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

Jennifer Lewis-Wong, Satenik Mkhitaryan

AFCP

ATALA

Journées d'Études sur la Parole
Traitement Automatique des Langues Naturelles
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

PARIS Inalco du 4 au 8 juillet 2016

Organisé par les laboratoires franciliens

https://jep-taln2016.limsi.fr

Conférenciers invités:

Christian Chiarcos (Goethe-Universität, Frankfurt.)
Mark Liberman (University of Pennsylvania, Philadelphia)

Coordinateurs comités d'organisation

Nicolas Audibert et Sophie Rosset (JEP)
Laurence Danlos & Thierry Hamon (TALN)
Damien Nouvel & Ilaine Wang (RECITAL)
Philippe Boula de Mareuil, Sarra El Ayari & Cyril Grouin (Ateliers)

Table des matières

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

Jennifer Lewis-Wong[1, 2] Satenik Mkhitaryan[1]
(1) Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM - EA 2520), INALCO, 2 rue de Lille, 75343 Paris Cedex 07, France
(2) Langues et Civilisations à Tradition Orale - CNRS / Paris III / INALCO (LACITO - UMR 7107), 7 rue Guy Môquet (bât. D), 94801 Villejuif Cedex, France
jennifer.wong@inalco.fr, satenik.mkhitaryan@inalco.fr

RÉSUMÉ

Cet article a pour objet la création automatique de ressources pour l'apprentissage de langues étrangères peu enseignées et peu dotées en matériels pédagogiques à partir de textes authentiques. Il s'inspire du travail de Ghadirian (2002) et son logiciel *TextLadder*, une application qui classifie les textes d'un corpus selon un ordre qui maximise la facilité de lecture pour l'apprenant, en calculant la similarité lexicale entre les textes. La classification automatique de textes par progression lexicale constitue une méthode intéressante pour proposer une séquence de textes appropriée au niveau d'un lecteur en L2, aussi bien pour proposer des textes à des lecteurs autonomes que pour la création de matériels pédagogiques destinés à être utilisés en classe. Cette méthode est spécialement bien adaptée à la classification de textes qui portent sur une thématique particulière.

ABSTRACT

Text classification by lexical progression for L2 reading practice in Thai and Hindi

This article looks at the creation of teaching and learning resources for less commonly taught languages from unsimplified texts. The inspiration for this study comes from Ghadirian (2002) and the associated computer program TextLadder. *The program classifies a series of texts by their lexical similarity, introducing target vocabulary incrementally and thus making reading easier for the learner. This kind of automated text sequencing can be used to select sequences of texts appropriate to the level of lexical competence of the L2 reader, whether for independent readers or for creating teaching material for classroom use. The method is particularly suitable for classifying texts with a similar topic or theme.*

MOTS-CLÉS : ALAO, aide à la lecture, thaï langue étrangère, hindi langue étrangère, lisibilité, TextLadder.

KEYWORDS: CALL, Reading aides, TFL, HFL, Readability, TextLadder.

1 Introduction

La pratique autonome et précoce de la lecture par un apprenant de langue étrangère (L2) présente maints avantages, notamment une exposition accrue à la langue étudiée et un réel plaisir.

Aujourd'hui, le web propose, pour d'innombrables langues, quantité de textes de tout genre, thème et taille. Il est ainsi susceptible de répondre aux besoins de lecture d'apprenants, même de langues, moins enseignées. Cette abondance de textes peut pourtant s'avérer être un handicap, voire un cauchemar, pour des apprenants de niveau initial ou intermédiaire, incapables de trouver des textes adaptés à leur niveau de compétence, qu'elle soit grammaticale ou lexicale.

Nous présentons dans cet article un dispositif automatique d'aide à la lecture, proposant un parcours de lecture au sein d'un corpus de textes optimisant l'acquisition de vocabulaire nouveau par le lecteur. Ce dispositif, inspiré de Ghadirian (2002), a été appliqué à deux langues fort diverses tant du point de vue linguistique (isolante / flexionnelle) que de leur système d'écriture (alpha-syllabaires sans / avec espaces), le thaï et le hindi, par ailleurs peu dotées en ressources pédagogiques.

Nous ferons tout d'abord, au chapitre 2, un rapide état de l'art concernant d'une part la notion de lisibilité des textes, c'est-à-dire d'évaluation de leur niveau de difficulté pour un apprenant, concernant d'autre part des études analogues portant aussi sur la classification de textes selon des critères lexicaux. Au chapitre 3 nous présenterons notre méthodologie : la constitution de listes lexicales de vocables, le choix des corpus de textes thaï et hindi et leurs prétraitements spécifiques, puis les principes de la classification et de création d'un parcours de lecture. Dans le chapitre 4, nous rendrons compte des tests de classification des deux corpus et des résultats obtenus. Enfin, le chapitre 5 proposera un regard critique sur l'étude et diverses pistes d'amélioration.

