Le thaï. De la segmentation aux maux

[français-thaï]

Christian Jean
chr_jean2000@yahoo.fr

Résumé : Le thaï ou siamois[1] est une des langues d'Asie-du-Sud-Est à écriture non segmentée dérivée de la dévanagari indienne.

Pour le chercheur qui tente de pénétrer le domaine des études thaïes, la mise à disposition, sur des sites webs thaïlandais, de traductions de textes français réalisées par des traducteurs dont le thaï est la langue maternelle, constitue une occasion précieuse d'avancer dans la compréhension de la langue et de la culture thaïes.

La présente étude est consacrée à l'exploration en corpus à l'aide des outils fournis par ***Lexico3*** des problèmes de segmentation du thaï dans l'optique d'une étude textométrique comparative ultérieure. En effet, des études portant sur le thaï dans le domaine du traitement automatique des langues sont de plus en plus présentées en France. Toutes introduisent une spécificité du thaï à savoir l'utilisation d'une écriture non segmentée mais rares sont celles montrant les intrications entre les notions de syllabe, de morphème et d'unité lexicale dans le système de la langue thaïe.

Pour réaliser cette étude nous disposons d'un segmenteur automatique permettant de segmenter les textes thaïs en trois niveaux : la syllabe, le morphème lexical et l'unité lexicale. Les méthodes de segmentation de cet outil ont fait l'objet d'une publication en thaï [Asa2003]. Nous nous appuierons sur cette étude pour définir les notions de syllabes, de morphèmes lexicaux et d'unités lexicales. Acquiesçons que sans cet outil et sans cette publication, la présente étude aurait été impossible à réaliser.

Nous disposons par ailleurs d'un corpus parallèle de nouvelles françaises traduites en thaï. Ce corpus initialement préparé dans le but de faire une étude textométrique comparative entre le français et le thaï, permettra d'apprécier le sens des mots thaïs en fournissant le référentiel sémantique d'origine en plus de fournir des mots inconnus au segmenteur.

La section §1 présente les particularités du système d'écriture thaï ainsi que les trois niveaux de segmentation utilisés. La section §2 présente le corpus sélectionné. La navigation dans les syllabes, les morphèmes et les unités lexicales débute véritablement dans la section §3. La dernière section §4 est consacrée à un approfondissement des problèmes de segmentation en unités lexicales.

1 Présentation du thaï

Nous commencerons par décrire quelques propriétés du thaï sur lesquelles les chercheurs s'accordent en général et qui nous seront utiles pour notre étude.

---

[1] Le terme thaï (ไทย) est la manière dont les Thaïs nomment leur langue, leur pays et eux-mêmes depuis 1939. Le siamois est le dialecte du centre de la Thaïlande (ancien royaume du siam) promu au rang de langue officielle, on l'appelle aussi thaï standard.

La langue et son système d'écriture

Le thaï est une langue isolante c'est-à-dire que tous les mots sont invariables : le masculin, féminin, singulier et pluriel ne sont pas morphologiquement marqués. Les verbes ne se conjuguent pas. C'est une langue à tendance monosyllabique dont les nombreux emprunts au sanskrit, au pâli et plus récemment à l'anglais ont introduit de nombreux mots constitués de plusieurs syllabes.

Comme on le voit sur l'extrait de traduction présenté ci-dessous, le thaï possède une écriture non segmentée. Les mots ne sont pas séparés les uns des autres par des espaces. Il n'y a pas de délimiteur de phrase comparable aux signes de ponctuation de l'alphabet latin bien que l'espace[Tha1978] puisse sembler jouer parfois ce rôle.

L'écriture thaïe[2] utilise 44 signes consonnes et 19 signes supplémentaires qui en se combinant permettent de représenter 32 voyelles. À cela il faut ajouter 4 marques tonales, 2 diacritiques, 10 chiffres traditionnels, 3 marques additionnelles pour les mots pâli/sanskrit et 6 signes typographiques utilisés principalement dans les œuvres versifiées. Dans le corpus que nous avons réuni, on remarque aussi la présence de guillemets.

Segmentations préalables des textes thaïs

Afin de rendre le texte thaï analysable par ***Lexico3*** nous l'avons préalablement segmenté en utilisant l'outil Kucut[3] développé par l'unité de recherche NaiST[4] de l'université Kasetsart spécialisée dans le traitement automatique des textes écrits en thaï.

La méthode de segmentation utilisée par ce segmenteur est décrite dans [Asa2003]. Le taux de reconnaissance des mots déclaré est d'environ 80% pour la segmentation des mots inconnus et de 65% pour la fixation des frontières de l'unité lexicale. Cet outil permet de réaliser la segmentation sur trois niveaux différents.

Le premier niveau est la syllabe. Cette segmentation consiste à regrouper des caractères afin de former une syllabe prononçable. Par exemple :

- Le mot ทรราช /thorara:t/[5] sera découpé en 2 syllabes ทร/thon/-[6]ราช/ra:t/ (mot d'origine sanskrite)

---

[2] Tous les caractères thaïs sont répertoriés dans le seul standard existant : le TIS 620-25335 défini en 1990 par l'Institut des Standards Industriels Thaïlandais. Il est encodé principalement par deux tables d'encodage 8 bits très similaires : la tis620, table officielle et la Windows-874 très utilisée dans le monde Microsoft. Ce jeu de caractères est aussi représenté dans Unicode.

[3] Kucut est un programme écrit en Python et téléchargeable gratuitement : http://naist.cpe.ku.ac.th/wordcut/static/kucut-1.2.2.tar.gz

[4] Natural Language Processing and Intelligent Information SystemTechnology Research http://naist.cpe.ku.ac.th/

[5] Notes sur la translittération : la translittération utilisée ici est une solution ad hoc ayant pour but l'identification des mots par le lecteur. Elle renseigne peu sur la façon de lire car ni les tons, ni les valeurs et ni les longueurs de voyelles ne sont vraiment représentés.

