaitable de redéfinir les listes de vocabulaire connu et cible, non seulement à partir de listes de fréquence lexicale basées sur un corpus plus représentatif, mais aussi prenant en compte le contenu des manuels et matériels pédagogiques existants déjà classés par niveau.

Le dernier aspect que nous voulons signaler concerne le vocabulaire cible nouveau de chaque texte sur la séquence de lecture. Comme nous avons vu précédemment (sections 5.3 et

[106] La *New General Service List* de Browne et coll. (2013) est une mise à jour de la *General Service List* (*GSL*) de West (1953), la première liste de fréquence lexicale élaborée pour la langue anglaise.

5.2), une séquence générée d'un corpus d'articles de presse propose un certain nombre de textes sans vocabulaire nouveau. Peu ou pas de vocabulaire nouveau pourrait être désirable, tout dépend de l'usage que l'on fait de la séquence. S'il s'agit d'un lecteur autonome, un décroissement uniforme à travers les textes comme dans la première version de *ThaiTextLadder* peut s'avérer motivant, les textes suivants constituant des textes de révision en quelque sorte. Si par contre un professeur de langue voulait générer une séquence de lecture pour un programme de cours, ou un manuel de lecture, une option de maîtrise de la quantité du nouveau vocabulaire par texte serait très utile. Même si notre programme favorise globalement la répétition du vocabulaire connu, nous sommes aussi conscients du fait que le système ne tient pas compte spécifiquement des intervalles de répétition du vocabulaire cible. Une amélioration de l'algorithme qui spécifie un minimum (et maximum) taille de vocabulaire cible qui prendrait en compte les intervalles entre les répétitions est à prévoir.

En dépit de ces limitations méthodologiques, ce travail ouvre des perspectives pour d'autres langues. L'utilisation de listes de vocabulaire connu et cible est intéressante dans la mesure où ces listes ne doivent pas être obligatoirement élaborées à partir de listes de fréquence lexicale basée sur corpus, mais peuvent se créer à partir de textes déjà lus et de vocabulaire cible défini par l'enseignant. Des modifications sont nécessaires pour intégrer des langues à morphologie plus complexe, qui demandent un traitement de lemmatisation. Un système sans lemmatisation intégrant des listes de fréquence lexicale basées sur un corpus pourrait s'avérer intéressant, car ceci introduirait les éléments de difficulté grammaticale dans la sélection de textes.

# 7. Conclusion

La création automatique de ressources pour l'apprentissage de langues étrangères peu enseignées et peu dotées en matériels pédagogiques est particulièrement séduisante. Toutefois, en ce qui concerne notre adaptation de *TextLadder*, logiciel de classification automatique de textes en anglais, à la langue thaïe, nous avons trouvé qu'il est nécessaire de prendre en compte les particularités de la langue. Notre tentative d'adapter *TextLadder* aux textes en thaï nous a permis d'élucider un nombre de précautions méthodologiques qu'il faut prendre en considération. Certaines de ces précautions s'appliquent spécifiquement au thaï, d'autres sont d'ordre plus général et peuvent s'appliquer à l'adaptation de *TextLadder* à d'autres langues.

Bien que la majeure partie de la recherche sur l'acquisition du vocabulaire en L2 soit focalisée sur l'acquisition de l'anglais, et nous sommes conscient que certaines de nos prémisses méthodologiques basées sur cette recherche devraient être validées par la recherche sur d'autres langues, les résultats que nous avons obtenus nous permettent de démontrer qu'il est possible de créer des ressources pédagogiques pour la lecture de façon automatique avec peu d'intervention humaine.

La classification automatique de textes par la progression lexicale constitue une méthode intéressante pour proposer une séquence de textes appropriée au niveau d'un lecteur en L2, surtout en ce qui concerne les textes qui portent sur une thématique particulière, que ce soit pour proposer des textes à des lecteurs autonomes, ou pour la création de matériels pédagogiques destinés à être utilisés en classe. De telles méthodes ont l'avantage de baser les matériels pédagogiques sur des données de connaissances réelles des étudiants plutôt que sur la seule intuition des enseignants.