2 Travaux antérieurs

2.1 Sur la lisibilité des textes

Si les bienfaits de la lecture personnelle régulière sur l'acquisition d'une langue étrangère, surtout en ce qui concerne l'élargissement du vocabulaire, sont bien établis (Krashen, 2004), l'abondance de textes ne garantit pas que l'apprenant puisse trouver aisément des textes adaptés à son niveau. En effet, l'apprenant peut se démotiver face à un texte qui ne correspond pas à son niveau de connaissances lexicales. Pour les enseignants de langues peu enseignées et peu dotées en matériel pédagogique, choisir des textes qui correspondent au(x) niveau(x) de ses étudiants est également un défi. Selon Liu et Nation (1985) le lecteur en L2 doit connaître 95 % des mots d'un texte avant de pouvoir déduire le sens des mots inconnus. C'est ce qu'on nomme la couverture textuelle.

En règle générale, les chercheurs en lisibilité ont recours à deux types de stratégies pour évaluer la difficulté des textes. La première est l'élaboration de formules de lisibilité qui s'appuient sur des mesures de caractéristiques superficielles des textes, la deuxième stratégie est le développement de modèles statistiques plus complexes, basés sur des corpus de textes dont la difficulté a déjà été mesurée, comme les textes des manuels scolaires.

Ces méthodes sont peu adaptées aux langues peu enseignées en L2. En premier lieu, la plupart des travaux sur la lisibilité mesurent la difficulté des textes pour les locuteurs natifs. Or, François (2011) a montré que la lisibilité d'un texte pour un lecteur en L1 n'est pas la même que pour un lecteur en L2. Heilman et al. (2007, voir plus loin) ont trouvé que la difficulté grammaticale est plus décisive dans la lisibilité de textes en L2 qu'en 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.

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 recherche sur la lisibilité en L2 s'est concentrée surtout sur l'anglais[1]) 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[2]. Le développement de modèles de lisibilité spécifiques à l'apprentissage en L2 des langues peu outillées et avec peu d'apprenants peut s'avérer impossible pour des raisons pratiques de temps et de ressources.

Le logiciel *TextLadder*, de Ghadirian (2002), développé pour des apprenants d'anglais langue étrangère, a l'avantage de ne pas dépendre de ces méthodes de lisibilité. Le dispositif sélectionne une séquence de textes à partir d'un corpus et les agence dans un ordre qui optimise la facilité de lecture, en prenant en compte une liste de vocabulaire connu et une liste de vocabulaire « cible » que le lecteur souhaiterait acquérir. Au lieu d'attribuer aux textes une note de difficulté, la classification est relative et plus fine ; le dispositif peut être paramétré pour prendre en compte les connaissances lexicales réelles de l'apprenant.

2.2 Applications de classification textuelle selon le lexique

REAP[3] est un système d'aide pédagogique pour des professeurs d'anglais L1 ou L2. Il a la particularité de prendre en compte les connaissances spécifiques de l'étudiant et son niveau d'études dans le choix des textes, des informations qui permettent la personnalisation du parcours de lecture. *REAP* a été porté en portugais et en français (Marujo et al., 2009).

La classification des textes par niveau de difficulté utilise une stratégie hybride, combinant modèles de langue basés sur la statistique lexicale et une classification par difficulté grammaticale modélisée sur des constructions grammaticales trouvées dans des manuels d'anglais langue étrangère de niveaux différents. La recherche sur laquelle *REAP* est basée (Heilman et al., 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.

1 Pour un état de l'art de la recherche sur la lisibilité en L2, lire François (2011).

2 Das et Roychudhury (2004; 2006), cités par Islam (2012) notent que l'indice de lisibilité de l'anglais Flesch–Kincaid donne des résultats satisfaisants pour le bengali, par exemple.

3 Le projet *REAP* est sous-titré *Reader-Specific Lexical Practice for Improved Reading Comprehension*. Voir Brown & Eskenazi (2004).

TextLadder est un logiciel conçu par Ghadirian (2002) qui classe des textes en anglais selon un ordre de lecture donné 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* propose un corpus des textes d'anglais simplifié tirés du site de la Voice of America (VOA) classifiés de la même manière que *TextLadder*. Le site du projet taïwanais d'apprentissage Internet *Candle* inclut un module de pratique de la lecture appelé ***TextGrader***, conçu par Huang et Liou (2007) et basé sur le travail de Ghadirian (2002).