[6] On utilisera tout au long de l'article le – pour marquer la segmentation des syllabes.

- Le mot เขลา /khlaw/ sera découpé en une seule syllabe bien qu'on aurait pu le découper en deux syllabes เข/khe:/-ลา/la:/ mais dans ce cas, on aurait eu soit deux mots thaïs, soit un mot d'origine étrangère. Le ข/kh/ et le ล/l/ forment un groupe consonantique.

Le second niveau de segmentation est celui du morphème lexicale[7]. Il est défini comme la plus petite unité ayant un sens et apparaissant dans le dictionnaire de mots du segmenteur. Par exemple :

- พ่อ /phau:/, père; แม่ /mè:/, mère; หุง /hung/, cuire; สะพาน /sapha:n/, pont.

Le troisième niveau est celui de l'unité lexicale. L'unité lexicale est soit un morphème lexical, soit un mot composé[8]. Un mot composé est la fusion de plusieurs morphèmes dont le sens est changeant par rapport à ces morphèmes. Par exemple :

- Simple : พ่อ : père; น้ำ /nam/ : eau;
- Composée **พ่อ-แม่** : parents; แม่-น้ำ /mè- nam/ : rivière, fleuve;.

Le but de notre étude est de pouvoir observer en corpus les formes les plus et les moins spécifiques de chacun de ces niveaux, d'initier le lecteur à la complexité de différencier un mot composé d'un syntagme nominale et de déterminer à quoi correspond réellement ce niveau d'unité lexicale.

2 Le corpus

Nous présentons dans cette partie le corpus de travail, sa structure logique ainsi que les problèmes d'encodage.

Constitution

Ce corpus est constitué d'un ensemble de nouvelles françaises[9] ainsi que de leurs traductions en thaï. Elles sont disponibles sur le site http://www.wanakam.com. Un travail de normalisation ainsi qu'un alignement manuelle en unités de traduction a été effectué. Celle-ci varie d'une phrase à plusieurs paragraphes selon les nouvelles.

Nous disposons de deux fichiers de travail qui ont chacun une finalité et donc une structure différente.

Le premier fichier de travail *sylmorwor-corpus-th-cp874* a pour but l'étude des différents niveaux de segmentation du thaï. Il contient les textes thaïs en trois exemplaires divisés en parties selon leur niveau de segmentation. Elles sont identifiées par la clé <langue> dont les valeurs sont thsyl pour la partie segmentée en syllabes, thmor pour celle segmentée en morphèmes et thlex pour celle segmentée en unités lexicales. Chacune de ces parties est

---

[7] Nous employons le terme morphème lexical bien qu'il puisse s'agir de mots outils pour indiquer qu'il n'est pas question de morphèmes comme dans les langues flexionnelles ou agglutinantes.

[8] Pour le lecteur curieux, ouvrir un dictionnaire thai-anglais à l'entrée /kham/ que l'on donne comme traduction du mot « mot » peu impressionner tant la liste de mots composés à partir de ce morphème lexical est longue. Par exemple dans le SE-ED's thai-english dictionnary la liste débute à la fin de la page 133 et s'achève à la fin de la page 136.

[9] Auteurs de ces nouvelles : Alphonse Allais, Apollinaire, Aragon, Baudelaire, Bertot, Bloy, Daudet, Didier Daeninckx, Flaubert, Jean Hourgon, JMG Leclezio, Marcel Schwob, Maupassant, Perrault, Raymond Radiguet, Renard, Roegiers, Sagan, Sartre, Zola.

divisée en nouvelles identifiées par la clé <nouvelle> dont la valeur est composée d'un numéro et préfixée de la lettre A, B ou C pour les nouvelles segmentées respectivement en syllabes, morphèmes et unités lexicales. Par exemple la nouvelle 001 est identifiée par les valeurs <nouvelle=A001>, <nouvelle=B001> et <nouvelle=C001>.

Illustration 1: Extrait de la traduction de la nouvelle Arfled d'Alphonse Allais

```
<langue="frth">
<nouvelle="001"><auteur="1">
<par="00001">
|
```

Le Dr Joris-Abraham-W. Snowdrop, de Pigtown (U.S.A.), était arrivé à l'âge de cinquante-cinq ans, sans que personne de ses parents ou amis eût pu l'amener à prendre femme.

หมอ จอรีส __ อับราฮัม __ ดับเบิลยู __ สโนว์ดรอป __ เมือง พิกทาวน์ __ (สหรัฐอเมริกา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โกโหติกา หรือ เพื่อน สนิท ผู้ใด สามารถ โน้มน้าว ให้ เขา แต่ง ภรรยา ได้

```
<par="00002">
```

Texte 1: Extrait du fichier en relation de traduction.

Le deuxième fichier de travail dont nous disposons, corpus-frth-al-win contient la version française en relation de traduction avec la version thaïe segmentée en unités lexicales. La partition langue unique est identifiée par la clé <langue=frth>. La valeur de la clé nouvelle est seulement composée du numéro de la nouvelle sans être préfixée d'une lettre.

La clé <par> ainsi que le symbole délimiteur de section sont utilisés de manière à maintenir la relation de traduction. L'intérêt de cette structure est de pouvoir retrouver facilement les unités lexicales en relation de traduction à l'aide de la carte des sections.

Encodage des textes thaïs pour Lexico3

Le couteau suisse de ***Lexico3*** permet d'afficher les caractères thaïs lorsqu'ils sont encodés avec win874. Cependant, on doit prendre quelques précautions car les caractères § et ฯ partagent le même code 8 bits. Il faut donc exclure § de la liste des séparateurs et y ajouter le caractère | qui sert de délimiteur de sections dans notre corpus. Comme on veut garder la trace des espaces originaux, on exclu aussi le caractère _ de la liste des délimiteurs.