# 8. Table de figures

Figure 1 La forme de base de la syllabe écrite en thaï ..... 14
Figure 2 Échantillon de texte en thaï avec sa traduction en anglais ..... 17
Figure 3 *SEAlang Lab Reader's Helper* ..... 23
Figure 4 *Scola* - Extraction automatique d'un article de presse ..... 25
Figure 5 *Readlang* avec un texte thaï ..... 26
Figure 6 *AideMoi* en cours de développement ..... 27
Figure 7 *Thai Text Reader* avec dictionnaire web intégré ..... 28
Figure 8 Exemple de segmentation résultant d'un traitement avec *SWATH* ..... 38
Figure 9 Courbe de couverture textuelle par taille de liste *ThaiWaC* pour chaque corpus .... 42
Figure 10 Balisage du vocabulaire cible par *ThaiTextLadder* ..... 46
Figure 11 Fonctionnement de *ThaiTextLadder* ..... 47
Figure 12 Tri par quantité de vocabulaire connu - évolution du nouveau vocabulaire cible (corpus *Business* de 300 textes) ..... 50
Figure 13 Tri par quantité de vocabulaire connu - longueur de texte (corpus *Business* de 300 textes) ..... 50
Figure 14 Tri par quantité de vocabulaire connu - couverture textuelle (corpus *Business* de 300 textes) ..... 51
Figure 15 Tri par couverture textuelle - évolution du nouveau vocabulaire cible (corpus *Business* de 300 textes) ..... 52
Figure 16 Tri par couverture textuelle - longueur de texte (corpus *Business* de 300 textes) 52
Figure 17 Tri par couverture textuelle - couverture textuelle (corpus *Business* de 300 textes) ..... 53
Figure 18 Distribution des répétitions de vocabulaire (corpus *Business* de 300 textes) ..... 54
Figure 19 Distribution des répétitions de vocabulaire (corpus de six rubriques différentes de *ThaiRath2013* de 300 textes) ..... 54
Figure 20 Distribution des répétitions de vocabulaire (corpus *GLOSS* de 53 textes) ..... 55
Figure 21 Distribution de répétitions de vocabulaire (corpus de nouvelles) ..... 56
Figure 22 Évolution du nouveau vocabulaire (corpus de nouvelles) ..... 56
Figure 23 Répétitions de vocabulaire nouveau (corpus de nouvelles) ..... 57
Figure 25 Répétitions de vocabulaire nouveau (corpus *Business* de 300 textes) ..... 58
Figure 24 Longueur de texte (corpus de nouvelles) ..... 58
Figure 26 Longueur de texte (corpus *Business* de 300 textes) ..... 59
Figure 27 Séquence de lecture et niveau ILR (corpus *GLOSS* entier) ..... 60
Figure 28 Séquence de lecture et niveau ILR (textes société du corpus *GLOSS*) ..... 60

# 9. Liste des tableaux

Tableau 1 Couverture textuelle des tranches successives de 1000 lemmes dans le corpus Brown ..... 32
Tableau 2 Corpus utilisés dans l'élaboration des dictionnaires de fréquence lexicale Routledge ..... 34
Tableau 3 Les premiers éléments de la liste de fréquence du corpus *ThaiWaC* ..... 36
Tableau 4 Composition du corpus de presse *ThaiRath2013* ..... 39
Tableau 5 Évaluation de la couverture de listes de fréquence lexicale *ThaiWaC* ..... 41
Tableau 6 Couverture de corpus et fréquence lexicale (détail) ..... 42
Tableau 7 Seuils de vocabulaire du *Thai Reader Project* ..... 43
Tableau 8 Couverture d'une liste de vocabulaire spécifique au corpus de presse ..... 45
Tableau 9 Comparaison de classement de mots anglais de la liste de fréquence *ThaiWaC* et de la *NGSL* (anglais) ..... 63

# 10. Bibliographie


Aroonmanakun, Wirote (2002). *Collocation and Thai Word Segmentation*. Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop.

Aroonmanakoon, Wirote (2006). *Basic statistics of Thai*. International Conference : Wisdom and Dynamism of Thai Language and Literature, 10–12 Nov 2006.[107]

Aroonmanakoon, Wirote (2007a). *Creating the Thai National Corpus*. MANUSYA : Journal of Humanities (Special Issue No. 13. 2007).