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 et Nation (1984). Ces listes représentent respectivement le vocabulaire le plus fréquent de l'anglais et le vocabulaire supplémentaire le plus fréquent des textes académiques. 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). 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*.

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 déjà mentionné que le lecteur en L2 doit connaître 95 % des mots d'un texte avant de pouvoir déduire le sens des mots inconnus (Liu et Nation 1985). Ensuite sont créées deux listes : une liste du 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 servent à constituer la liste de vocabulaire cible dans chaque cas.

Après élimination de textes qui ne sont pas couverts à 95 % par les listes lexicales initiales, le dispositif 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 du 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, *TextLadder* permettant au lecteur de rechercher la définition des tous les mots du texte, alors que *TextGrader* ne glose que le vocabulaire cible.

3 Méthodologie

Comme nous avons mentionné plus haut, cette étude s'inspire des travaux de Ghadirian (2002), l'objet est de créer un parcours de lecture dans un corpus de textes, proposant à chaque étape le texte optimisant au mieux l'acquisition du vocabulaire du corpus. Nous présentons, dans ce qui suit, la création des différentes listes, les corpus, les prétraitements spécifiques à chaque langue, ainsi que les étapes de l'implémentation de la méthode.

3.1 Listes lexicales

Nous avons pour objectif l'acquisition du vocabulaire le plus utile, c'est-à-dire le lexique le plus fréquent et le vocabulaire spécifique au corpus. Nous avons trouvé que le seuil de lexique de haute fréquence se trouve à environ 5000 mots. Ce sont ces mots que nous avons utilisés pour former nos listes de vocabulaire connu et cible, complétées par d'autres listes de vocabulaire de langue parlée que nous avons jugé connu d'étudiants de niveau intermédiaire. Nous avons aussi utilisé notre propre intuition en tant qu'apprenantes de ces langues pour situer la frontière entre le vocabulaire connu et cible en nous référant aux manuels d'apprentissage de thaï et de hindi. Au-delà de ces 5000 mille mots se trouve la majorité du lexique d'une langue qui est de basse fréquence. Notre système sélectionne du vocabulaire de basse fréquence qui est fréquent dans le corpus, afin de créer une liste de vocabulaire spécifique au corpus utile pour l'apprenant. Cette liste de vocabulaire spécifique est rajoutée à la liste de vocabulaire cible.

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 al., 2010)[4] sur le site de *Sketch Engine*[5] pour notre projet. Il se trouve que l'outil de segmentation du thaï et le lemmatiseur du hindi utilisés par *Sketch Engine* (*SWATH*, Hindi POS Tagger, voir ci-dessous) sont les mêmes que nous utilisons pour segmenter le corpus thaï et lemmatiser le corpus hindi.

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 environ 108 M tokens et le corpus hindi *HiWaC* en contient environ 65 M. A partir de ces corpus, nous avons obtenu des listes de fréquence triées par fréquence lexicale, chaque élément accompagné du nombre de ses occurrences dans le corpus de référence.

Lors de nos essais sur le corpus de presse thaï, nous avons trouvé que nos listes de vocabulaire de haute fréquence ne couvrent en moyenne que 72 % du vocabulaire des textes de nos corpus, très loin des 95 % recommandé par Liu et Nation (1985) pour la déduction du sens des mots inconnus. Afin de couvrir suffisamment le vocabulaire de basse fréquence pour atteindre les 95 %, notre dispositif complète la liste de vocabulaire cible avec une liste de vocabulaire spécifique au corpus de textes à classifier. Créée à la volée, cette liste, en combinaison avec la liste de haute fréquence, couvre plus aisément 95% des textes. Ceci diffère de la stratégie employée par Ghadirian (2002) et Huang et

[4] La méthode *Corpus Factory* détaillée dans Kilgarriff et al. (2010) consiste à télécharger une sauvegarde 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 ».

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

Liou (2007) dont les systèmes de classification ont recours à des listes préfabriquées spécifiques non seulement au type de texte, mais spécifiques au corpus. Notre stratégie permet d'éviter les inconvénients de listes spécifiques figées : l'élaboration de telles listes de vocabulaire spécifique exige un corpus de textes représentatif (qui n'est pas toujours disponible pour des langues peu dotées) et limite le choix de textes que le système peut classifier.

3.2 Corpus

Pour le thaï, le corpus principal que nous avons utilisé pour nos tests est composé d'articles apparus entre 2013 et 2014 dans un journal quotidien populaire à grand tirage, *ThaiRath*, divisé en huit rubriques. Le corpus hindi contient des articles de presse de l'année 2013 apparus sur le site de NDTV (New Delhi Television Limited) dans la rubrique « India ».