La table win874, idéale pour des textes bilingues anglais-thaï, permet de travailler simultanément avec les caractères ASCII et les caractères thaï mais pas avec les caractères français accentués. Ainsi il faudra faire un choix d'affichage lorsqu'on travaillera avec les fichiers contenant à la fois les versions françaises et les versions thaïes des nouvelles.

<langue="fr"><nouvelle="001"><auteur="1"> Le Dr Joris-Abraham-W. Snowdrop, de Pigtown (U.S.A.), était arrivé à l'âge de cinquante-cinq ans, sans que personne de ses parents ou amis eût pu l'amener à prendre femme. L'année dernière, quelques jours avant Noël, il entra dans le grand magasin du 37th Square (Objets artistiques en Banaloïd), pour y acheter ses cadeaux de Christmas.
<langue="th"> <nouvelle="001"><auteur="1"> หมอจอรีส อับราฮัม ดับเบิลยู สโนว์ครอป เมืองพิกทาวน์ (สหรัฐอเมริกา) ย่างเข้าสู่วัยห้าสิบห้าโดยไม่มีญาติโกโหติกาหรือเพื่อนสนิทผู้ใดสามารถโน้มน้าวให้เขาแต่งภรรยาได้ ปีที่แล้ว สามสี่วันก่อนวันคริสต์มาส หมอจอรีสเข้าไปซื้อของขวัญคริสต์มาสในห้างสรรพสินค้าย่านจัตุรัสสามสิบเจ็ด (ชิ้นงานศิลปะที่ทำด้วยพลาสถ่อย)
<langue="thsyl"><nouvelle="A001"><auteur="1"> หมอ จอ รีส __ อับ รา ฮัม __ ดับ เบิล ยู __ สโนว์ ครอป __ เมือง พิก ทาวน์ __ (สหรัฐ อเม ริ กา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โก โห ติ กา หรือ เพื่อน สนิท ผู้ ใด สา มารถ โน้ม น้าว ให้ เขา แต่ง ภรร ยา ได้ ปี ที่ แล้ว __ สาม สี่ วัน ก่อน วัน คริสต์ มาส หมอ จอ รีส เข้า ไป ซื้อ ของ ขวัญ คริสต์ มาส ใน ห้าง สรรพ สิน ค้า ย่าน จัตุ รัส สาม สิบ เจ็ด __ (ชิ้น งาน ศิล ปะ ที่ ทำ ด้วย พลาส ถ่อย)
<langue="thmor"><nouvelle="B001"><auteur="1"> หมอ จอ รีส __ อับ รา ฮัม __ ดับ เบิล ยู __ สโนว์ ครอป __ เมือง พิก ทาวน์ __ (สหรัฐอเมริกา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โก โห ติ กา หรือ เพื่อน สนิท ผู้ใด สามารถ โน้มน้าว ให้ เขา แต่ง ภรรยา ได้ ปี ที่ แล้ว __ สาม สี่ วัน ก่อน วัน คริสต์ มาส หมอ จอ รีส เข้า ไป ซื้อ ของขวัญ คริสต์ มาส ใน ห้างสรรพสินค้า ย่าน จัตุรัส สาม สิบ เจ็ด __ (ชิ้น งาน ศิลปะ ที่ ทำ ด้วย พลาส ถ่อย)
<langue="thlex"><nouvelle="C001"><auteur="1"> หมอ จอรีส ___ อับราฮัม ___ ดับเบิลยู ___ สโนว์ครอป ___ เมือง พิกทาวน์ ___ (สหรัฐอเมริกา) ย่าง เข้า สู่ วัย ห้า สิบ ห้า โดย ไม่ มี ญาติ โกโหติกา หรือ เพื่อน สนิท ผู้ใด สามารถ โน้มน้าว ให้ เขา แต่ง ภรรยา ได้ ปี ที่แล้ว ___ สาม สี่ วัน ก่อน วันคริสต์มาส หมอ จอรีส เข้า ไป ซื้อ ของขวัญ คริสต์มาส ใน ห้างสรรพสินค้า ย่าน จัตุรัส สาม สิบ เจ็ด ___ (ชิ้น งาน ศิลปะ ที่ ทำ ด้วย พลาส ถ่อย)

Tableau 1: Les différentes versions d'une nouvelle.

Guide de lecture du tableau 1

La première partie du Tableau 1 correspond à la version originale de la nouvelle Collage d' *Alphonse Allais*. La deuxième partie du tableau correspond à la version traduite en thaï. On remarque que le texte n'est globalement pas segmenté hormis quelques espaces ici ou là.

Chacune des parties suivantes a été segmentée par l'outil Kucut. Il a remplacé les espaces originels par la suite de caractères __ puis il a ajouté des espaces afin de délimiter les segments. La troisième, quatrième et cinquième partie du tableau correspondent aux versions thaïes segmentées respectivement en syllabes, morphèmes et unités lexicales.

3 Navigation dans les segmentations du thaï

Nous essayons de caractériser dans cette partie les différents niveaux de segmentation en observant leurs formes avec les outils statistiques de ***Lexico3***.

Principales caractéristiques

Partie	occurrences	formes	hapax	Fréq.Max	Forme
thsyl	110235	3991	1083	4125	__
thmor	98199	5978	2276	4125	__
thlex	89178	6493	2656	4125	__
Corpus	297612	8050	1353	12375	__

Tableau 2: Principales Caractéristiques Lexicographiques.

On observe[10] dans le Tableau 2 conformément à ce que l'on pouvait supposer que plus l'unité est petite telle la syllabe, plus la forme est en moyenne répétée et moins elle est susceptible d'être hapax. Inversement, plus l'unité est grande comme l'unité lexicale, moins la forme est répétée et plus il y a d'hapax. Le nombre élevé de syllabes différentes peut frapper mais sachant que le système d'écriture thaï peut théoriquement produire plus de 1.400.000 syllabes différentes[Ber2004], le nombre attesté est relativement faible.