Aroonmanakun, Wirote (2007b). *Thoughts on Word and Sentence Segmentation in Thai*. Proceedings of the Seventh Symposium on Natural Language Processing, Dec 13-15, 2007, Pattaya, Thailand.[108]

Barry, Coeli. *Rights to Culture: Heritage, Language and Community in Thailand*. Silkworm Books, 2013.

Baptista, J., N. Costa, J. Guerra, M. Zampieri, M. Cabral & N. Mamede (2010). *P-AWL: Academic Word List for Portuguese*. Computational Processing of the Portuguese Language, Lecture Notes in Computer Science, 2010, Volume 6001/2010, 120–123.

Berment, Vincent. *Méthodes pour informatiser les langues et les groupes de langues « peu dotées »*. Thèse de doctorat, Université Joseph-Fourier - Grenoble I, 2004.[109]

Bofman, Teddy & Paul Prez (2008). *Thai Pop Music: Corpus Analysis and Second Language Learning.* Journal of Southeast Asian Language Teaching Volume 14, 2008.[110]

Brown, Jonathan & Maxine Eskenazi (2004). *Retrieval of Authentic Documents for Reader-Specific Lexical Practice*. Proceedings of InSTIL/ICALL Symposium, Venice, Italy. 2004.

Brown, Marie-Hélène. *Lire et écrire le thaï.* Bangkok, Thaïlande : Kuang-kamol, 1991.

Browne, C., Culligan, B. & Phillips, J. (2013). *The New General Service List.* Tirée de http://www.newgeneralservicelist.org.

Browne, C. (2014). *A new general service list: The better mousetrap we've been looking for?* Vocabulary Learning and Instruction, 3 (1), 110.

Cembreros Castaño, Diana. *A Software Design/Method for Predicting Readability for ESL Students*. Mémoire de Master, Universidad Complutense de Madrid, 2010.

Cobb, T. & Horst, M. (2004). *Is there room for an AWL in French?* Publié dans Bogaards, P. & Laufer, B. (éds.), Vocabulary in a Second Language: Selection, acquisition, and testing (pp. 15-38). Amsterdam : John Benjamins.

Conjeaud, Michèle (2009). *De l'usage de ๋ (cà) dans la langue siamoise (ou thaï)*, Colloque international sur les langues d'Asie du Sud-est, Paris, 2009.


[107] http://pioneer.chula.ac.th/~awirote/ling/ThaiStat.pdf
[108] http://pioneer.netserv.chula.ac.th/~awirote/ling/snlp2007-wirote.pdf
[109] https://tel.archives-ouvertes.fr/tel-00006313/document
[110] http://www.seasite.niu.edu/jsealt/Volume2008/JSEALT_08_Teddy_Final%20.pdf

Crossley, Scott A., Tom Cobb, Danielle S. McNamara. (2013). *Comparing count-based and band-based indices of word frequency : Implications for active vocabulary research and pedagogical applications*. System, Volume 41, Issue 4, December 2013.

Coxhead, Averil (2000). *A New Academic Word List*. TESOL Quarterly, 34(2): 213-238.

Dānwiwat, Nanthanā. *The Thai Writing System*. Buske Verlag, 1987.

Dörnyei, Z. (1998). *Motivation in second and foreign language learning*. Language Teaching, 31, 117-135.

DuBay, William H. *The Principles of Readability*. CA: Impact Information, 2004.

Fels, Jacqueline de. *Promotion de la littérature en Thaïlande: vers les prix littéraires (1882-1982)*. Paris : INALCO, 1993.

Ferlus, M. (1999). *Sur l'ancienneté des écritures thaï d'origine indo-khmère*. Publié dans Colloque Georges Cœdès aujourd'hui, Bangkok, 1999.

François T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*, Thèse de doctorat, Université Catholique de Louvain.

Ghadirian, Sina (2002). *Providing controlled exposure to target vocabulary through the screening and arranging of texts*. Language Learning & Technology. Vol.6, No.1, 147-164.[111]

Gillam, Richard. *Unicode Demystified: A Practical Programmer's Guide to the Encoding Standard*. Addison-Wesley Professional, 2003.

Haruechaiyasak, Choochart, Sarawoot Kongyoung & Matthew N. Dailey (2008). *A comparative study on Thai word segmentation approaches*. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008, vol.1, no., pp.125-128, 14-17, May 2008.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). *Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts*. Proceedings of the Human Language Technology Conference. Rochester, NY.