3.3 Prétraitements de textes

Les prétraitements à effectuer sur les textes sont nécessairement dépendants des particularités d'une langue et de son système d'écriture. Comme mentionné précédemment, il est préférable que les prétraitements effectués sur les textes soient les mêmes que ceux utilisés dans la création des listes.

3.3.1 En thaï

Comme tous les systèmes d'écriture de l'Asie du Sud-est qui utilisent des alphasyllabaires, ainsi que le chinois et le japonais, le thaï s'écrit en continu sans séparation entre les mots (*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 (c'est-à-dire les noms de personnes, les toponymes et les noms d'organisations). 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. Nous utilisons l'outil de segmentation nommé *SWATH* (*Smart Word Analysis for THai*) de Meknavin et al. (1997). Cet outil sépare les mots d'éventuels signes de ponctuation thaï, comme le symbole de répétition. *SWATH* améliore la segmentation des mots inconnus avec une analyse linguistique qui prend en compte l'environnement des mots, cherchant des mots de contexte et des collocations pour déterminer la segmentation la plus probable.

วัน|หนึ่ง|_|เขา|ตื่น|ลืมตา|ขึ้น|มา|จาก|ความ|ฝัน|อัน|สับสน|_|รู้สึก|อ่อน|_|เพลีย|จาก|การ|พักผ่อน|ไม่|เพียงพอ|_|เขา|เดิน|โซเซ|ไป|ชนก|อง|หนังสือ|ใน|ห้องรับแขก|ที่ตั้ง|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 1: Exemple de segmentation résultant d'un traitement avec SWATH (Le trait vertical, |, sépare les tokens et le tiret bas, _, représente l'espace dans le texte avant la segmentation)

Les erreurs de segmentation proviennent souvent de la présence d'entités nommées, mais il faut comprendre aussi que la segmentation est ambiguë dans le cas de mots composés, dans ces cas seul le contexte permet de décider définitivement. Compte tenu du fait qu'il s'agit d'une langue sans flexions, une étape de lemmatisation n'est pas nécessaire dans le prétraitement de textes en thaï.

3.3.2 En hindi

Le hindi utilise le système d'écriture devanāgarī (देवनागरी), un système également alphasyllabique, qui s'écrit de gauche à droite avec des espaces entre les mots. La devanāgarī dispose de ses propres symboles pour représenter les chiffres (० १ २ ३ ४ ५ ६ ७ ८ ९), mais les chiffres arabes sont de plus en plus souvent employés. Le hindi, comme la thaï, ne fait pas de distinction entre majuscules et minuscules ce qui rend opaques les entités nommées. La devanagari utilise les signes de ponctuation de l'alphabet latin, sauf la fin de phrase qui est marquée par une barre verticale « । » propre à cette écriture).

Le hindi est une langue à flexion nominale, adjectivale et verbale, donc une phase de lemmatisation est nécessaire pour le corpus hindi. Nous avons utilisé le lemmatiseur intégré à l'étiqueteur morphosyntaxique de Reddy et Sharoff (2011). La lemmatisation résultante n'est pas dépourvue d'erreurs, mais les erreurs sont cohérentes avec nos listes de vocabulaire qui ont été lemmatisées avec le même outil.

3.4 Principes de classification

À l'instar du logiciel *TextLadder* de Ghadirian (2002), notre dispositif dispose d'une liste de vocabulaire connu et d'une liste de vocabulaire cible, c'est-à-dire le vocabulaire à acquérir. Après des prétraitements spécifiques à chaque langue, le classement automatique des textes suit les étapes suivantes :

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

2. Création d'une liste de vocabulaire spécifique au corpus. Cette liste comprend tous les éléments qui ne figurent pas dans les listes initiales, qui ont une fréquence suffisante et représentative (nous avons choisi respectivement huit occurrences et cinq textes).

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

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 et les logogrammes (tels que les chiffres arabes, les symboles monétaires, etc.).

5. Choix du texte le plus accessible d'un point de vue lexical. Deux critères de choix sont possibles : soit la maximisation du vocabulaire connu, soit la maximisation de la couverture textuelle.

6. Ajout du texte à la séquence de lecture.

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.