Les sommations sur l'ensemble du corpus montrent que les parties ne sont pas au sens strict des partitions. En effet, il existe des formes et des hapax communs aux différentes parties.

Une dernière remarque concerne la forme la plus fréquente, le symbole __ qui représente les espaces présents initialement dans le corpus. Son utilisation reste fréquente bien que l'espace ne sert pas à séparer les mots[11].

Accroissement de vocabulaire

L'Illustration 2 montre les courbes d'accroissement de vocabulaire pour chacune des parties. On observe une forte corrélation entre les courbes des morphèmes et des unités lexicales. L'écart entre ces deux courbes tend à se stabiliser plus on avance dans le corpus alors que la courbe des syllabes a un comportement différent, elle se tasse beaucoup plus rapidement. On observe cependant dans deux secteurs du corpus, entourés en gris, une accélération de l'accroissement du vocabulaire pour chacune des parties. Ceci indique que l'apport de nouveaux mots et de nouveaux morphèmes est en partie réalisé par l'apport de nouvelles syllabes. Peut-être s'agit-il de mots empruntés transcrits comme des noms propres ?

---

[10] Nous rappelons que les partitions thsyl, thmor, thlex correspondent au corpus segmenté respectivement en syllabes, morphèmes et unités lexicales.

[11] Une étude textométrique de son usage à travers par exemple des concordances serait intéressante à mener ultérieurement.

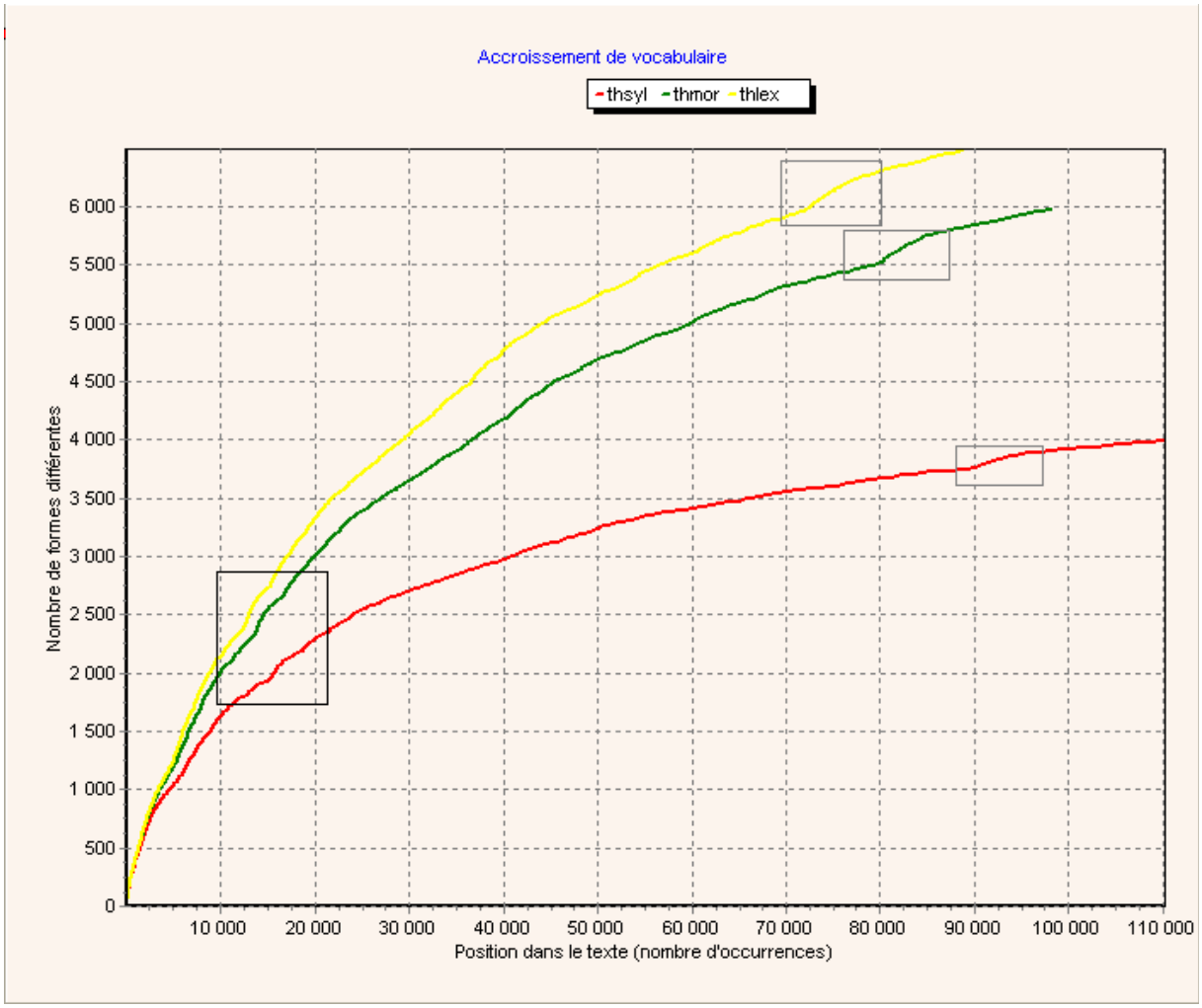


Illustration 2: Courbes d'accroissement de vocabulaire

Diagramme de Pareto

Le diagramme de Pareto, Illustration 3, montre que les syllabes, les morphèmes et les unités lexicales suivent à peu près la loi de Zipf. Il confirme que les syllabes sont plus utilisées que les morphèmes, ceux-ci plus utilisés que les unités lexicales. Cependant on observe que les courbes des morphèmes lexicaux et des unités lexicales sont très proches alors que celle des syllabes est un peu plus éloignée.