Hoonchamlong, Yuphaphann (2012). *Issues in determining and annotating linguistic information for the corpus-based 100 high frequency Thai words for learners*. SEALS XXII, 2012.

Hoonchamlong, Yuphaphann (2013). *Punctuation and other text-category indicators in written Thai text: Issues and implications for Thai L2 reading instruction*. SEALS XXIII, 2013.

Hu, M., & Nation, I.S.P. (2000). *Vocabulary density and reading comprehension*. Reading in a Foreign Language, 13(1), 403–430.

Huang, Hung-Tzu & Hsien-Chin Liou (2007). *Vocabulary Learning in an Automated Graded Reading Program*. Learning & Technology. Vol.11, No.3, 64-82.[112]

Islam, Zahurul, Alexander Mehler & Rashedur Rahman (2012). *Text Readability Classification of Textbooks of a Low-Resource Language*. Proceedings of the 26th Pacific Asia Conference

[111] http://llt.msu.edu/vol6num1/ghadirian/default.html

[112] http://llt.msu.edu/vol11num3/huangliou/

on Language, Information and Computation, 2012.

Jantharat, Prawet (2012). *Thai Language : Teaching and Trends.* International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, August 17-18, 2012, Bangkok.

Jean, Christian (2009). *Le thaï. De la segmentation aux maux.* Lexicometrica. Numéro spécial - Explorations textométriques. 2009.[113]

Johansson Kokkinakis, Sofie, Emma Sköldberg, Birgit Henriksen, Kari Kinn, Janne Bondi Johannesse (2012). *Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project Fjeld*, R. V. & J. M. Torjusen (red.) (2012): Proceedings of the 15th EURALEX International Congress 7–11 August, 2012, Oslo, Oslo : Department of Linguistics and Scandinavian Studies, University of Oslo. 563-569.

Kasisopa, Benjawan. *Reading Without Spaces Between Words: Eye Movements in Reading Thai*. Thèse de doctorat, University of Western Sydney, 2011.

Kilgarriff, Adam, Siva Reddy, Jan Pomikálek & Avinesh PVS. (2010). *A Corpus Factory for many languages*. Proc. LREC, Malta, 2010.

Koldunova, Ekaterina. *Thai language programs in Russia go from strength to strength*, The Nation, April 9, 2014.[114]

Kosawat, Krit. *Méthodes de segmentation et d'analyse automatique de textes thaïs*. Thèse de doctorat, Université Paris-Est, 2003.

Krashen, Stephen (2004a). *The Power of Reading: Insights from the Research*. Libraries Unlimited, Connecticut & London.

Krashen, Stephen (2004b). *The Case for Narrow Reading*. Language Magazine 3(5):17-19.

Krawtrakul, A. Thumkanon, C., Poovorawan, Y., & Suktarachan, M. (1997). *Automatic Thai Unknown Word Recognition*. In Proceedings of the natural language Processing Pacific Rim Symposium 1997 (NLPRS'1997).

Laungaramsri, P. (2003). *Ethnicity and the politics of ethnic classification in Thailand.* Ethnicity in Asia, 157-173.

Liu, N., & Nation, I. S. P. (1985). *Factors affecting guessing vocabulary in context*. RELC Journal, 16(1), 33-42.

Luo, Yiyuan (2012). *Thai Language Proficiency Testing and The Teaching of Thai as a Foreign Language*. International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, August 17-18, 2012, Bangkok.

MacRostie, Ehry, Rohit Prasad, Stephen Rawls, Matin Kamali, Huaigu Cao, Krishna Subramanian, & Prem Natarajan (2010). *The BBN Document Analysis Service: A Platform for Multilingual Document Translation*. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 447‑54. DAS '10. New York, NY, USA: ACM, 2010.

Marujo, L., Lopes, J., Mamede, N., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J. & Viana,

[113] http://lexicometrica.univ-paris3.fr/numspeciaux/special8/Mult5.pdf

[114] http://rbth.co.uk/society/2015/04/09/russia_has_four_main_centres_for_thai-language_studies_45099.html

C. (2009). *Porting REAP to European Portuguese.* Proceedings of the SLaTE Workshop on Speech and Language Technology in Education.