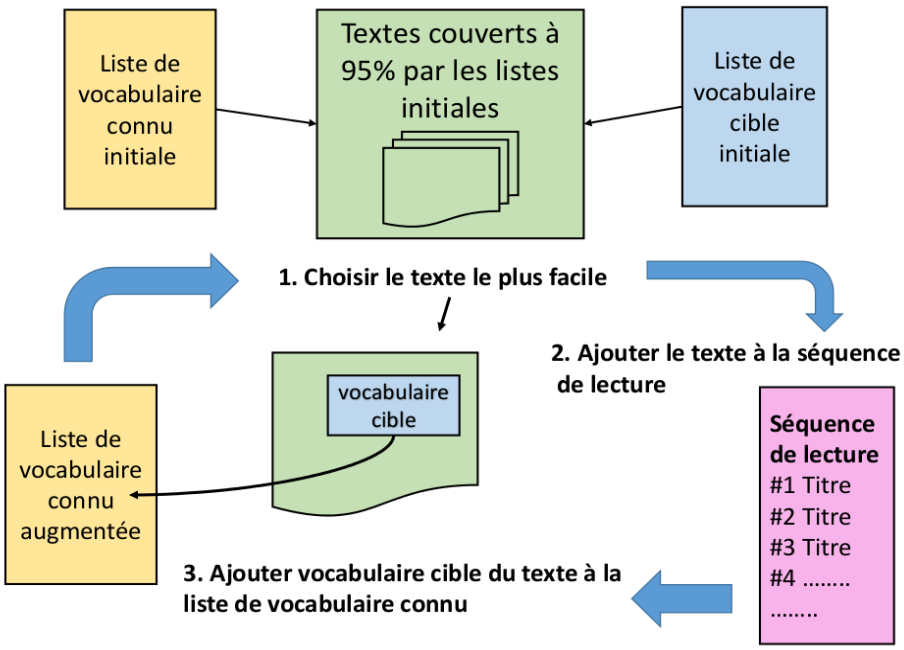

FIGURE 2: Fonctionnement de base

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

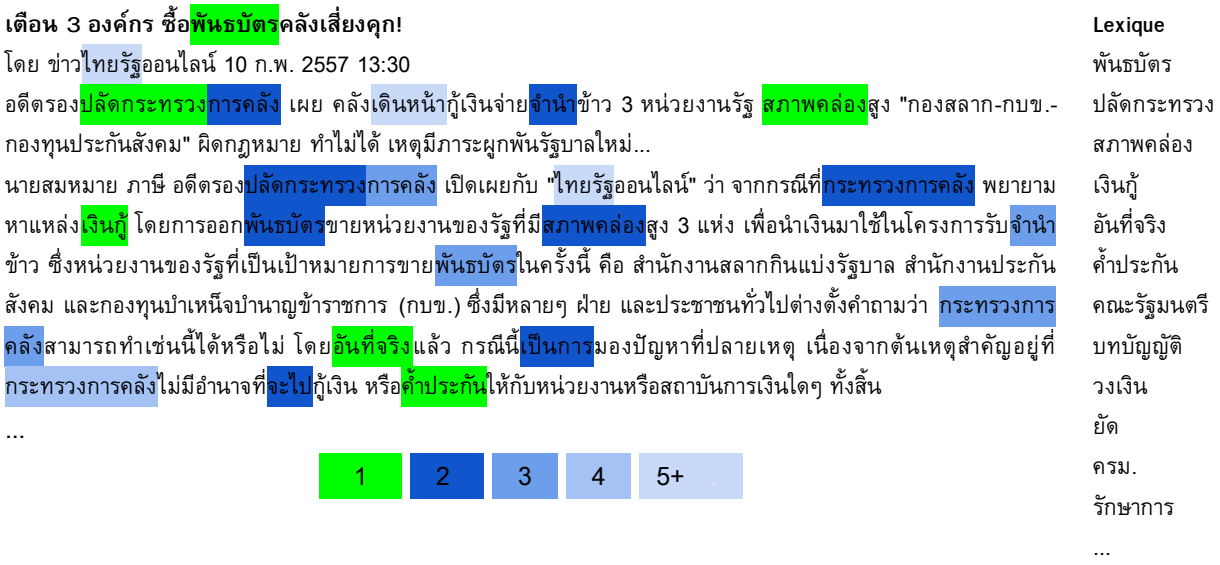

FIGURE 3: Balisage du vocabulaire cible d'un texte du corpus *ThaiRath2013* rubrique *Business* issu d'une séquence de lecture