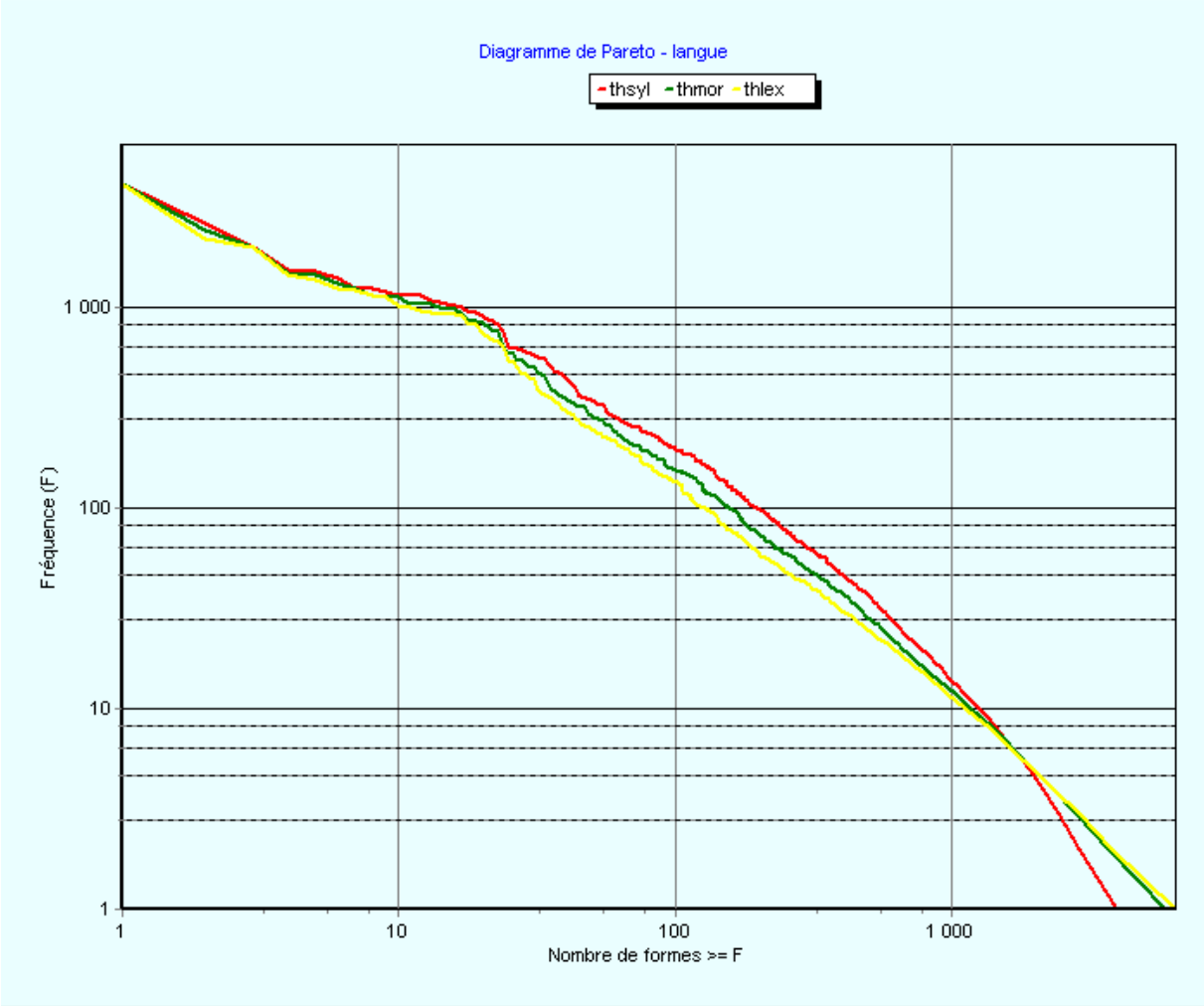


Illustration 3: Diagramme de Pareto

Les segmentations vues par les spécificités

Nous avons observé dans les parties précédentes que les syllabes et plus encore les morphèmes et les unités lexicales ont des comportements textométriques assez similaires. Par conséquent, nous allons utiliser les spécificités[12] de chacune des parties par rapport au corpus entier pour faire ressortir les formes spécifiques à chaque partie.

Les syllabes

Le Tableau 3 montre dans le volet gauche les cinq formes les plus spécifiques de la partie thsyl qui n'est autre que les traductions thaïes segmentées en syllabes. Il montre dans le volet droit les cinq formes les moins spécifiques de cette partie. On observe que toutes ces formes sont monosyllabiques. De plus, on remarque que les occurrences des formes du volet gauche sont presque exclusivement dans cette partie alors que les occurrences des formes du volet droit ne compte que pour un tiers des occurrences totales.

---

12 Nous avons retenu un seuil de probabilité de 5 et une fréquence minimale de 10.

Syllabes	*Spécificités positives*					*Spécificités négatives*		
Formes	Fréq.	Fréq.Tot.	Coef.		Formes	Fréq.	Fréq.Tot.	Coef.
ระ	256	301	***		ผม	1980	5938	-10
กระ	437	556	***		และ	1248	3735	-7
อะ	211	216	***		จะ	1144	3400	-6
ประ	423	536	***		เธอ	1151	3441	-6
สา	174	181	***		เขา	1013	3028	-6

Tableau 3: Spécificités pos/nég thsyl sur thsyl+thmor+thlex

On peut corroborer ce constat numérique par des connaissances linguistiques. En effet, il semble difficile d'attribuer un sens aux formes de gauche alors qu'on sait par connaissance du thaï qu'elles sont présentes dans de nombreuses unités lexicales. On peut confirmer cette affirmation par une recherche à l'aide de l'outil groupe de formes. Quelques exemples sont donnés dans le Tableau 4. Quant à la présence d'occurrences de ces formes dans les parties thmor ou thlex, il peut s'agir d'erreur de segmentation.