Meknavin, Surapant, Paisarn Charoenpornsawat, & Boonserm Kijsirikul (1997). *Feature-based Thai Words Segmentation*, NLPRS, Incorporating SNLP-97.

Milton, J. (2010). *The development of vocabulary breadth across the CEFR levels*. Second Language Acquisition and Testing in Europe, I. Bartning, Martin, M., and Vedder, I. Eurosla, 2010, pp. 211-232.

Milton, J., & Alexiou, T. (2009). *Vocabulary size and the Common European Framework of Reference for Languages*. Dans B. Richards et coll. (éds.), Vocabulary studies in first and second language acquisition (pp. 194–211). Basingstoke : Palgrave.

Muller, Charles. *Principes et méthodes de statistique lexicale*. Hachette université. Paris: Hachette, 1977.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2012). *Measuring vocabulary size in an uncommonly taught language*. International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, August 17-18, 2012, Bangkok.

Nilsen, Jørgen. (sans date). *Thai Academic Word List (TAWL): A list of 548 Thai words occurring frequently in academic writing*.[115]

Nilsen, Jørgen. *Thai Frequency Dictionary. 2nd Edition*. 2014.[116]

Noyunsan, Chaluemwut, Choochart Haruechaiyasak, Seksan Poltree & Kanda Runapongsa Saikaew (2014). *A Multi-Aspect Comparison and Evaluation on Thai Word Segmentation Programs*. JIST (Workshops & Posters) 2014: 132-135.

Poltree, S., K. Saikaew (2011). *Thai Word Segmentation Web Service*. Proceedings of the Joint International Symposium on Natural Language Processing and Agricultural Ontology Service 2011 (SNLP-AOS 2011).

Ponmanee, S. (2002). *พื้นฐานการสอนภาษาไทยในฐานะภาษาต่างประเทศ* [*Foundation of teaching Thai as a foreign language*]. Bangkok, Thailand: C.U. Book.

Poowarawan, Y. (1986). *Dictionary-based Thai Syllable Separation*, In Proceeding of the Ninth Electronics Engineering Conference.

Schmitt, Norbert. *Researching Vocabulary : A Vocabulary Research Manual*. Palgrave Macmillan. 2010.

Schmitt, N. and Schmitt, D. (2012). *A reassessment of frequency and vocabulary size in L2 vocabulary teaching.* Language Teaching.

สำนักงานคณะกรรมการการอุดมศึกษา, *นักศึกษาต่างชาติที่ศึกษาในสถาบันอุดมศึกษาสังกัดสำนักงานคณะกรรมการการอุดมศึกษา ปีการศึกษา 2555*, กรุงเทพฯ : 2557. [Office of the Higher Education Commission, *Foreign Students Enrolled in Higher Education Institutions Regulated by the Higher Education Commission, Academic*

[115] http://www.thaifrequency.com/wp-content/uploads/2015/03/THAI-ACADEMIC-WORD-LIST.pdf

[116] http://www.thaifrequency.com/wp-content/uploads/2014/12/Nilsen-J.-2014.-Thai-Frequency-Dictionary.-2.-Edition.pdf

*Year 2012*, Bangkok : 2014.]

Sornlertlamvanich, Virach (1993). *Word Segmentation for Thai in a Machine Translation System* NECTEC.

Sornlertlamvanich, Virach, Tanapong Potipiti, Chai Wutiwiwatchai, & Pradit Mittrapiyanuruk (2000). *The State of the Art in Thai Language Processing*. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 1‑2. ACL '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000.

Theeramunkong, T., Usanavasin, S., Machomsomboon, T., et Opasanont, B. (2000). *Thai Word Segmentation without a Dictionary by Using Decision Trees*. The fourth Symposium on Natural Language Processing 2000.

The Unicode Consortium. *The Unicode Standard*. http://www.unicode.org/versions/latest/

Unicode, Inc. *Thai Code Chart*. 1991-2013 http://www.unicode.org/charts/PDF/U0E00.pdf

West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.

Xue, G., & Nation, I. S. P. (1984). *A university word list.* Language Learning and Communication, 32, 215-219.

# 11. Œuvres de référence

## 11.1. Dictionnaires de thaï

*พจนานุกรมฉบับมติชน* พ.ศ. ๒๕๔๗ (*Matichon Dictionary of the Thai Language, 2004*).