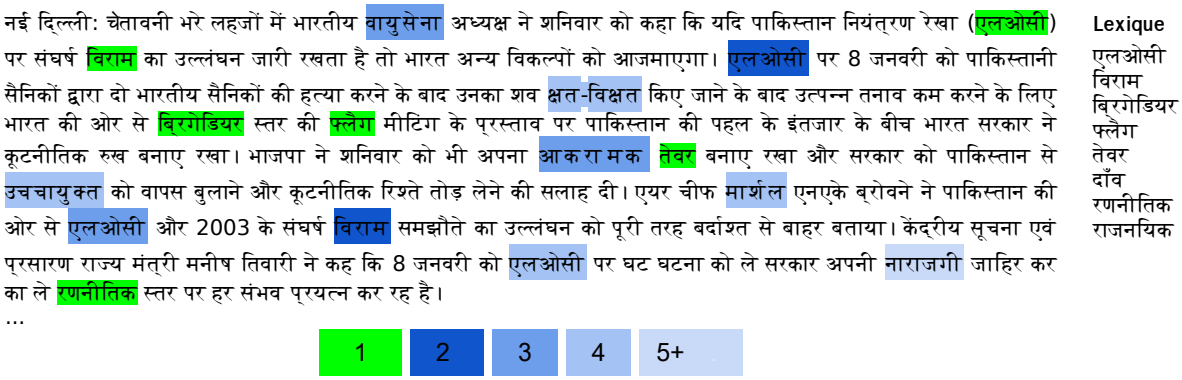

FIGURE 4: Balisage du vocabulaire cible d'un texte du corpus *NDTV Hindi* rubrique *India* issu d'une séquence de lecture

4 Tests sur corpus et résultats

4.1 La segmentation du thaï

Tout comme Jean (2009), nous avons trouvé beaucoup d'instances de sursegmentation d'entités nommées dans le corpus thaï. L'outil de segmentation, qui n'a pas ces mots dans son dictionnaire, les divise en syllabes. Notre dispositif 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 segmentation ne concerne pas uniquement les noms propres, mais aussi les mots anglais écrits en thaï et les mots composés thaï-anglais. 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.

Ayant constaté que beaucoup d'entités nommées sont entourées d'espaces typographiques, nous avons introduit un dispositif d'amélioration de leur segmentation avant l'étape de la création de la liste de vocabulaire spécifique au corpus, à l'aide d'une liste d'entités nommées extraite d'une sauvegarde des entêtes des pages Wikipédia en thaï. 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.

4.2 Influence de types de tri

Nous avons pris un échantillon de 300 textes thaï pris au hasard de notre corpus d'articles de presse *ThaiRath*, rubrique *Business* pour un premier aperçu de la classification de textes en séquence de lecture. Le tri par quantité de vocabulaire connu, qui favorise le texte avec le plus grand nombre de mots connus est comparé avec le tri par couverture textuelle, qui lui favorise le texte avec le plus grand pourcentage de couverture textuelle de la liste de vocabulaire connu. Pour cette dernière, si plusieurs textes ont le même pourcentage de couverture textuelle, le dispositif 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 de répartir le nouveau vocabulaire plus équitablement.

Les résultats de ces deux types de tri sont illustrés par les figures 5 et 6. La figure 6 montre l'évolution de la taille du nouveau vocabulaire sur la séquence de lecture, la figure 5 représente la couverture textuelle par la liste de vocabulaire connu ; pour chaque graphique l'axe horizontal représente la séquence de lecture, par numéro de texte sur la séquence.