Motif : ^ระ 40 formes.	Fréq.	Motif : ระ$ 12 formes.	Fréq.	Motif : .+ระ.+ 198 formes.	
ระหว่าง	60	พระ	39	กระ.+	85
ระฆัง	26	ธุระ	7	ประ.+	66
ระดับ	13	กระ	6	Motif : ^[^ค|ป].+ระ.+ 33 formes.	
ระบม	1	วาระ	2	จนกระทั่ง	8
ระบาด	1	ขรุขระ	1	ถึงกระนั้น	8
ระบายสี	1	ตระกรุมตระ	1	หลักประกัน	1

Tableau 4: Extraits de résultats de recherche de groupe de formes dans la partie thlex

En revanche, les formes du volet droit sont bien connues comme unité lexicale. Par exemple les formes ผม /phom/, เธอ /theu/ et เขา /khaw/ peuvent être utilisées comme des substituts du nom (je, tu/elle/il, il/elle/ils/elles) ou avoir une valeur lexicale (cheveux, ,montagne) quant à และ /lè/ c'est une conjonction de coordination étant presque équivalant à notre « et ». Le จะ /ja/ est une particule marquant l'inaccompli.

On peut confirmer cette connaissance linguistique par l'utilisation du concordancier pour décompter ces formes par partie. Par exemple, cela donne pour la forme ผม les résultats suivants : thsyl, 1980 ; thmor, 1979 ; thlex, 1979, confirmant ainsi le statut de syllabe, morphème lexicale et unité lexicale de cette forme.

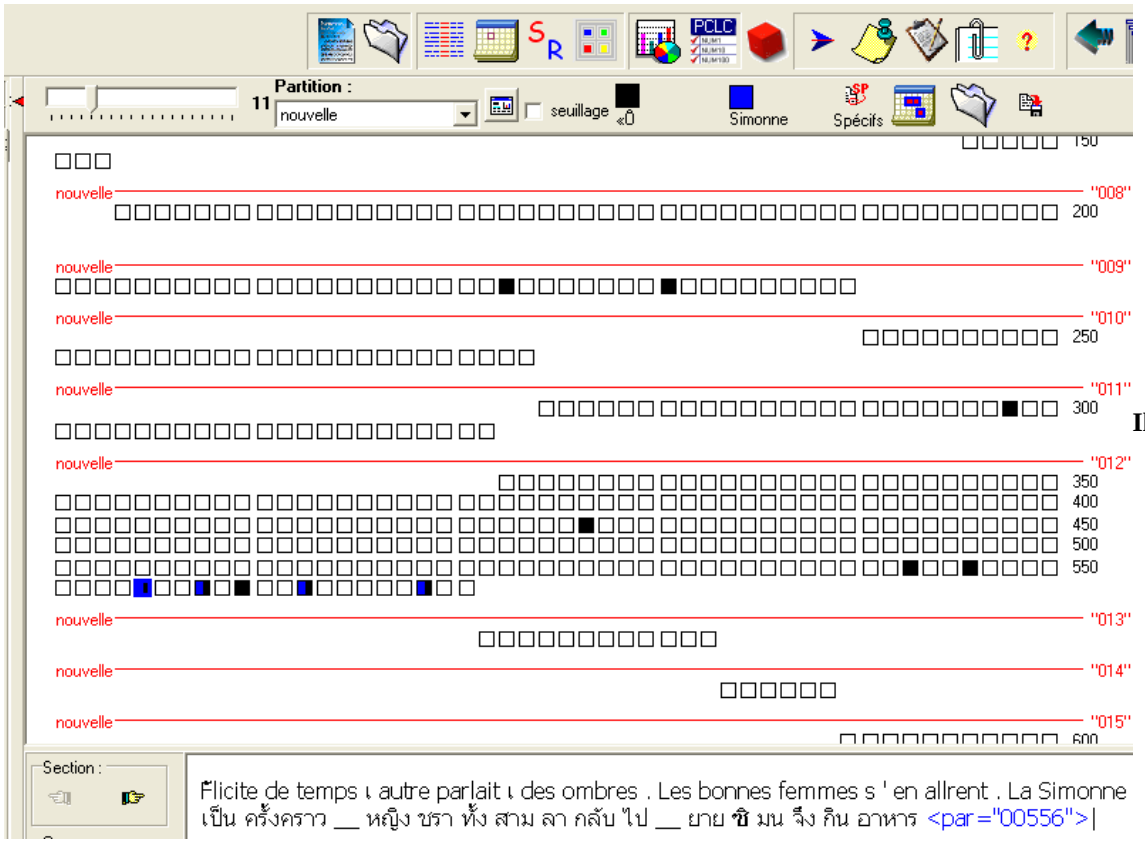
Les morphèmes lexicaux

Le Tableau 5 montre dans le volet gauche les cinq formes les plus spécifiques de la partie thmor qui n'est autre que les traductions thaïes segmentées en morphèmes lexicaux. Ceux-ci étant défini par le segmenteur comme la plus petite unité ayant un sens selon son dictionnaire. Il montre dans le volet droit les cinq formes les moins spécifiques.