*พจนานุกรมฉบับราชบัณฑิตยสถาน* พ.ศ.๒๕๔๒, พ.ศ.๒๕๕๔ (*Royal Institute Dictionary* 1999, 2011).[117]

*หลักเกณฑ์การใช้เครื่องหมายวรรคตอนและเครื่องหมายอื่น ๆ หลักเกณฑ์การเว้นวรรค หลักเกณฑ์การเขียนคำย่อ* ฉบับราชบัณฑิตยสถาน พิมพ์ครั้งที่ ๖ หน้า ๕๖-๖๖ (*Principes d'emploi des signes de ponctuation et d'autres signes, principes d'espacement et d'abréviation*, Royal Institute, 6e édition, pp. 56-66)[118]

Haas, Mary R., *Thai-English Student's Dictionary*, Stanford University Press, 1964.

*SEAlang Library Thai Dictionary Resource*[119]

*Longdo Thai Dictionary Search and Compilation Service*[120]

Smyth, David, *Thai: An Essential Grammar*, Routledge, 2002.

Iwasaki, Shoichi & Ingkaphirom, Preeya, *A Reference Grammar Of Thai*, Cambridge University Press, 2005.

## 11.2. Dictionnaires de fréquence lexicale

Buckwalter, Tim, et Dilworth Parkinson. *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*. Routledge, 2014.

Cermák, František, et Michal Kren. *A Frequency Dictionary of Czech: Core Vocabulary for Learners*. Routledge, 2011.

Davies, Mark. *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. Routledge, 2006.

Davies, Mark, et Dee Gardner. *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates and Thematic Lists*. Routledge, 2013.

Davies, Mark, et Ana Maria Raposo Preto-Bay. *A Frequency Dictionary of Portuguese*. Routledge, 2007.

Jones, Randall, et Erwin Tschirner. *A Frequency Dictionary of German: Core Vocabulary for Learners*. Routledge, 2015.

Lonsdale, Deryle, et Yvon Le Bras. *A Frequency Dictionary of French: Core Vocabulary for Learners*. Routledge, 2009.

Sharoff, Serge, Elena Umanskaya, et James Wilson. *A Frequency Dictionary of Russian: Core Vocabulary for Learners*. Routledge, 2014.

Tiberius, Carole, et Tanneke Schoonheim. *A Frequency Dictionary of Dutch: Core Vocabulary*

[117] http://rirs3.royin.go.th/dictionary/
[118] http://www.royin.go.th/?page_id=629
[119] http://www.sealang.net/thai/dictionary.htm
[120] http://dict.longdo.com/

*for Learners*. Routledge, 2013.

Tono, Yukio, Makoto Yamazaki, et Kikuo Maekawa. *A Frequency Dictionary of Japanese*. Routledge, 2013.

Xiao, Richard, Paul Rayson, et Tony McEnery. *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners*. Routledge, 2015.

### 11.3. Manuels de thaï langue étrangère

Becker, Benjawan Poomsan. *Thai for Advanced Readers*. Bilingual edition. Paiboon Publishing, 2000.

Brown, J. Marvin. *A.U.A. Language Center Thai Course, Book 1*. 1 edition. Southeast Asia Program Publications, 1974.

Brown, J. Marvin. *A.U.A. Language Center Thai Course: Reading and Writing–Mostly Reading*. 1 edition. Southeast Asia Program Publications, 1986.

Butori, Bernard Noonpackdee et Watana Butori. *Introduction au Thaï*. Assimil France, 2002.

Conjeaud, Michèle, et Wanee Pooput. *Pratique du thaï. Volume 1*. L'Asiathèque - Maison des langues du monde, 2010a.

Conjeaud, Michèle, et Wanee Pooput. *Pratique du thaï. Volume 2*. Langues et Mondes L'Asiathèque, 2010b.

Delouche, Gilles. *Méthode de thaï : Volume 1*. L'Asiathèque - Maison des langues du monde, 2009a.

Delouche, Gilles. *Méthode de thaï : Volume 2*. L'Asiathèque - Maison des langues du monde, 2009b.

Haas, Mary. *Thai Reader*. Ithaca, N.Y.: Spoken Language Services, 1978.