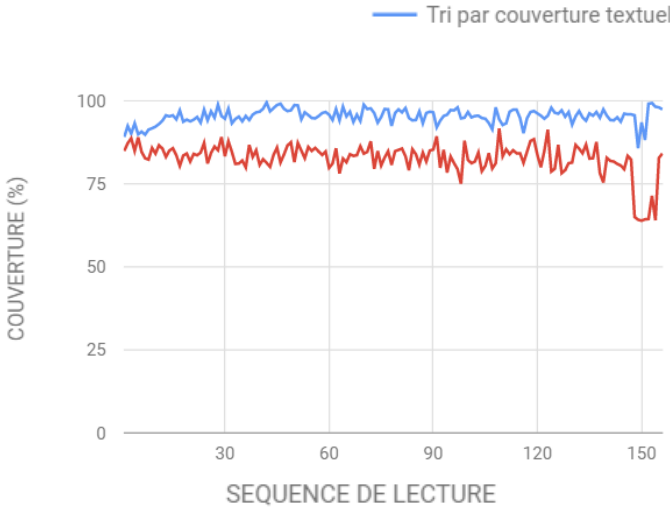

FIGURE 5: Couverture textuelle par liste de vocabulaire connu

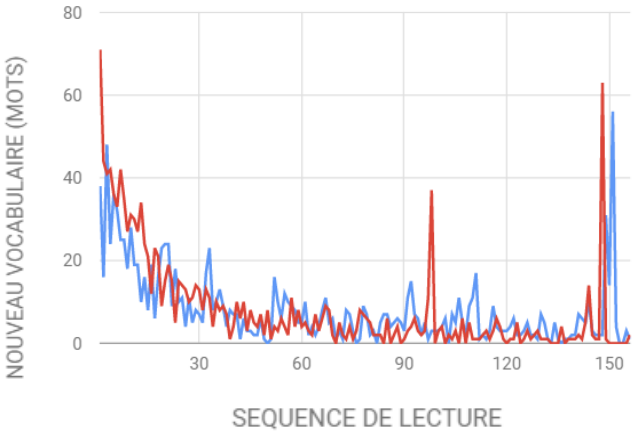

FIGURE 6: Nouveau vocabulaire (corpus ThaiRath rubrique Business de 300 textes)

On constate que l'évolution de la taille du nouveau vocabulaire est moins abrupte pour le tri par couverture textuelle que le tri par quantité de vocabulaire. Dans le cas du tri par couverture textuelle, le premier texte a un vocabulaire de 38 éléments alors que le tri par taille du vocabulaire mettait un texte d'un vocabulaire de 71 éléments en première position. Le nombre de textes sans nouveau vocabulaire a diminué légèrement (8 textes, contre 11 avec le tri par taille) et le nombre de textes avec un seul vocable nouveau a aussi diminué (le nombre de textes est passé de 19 à 15). Le tri par couverture textuelle permet également d'assurer une meilleure couverture textuelle, 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.

4.3 Homogénéité du corpus

Pour tester un classement de textes plus hétérogènes, nous avons testé notre système sur un corpus de 300 articles en thaï de *ThaiRath* 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 du 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. Pour examiner la répétition de vocabulaire, nous avons réduit l'exigence de couverture des listes initiales à 85 % pour disposer de plus d'articles. Nous avons constaté une plus grande proportion d'hapax dans le corpus hétérogène, mais il faut noter qu'il s'agit en grande partie d'entités nommées.

4.4 Application au corpus hindi

Nous avons choisi un sous-corpus de 1030 articles du corpus *NDTV Hindi*. Afin d'observer l'impact de la lemmatisation des textes, nous avons testé d'abord avec un corpus et des listes de vocabulaires non lemmatisés. Le système a sélectionné 160 textes couverts à plus de 95 %. Après la lemmatisation du corpus et des listes de vocabulaires, le nombre de textes sélectionnés a augmenté jusqu'à 183. Ceci montre que la lemmatisation n'a pas eu beaucoup d'effets sur les résultats, car le lemmatiseur choisi ne lemmatise qu'un faible pourcentage de mots et, de surcroît, fait de nombreuses erreurs. Néanmoins nous avons continué les tests avec le corpus lemmatisé, qui est plus intéressant qualitativement en termes acquisition de vocabulaire.

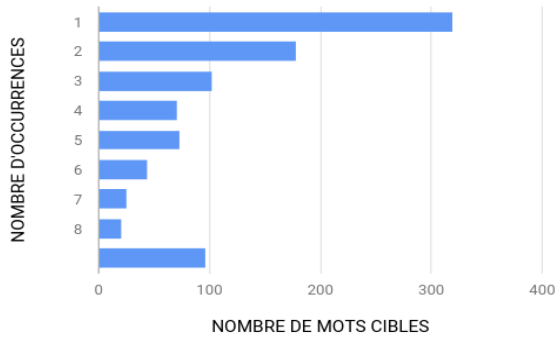

FIGURE *7:* Distribution du vocabulaire cible

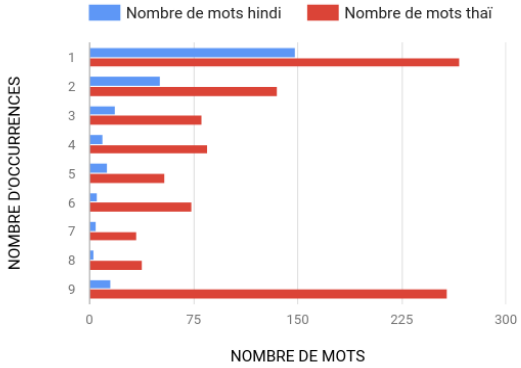

FIGURE *8:* Répétitions du vocabulaire cible

La figure 7 montre un grand nombre d'hapax. Ceci est dû en grande partie à la présence des entités nommées, notamment des toponymes, et de mots anglais qui pourraient être considérés comme du vocabulaire connu. L'utilisation des mots anglais en hindi peut être aléatoire donc il est difficile de les recenser. Nous avons rencontré ce phénomène en thaï aussi, dans une moindre mesure. Un détecteur de mots anglais translittérés dans les langues traitées améliorerait la couverture des listes de vocabulaire connu. Le corpus thaï *Business* serait plus homogène que le corpus hindi *India* donc il contiendrait plus de répétitions (figure 8).