Morphèmes	*Spécificités positives*						*Spécificités négatives*		
Forme	thmor	thlex	thsyl	Fréq.Tot.	Coef.	Forme	thmor	Fréq.Tot.	Coef.
อะไร	205	205	0	410	13	อะ	4	216	-32
ซิ	184	12	184	381	10	สา	6	181	-24
เวลา	140	143	0	283	9	เว	8	172	-20
ลิ	108	0	108	216	8	ตะ	1	117	-20
มาดาม	147	167	0	314	8	วิด	2	119	-18

Tableau 5: Spécificités pos/neg thmor sur thsyl+thmor+thlex

On observe dans le volet gauche trois formes composées de deux syllabes อะ-ไร/a-raj/, เว-ลา/we-la/ et มา-ดาม /ma-dam/ et deux formes composées d'une seule syllabe : ซิ /si/ et ลิ /li/. Le nombre de syllabes est aussi déductible par l'observation de la distribution des fréquences selon les parties. Les morphèmes dissyllabiques sont clairement des morphèmes lexicaux, en effet on a อะไร (pronom interrogatif), เวลา (Le temps) et มาดาม qui est une translittération de madame.



ation 4: Carte des sections pour ชิ et Simon-ne

Que sont les 12 occurrences de ชิ dans la partie thlex ? Une concordance groupée par nouvelle montre que sur les douze occurrences de ชิ, cinq, répartie dans quatre nouvelles, ont une autonomie réelle en tant que particule d'insistance. Comme le montre la carte des sections , Illustration 4, les sept autres occurrences sont localisées dans une seule nouvelle et n'ont qu'une valeur syllabique en tant que constituant d'un nom propre nom reconnu par le segmenteur ชิมน / Simonne. Les deux formes ชิ /si/ et ลิ /li/ sont apportées majoritairement par le prénom Félicité/เฟลิชิเต. Le segmenteur essaie de reconstituer les mots inconnus uniquement lors de la segmentation en unités lexicales, il est donc normal de retrouver les formes ลิ et ชิ dans la partie thmor lors de la segmentation en morphème lexical. En revanche, laisser telles quelles les occurrences de ชิ dans la partie thlex lorsqu'elles font parties du prénom ชิมน est clairement une erreur du segmenteur. Le problème spécifique de la reconnaissance des noms propres sera abordé ultérieurement.

Les formes du volet droit sont toutes monosyllabiques et ont une fréquence faible dans la partie thmor. On observe que deux des formes อะ /a/, เว /we/ sont des syllabes des formes อะไร /araj/ et เวลา /wela/ respectivement. Les quatre occurrences de อะ dans la partie thmor proviennent essentiellement d'emprunt dont certaines des syllabes sont connues comme des mots thaïes ainsi อะกู (aku, pronom malais signifiant je) où la syllabe กู signifie aussi je (familier) en thaï. La plupart de ces formes, à l'exception des instances de noms propres, seront reconstituées correctement dans la partie thlex.

Les unités lexicales

Le Tableau 6 montre dans le volet gauche les cinq formes les plus spécifiquement employées comme unité lexicale et dans le volet droit, les cinq formes les moins spécifiques.

Unité lexicale	Principaux sens	*Spécificités positives*					*Spécificités négatives*		
Forme		Fréq	Fréq.Tot.	Coef.		Forme	Fréq	Fréq.Tot.	Coef.
รู้สึก	Ressentir, sentir, sentiment,*	233	335	***		สึก	1	369	***
เฟลิซิเต	Félicité	76	76	41		อา	8	472	***
ทำให้	Causer, faire en sorte de ...	225	355	39		ประ	3	536	***
กำลัง	Modificateur d'aspect temporel/ pouvoir:N	159	255	27		กระ	6	556	***
ตัวเอง	Pronom personnel réflexif.	111	157	26		ใจ	82	988	***

Tableau 6: Spécificités pos/neg thlex sur thsyl+thmor+thlex

On observe que toutes les formes de gauche sont polysyllabiques alors que celles de droite sont monosyllabiques.

Les formes de gauche sont variées quant à leur nature. En effet, nous avons un verbe, un nom commun, un nom propre ainsi que des mots outils[13]. On remarque que le mot outils ทำ-ให้ /tham-haj/ est composé de deux syllabes dont l'une est principalement un verbe ทำ(faire) rentrant dans la composition d'un nombre assez important d'unités lexicales et l'autre est aussi un mot outil dérivé du verbe ให้ /haj/ (donner). Ils sont très fréquents. Par exemple ทำ apparaît dans les parties thsyl, thmor, thlex respectivement 581, 360, 210 fois et ให้ respectivement 1002, 854 et 748 fois.

Les fréquences des formes de droite, hormis celle de la forme ใจ[14], sont faibles dans cette partie. Ainsi il n'y a qu'une seule occurrence de สึก /seuk/ contre 132 dans la partie thmor et 236 dans la partie thsyl. On ne manquera pas de remarquer qu'elle rentre en composition dans la forme รู้สึก /ruuseuk/, celle-ci apparaît 102 fois dans la partie thlex, ce qui nous permet de déduire par calcul que la séquence รู้ สึก apparaît 131 fois dans la partie thmor[15]. Il existe donc

---

[13] Conformément à l'expression utilisée dans la méthode de langue intitulée « Pratique du Thaï » de Wanee Pooput et Michèle Conjeaud.

[14] La formeใจ /cai/ est une des plus belles formes du thaï dont le sens est associé à celui de cœur au sens propre comme figuré. Je laisse son étude aux doctorants ou au romancier amoureux du thaï (cf. http://www.learningthai.com/hearttalk.htm).

[15] On notera au passage que la segmentation en morphèmes lexicaux n'est pas stable puisqu'il n'y a pas de raison de découper la séquence รู้สึก tantôt en รู้สึก tantôt en รู้ สึก. Cela n'est pas très grave car l'étape morphème

une occurrence de สึก dans la partie thlex et une dans la partie thmor. L'utilisation des concordances groupées montre que c'est la même.

On a remarqué précédemment que les formes ประ /pra/ et กระ /kra/ participaient en tant que syllabe à la formation de nombreux mots (cf. Tableau 4) mais il s'avère que ces deux formes ont aussi une signification autonome attestée par l'existence d'entrée dans différents dictionnaires. Cependant, il reste à confirmer le statut de leurs occurrences dans nos textes.