Hoonchamlong, Yuphaphann. *Thai Language and Culture for Beginners Book 1*. University of Hawaii Press, 2007a.

Hoonchamlong, Yuphaphann. *Thai Language and Culture for Beginners Book 2*. University of Hawaii Press, 2007b.

Jones, Robert B., Ruchira C. Mendiones, and Craig J. Reynolds. *Thai Cultural Reader Book 1*. SEAP Publications, 1976.

Kesavatana-Dohrs, Wiworn. *Everyday Thai for Beginners*. Silkworm Books, 2007.

Phrajwan, Ratnitri, *Maanii Readers*, Bangkok: 1978- [121]

Ponmanee, Sriwilai. *45 Thai Stories for Listening and Reading Practice*. Thai Studies Center, Chiangmai University, 2001.

Ponmanee, Sriwilai. *Reading to learn Thai*. Thai Studies Center, Chiangmai University, 2000.

[121] http://www.seasite.niu.edu/Thai/maanii1/maaniireaders.htm

# 12. Annexes

## 12.1. Les niveaux de compétences du test CU-TFL et ses équivalents[122]

<table>
<tr><th>CU-TFL</th><th>ACTFL</th><th>ILR/FSI</th><th>CECR</th></tr>
<tr><td rowspan="3">ดีเด่น<br>Chula Distinguished</td><td rowspan="3"></td><td>5</td><td rowspan="3">C2</td></tr>
<tr><td>4+</td></tr>
<tr><td>4</td></tr>
<tr><td rowspan="2">ดีมาก<br>Chula Superior</td><td rowspan="2">Superior</td><td>3+</td><td rowspan="2">C1</td></tr>
<tr><td>3</td></tr>
<tr><td rowspan="2">ดี<br>Chula Advanced</td><td>Advanced Plus</td><td>2+</td><td rowspan="2">B2</td></tr>
<tr><td>Advanced</td><td>2</td></tr>
<tr><td rowspan="3">กลาง<br>Chula Intermediate</td><td>Intermediate High</td><td>1+</td><td rowspan="3">B1</td></tr>
<tr><td>Intermediate Mid</td><td rowspan="2">1</td></tr>
<tr><td>Intermediate Low</td></tr>
<tr><td>ฝึกพูด Chula Novice</td><td>Novice High</td><td rowspan="2">0+</td><td>A2</td></tr>
<tr><td></td><td>Novice Mid</td><td>A1</td></tr>
<tr><td></td><td>Novice Low</td><td>0</td><td></td></tr>
</table>

[122] Sources : http://www.sti.chula.ac.th/academic/non-native/CU-TFL et http://www.ctsynu.com/showArticle.php?id=1368

## 12.2. Règles d’utilisation de l’espace typographique en thaï

Cette liste est basée sur une publication du *Royal Institute of Thailand* (actuellement appelé *Royal Society of Thailand*) sur les principes d’emploi des signes de ponctuation.