5 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. D'abord, nous avons observé une forte corrélation entre les performances de la méthode (à la fois sur le plan du nombre de répétitions du vocabulaire et du choix des textes) et la qualité des prétraitements effectués sur nos

corpus, en l'occurrence la segmentation du thaï et la lemmatisation du hindi. Nous prévoyons de tester d'autre lemmatiseurs du hindi pour tenter d'améliorer les résultats. D'autres facteurs ont aussi influé les performances de notre dispositif, tels que la présence d'entités nommées et des mots anglais, qui selon nous pourrait être considérés comme du vocabulaire acquis. Les résultats pourraient donc être améliorés par l'intégration de traitements tels que la détection automatique de mots anglais translittérés et des entités nommées, afin de proposer à l'utilisateur l'option de les considérer comme du vocabulaire connu.

Nous sommes conscientes que 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, comme le thaï, dépourvus de repères que donne la syntaxe. Ghadirian (2002) signale lui-même que sa méthode n'analyse pas les collocations, ou les verbes à particule (*phrasal verbs*). Nous ne tenons pas compte non plus du fait que l'apprenant peut acquérir les différents sens d'un lexème à des niveaux différents. En effet, 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. Sur ce plan, les entités nommées et mots anglais écrits en thaï et en hindi nous ont donné beaucoup de matière à réflexion. L'efficacité du dispositif dépend de la capacité des listes de vocabulaire connu et cible de représenter les connaissances lexicales réelles de l'apprenant. Un dispositif de test de niveau avant, voire au cours de la lecture du parcours proposé est à l'étude.

Avant d'entreprendre des modifications possibles, il faudrait définir l'objectif de la lecture. S'il s'agit d'une lecture intensive destinée à l'apprentissage lexical 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 et 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 ? À ces questions s'ajoute le fait que le dispositif ne prend en compte que l'aspect lexical dans le calcul de difficulté des textes et vise principalement l'acquisition lexicale.

6 Conclusion

Les résultats que nous avons obtenus nous permettent de démontrer qu'il est possible de créer des ressources pédagogiques pour la lecture en L2 de façon automatique avec peu d'intervention humaine, sans avoir recours à des modèles ou des formules de lisibilité. Ceci est particulièrement utile dans le traitement d'une langue peu enseignée, pour laquelle les corpus d'entraînement déjà classifiés par niveau de difficulté n'existent pas. En dépit des difficultés méthodologiques rencontrées, le dispositif est un moyen efficace de proposer des textes au niveau de connaissance lexicale de l'apprenant et le parcours de lecture créé permet une répétition de vocabulaire suffisante pour faciliter l'acquisition du vocabulaire nouveau avec un effort moindre pour le lecteur. La méthode marche mieux pour un corpus de thématique homogène, idéal pour la création de matériels pédagogiques pour l'enseignement de langues sur objectifs spécifiques.

Références

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 S. (2002). Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning & Technology* 6(1), 147-164.

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. Actes de *The Human Language Technology Conference. Rochester, NY.*

HUANG H-T., LIOU H-C. (2007). Vocabulary Learning in an Automated Graded Reading Program. *Learning & Technology* 11(3), 64-82.

ISLAM Z., MEHLER A., RAHMAN R. (2012). Text Readability Classification of Textbooks of a Low-Resource Language. Actes du *26th Pacific Asia Conference on Language, Information and Computation, 2012.*

JEAN C. (2009). Le thaï. De la segmentation aux maux. *Lexicometrica. Numéro spécial - Explorations textométriques.* 2009.

KILGARRIFF A., REDDY S., POMIKÁLEK J., AVINESH PVS. (2010). A Corpus Factory for many languages. Actes de *LREC, Malta, 2010.*

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

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

MEKNAVIN S., CHAROENPORNSAWAT P., KIJSIRIKUL B. (1997). Feature-based Thai Words Segmentation. *NLPRS, Incorporating SNLP-97.*

MARUJO L., LOPES J., MAMEDE N., TRANCOSO I., PINO J., ESKENAZI M., BAPTISTA J., VIANA C. (2009). Porting REAP to European Portuguese. Actes du *SLaTE Workshop on Speech and Language Technology in Education.*

REDDY S., SHAROFF S. (2011). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. *Cross Lingual Information Access* 11.

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

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