L’espace typographique simple est employée pour

1. indiquer le début d’une proposition coordonnée qui commence par des connecteurs tels que และ /lɛ́ʔ/ (*et*), หรือ /rŭ̌ː/ (*ou*), แต่ /tɛ̀ː/ (*mais*), sauf en cas de phrase courte, où il n’y a pas de coupure.
2. séparer le prénom et le nom de famille.
3. pour les noms de membres de la famille royale, séparer le titre de noblesse et le nom.
4. séparer le nom d’une société et จำกัด /tɕamkàt/ (qui veut dire *limité* comme dans *SARL Société À Responsabilité Limitée*) et, de la même façon, entre le mot ห้างหุ้นส่วนจำกัด /hâːŋhûnsùːantɕamkàt/ (*société en commandite*) ou ห้างหุ้นส่วนสามัญนิติบุคคล /hâːŋhûnsùːansăːmanníʔtìʔbùkkʰon/ (*statut de partenariat enregistré simple*) et le nom de la société ou partenariat.
5. séparer les noms de lieux commençant avec des mots tels que ถนน /tʰànǒn/ (*rue*) ตำบล /tambon/ (*district*) จังหวัด /tɕaŋwàt/ (*province*) du reste de la phrase.
6. séparer un titre de profession ou son abréviation et le nom de famille, comme ศาสตราจารย์ /sàːtsàʔtraːtɕaːn/ (*Professeur*).
7. séparer le grade militaire et le nom.
8. entourer des groupes de lettres suivies d’un point.
9. séparer les chiffres et les lettres.
10. séparer le jour et l’heure.
11. séparer les dimensions (de largeur, profondeur ou hauteur).
12. indiquer un changement d’écriture, par exemple entre la fin d’un mot en thaï et le début d’un mot en lettres latines.
13. séparer les éléments d’une liste. Le mot ได้แก่ /dâjkɛ̀ː/ (*y compris*) qui précède une liste est entouré d’espaces et เป็นต้น /pentôn/, qui apparaît à la fin d’une liste, est précédé d’une espace.
14. entourer des signes de ponctuation comme ๆ et ฯลฯ et les signes mathématiques (+, = ).
15. séparer un sujet de son propos.
16. démarquer les lettres ฉ et ธ, quand elles constituent une unité linguistique seule, ฉ en tant que préposition (prononcé /ná/) et ธ en tant que pronom (prononcé /tʰá/).
17. entourer le mot เช่น /tɕʰên/, quand il est utilisé dans le sens de « par exemple », mais sans espaces quand il est utilisé dans le sens de « comme » (dans une comparaison).
18. séparer le mot ว่า /wâː/ de ce qui le suit (une citation directe ou indirecte).

L’espace n’est pas utilisée

19. entre le titre de civilité et le prénom, comme นาย /naːj/ (*Monsieur*) นาง /naːŋ/ (*Madame*) ou คุณ /kʰun/, y compris les titres de civilité religieux et titres de noblesse donnés par le roi.
20. entre le prénom et le titre de civilité s’il s’agit d’un rang ou d’une profession.
21. avant le préfixe d’un nom d’un institut ou des groupes de personnes, c’est-à-dire entre le mot « Ministère » et le nom du ministère.
22. après le signe ฯ (ไปยาลน้อย /pajjaːnnɔ́ːj/) le symbole d’abréviation.
23. autour du trait d’union ou le tiret.

En règle générale, l’utilisation de l’espace avec les signes de ponctuation occidentaux suit les règles de typographie anglaise.

## 12.3. Liste des abréviations

| | |
|---|---|
| ACTFL | American Council on the Teaching of Foreign Languages |
| ARI | Automated Readability Index |
| ASEAN | Association of Southeast Asian Nations *Association des nations de l'Asie du Sud-Est* |
| AWL | Academic Word List. Liste de vocabulaire spécifique aux textes académiques en anglais de Coxhead (2000) |
| CECR | Cadre européen commun de référence pour les langues. Échelle d'évaluation standard de compétences linguistiques. |
| CL | Classificateur = spécificatif. Indique la classe d'un nom. |
| CU-TFL | Chulalongkorn University Proficiency Test of Thai as a Foreign Language |
| FKGL | Flesch-Kincaid Grade Level. Indice de lisibilité |
| FSI | Foreign Service Institute. Échelle d'évaluation standard de compétences linguistiques. |
| GFI | Gunning-Fog Index. Indice de lisibilité |
| GSL | General Service List |
| HSF | High School Frequency Word List |
| ILR | Interagency Language Roundtable[123] Échelle d'évaluation standard de compétences linguistiques, aussi connue comme l'échelle FSI. |
| L1 | Langue maternelle |
| L2 | Langue étrangère ou Langue seconde |
| NGSL | New General Service List (Browne et coll., 2013)[124] |
| ORCHID | Open Linguistic Resources Chanelled toward InterDisciplinary research. |
| REST | REpresentational State Transfer. Style d'architecture d'applications Web |
| RID | Royal Institute Dictionary |
| SMOG | Simple Measure of Gobbledygook. Indice de lisibilité |
| SWATH | Smart Word Analysis for THai. Outil de segmentation lexicale |
| TAL | Traitement automatique des langues |
| TFL | Thai as a Foreign Language *thaï langue étrangère* |
| TNC | Thai National Corpus |
| UWL | University Word List |
| VOA | Voice of America. Le service de radiodiffusion internationale officiel des États-Unis |

[123] http://www.govtilr.org/

[124] http://www.newgeneralservicelist.org/