Formes	Occurrences dans le texte		Entrées de dictionnaire
ประ	หลังคา **ประ** ทุน	capote	หลังคา N:toit **ประ**ทุน N:couverture
ประ	**ประ** เหมาะ	convenir	**ประ** V: ? เหมาะ Adj:être adapté **ประ**เหมาะ : Abs. dico.
ประ	ผู้ **ประ** เหมาะ เคราะห์ ซวย	La malheureuse	ผู้ : N:personne **ประ** : V:? เหมาะ : V: être fait pour เคราะห์ : N chance ou malchance ซวย : Adj malchanceux
สึก	เซาะ เสีย จน **สึก**	usés	เซาะ V:éroder เสีย : V: être abîmé จน : prép. Jusqu'à **สึก** : V: être éroder

Tableau 7: Occurrences en contexte d'unités lexicales les moins spécifiques

Le Tableau 7 montre les occurrences des formes ประ /pra/ et สึก /seuk/ de la partie thlex, c'est-à-dire considérées comme une unité lexicale après segmentation du texte original. On voit que leur statut respectif n'est pas simple puisqu'à chacune des séquences où apparaissent ces formes correspond un seul mot source français. La première ligne du tableau montre que la séquence est mal segmentée puisque les formes **ประ** /pra/ et ทุน /thun/ auraient dû être fusionnées en **ประ**ทุน/prathun/ conformément à l'entrée des dictionnaires. Quant à savoir si les formes หลังคา /langka/ et **ประ**ทุน doivent être fusionnées, il s'agit d'un autre problème.

La deuxième occurrence de ประ laisse à penser que ce sont bien deux unités séparées car la séquence **ประ** /pra/ เหมาะ /maw/ n'est attestée dans aucun de nos dictionnaires[16]. Cependant le sens de ประ est légèrement modifié par rapport aux différents sens donnés par ces dictionnaires.

La troisième occurrence de ประ ajoute encore au doute. En effet, on retrouve de nouveau la séquence **ประ** เหมาะ. En outre, on observe la séquence เคราะห์ /khraw/ et ซวย /suey/ qui est une accumulation de deux formes au sens proche ce qui légitimerait la composition en เคราะห์ซวย. Quant à la forme ผู้ /phou/ elle est souvent décrite dans les méthodes de langues comme un préfixe permettant la création de nombreux mots relatifs à une personne. Ainsi, si la forme

lexical pour le segmenteur est une sorte de pré-traitement pour constituer les unités lexicales. Ce n'est pas une analyse d'une unité lexicale en morphèmes.

16 Voir la liste des dictionnaires utilisés dans les références.

เขียน /khien/ signifiant écrire est précédée de ผู้ pour former ผู้เขียน, le tout signifie auteur, à ne pas confondre avec écrivain qui s'écrit นักเขียน. Ceci laisse à penser que la séquence complète de la troisième ligne constitue une seule unité lexicale construite à des fins littéraires mais dont le sens est parfaitement décomposable.

Nous voyons donc que la notion d'unité lexicale n'est pas simple et que les spécificités donc le segmenteur, ne se sont pas trompées en nous présentant la forme ประ comme peu représentative d'une unité lexicale et en nous présentant les noms propres et les mots outils comme des unités lexicales. Toutefois, on peut s'interroger sur la pertinence de la segmentation des séquences plus longues comme celles du Tableau 7.

Bilan de la navigation

Les observations faîtes sur les courbes d'accroissement de vocabulaire à savoir que les accroissements de syllabes, de morphèmes et d'unités lexicales sont corrélés, ont été confirmées par l'analyse des spécificités par partie. Ainsi on a vu que les syllabes les plus spécifiques rentrent dans la composition de nombreuses formes polysyllabiques ayant autant le statut de morphème lexicale que d'unité lexicale. On a aussi observé que certaines syllabes très fréquentes sont aussi des morphèmes et des unités lexicales notamment des mots à usage grammatical comme les substituts du noms.

On a aussi montré qu'il ne fallait pas trop se fier à la partie morphème lexicale lorsqu'il s'agissait d'analyser la composition d'une unité lexicale car bien souvent la segmentation était instable : soit l'unité lexicale apparaissait telle quelle, soit elle apparaissait segmentée.

Conformément à la description de cette méthode employée par le segmenteur[Asa2003], la segmentation en morphèmes lexicaux doit être vue comme une étape intermédiaire vers la construction des unités lexicales à partir des syllabes.

Enfin, l'observation des spécificités sur la partie unité lexicale a montré que si les mots outils, les noms propres semblent constituer le gros des unités lexicales c'est que les frontières des unités composées ne semblent pas très nette.

4 Les maux de l'unité lexicale

On vient d'observer que la nature des formes les plus spécifiques de la partie thlex est variée (noms propres, mots outils, verbe). Cependant, si on sélectionne les quinze premières formes au lieu de cinq, on remarque une large prédominance des noms propres. Ces formes complémentaires sont consignées dans le Tableau 8.

L'identification des noms propres et notamment des personnages est intéressante puisque notre corpus est constitué de nouvelles françaises traduites en thaï. L'enjeu est donc la restitution des noms de personnes, mots vraisemblablement inconnus des dictionnaires du segmenteur mais dont la limite signifiant/signifié est claire.

Par conséquent, nous utiliserons dans un premier temps les outils de ***Lexico3*** pour vérifier si les occurrences de noms propres ont été correctement identifiées et analyser, le cas échéant, les problèmes de non reconnaissance. Dans un deuxième temps nous essayerons de saisir la complexité de la notion d'unité lexicale en l'illustrant par un exemple tiré des formes les plus spécifiques, à priori simple, la forme อาหาร /ahan/(aliment, nourriture).

Forme	Principaux sens	Fréq/Fréq.Tot.		Forme	Principaux sens	Fréq/Fréq.Tot.
เฟลิซิเต	Félicité	76/76		รู้สึก	Ressentir, sentir, sentiment,*	233/335
โอแบง	Aubain	40/40				
มาร์เกอริต	Marguerite	39/39		อาหาร	Repas, diner, aliment,...	88/131
ปารีส	Paris	35/35		ประตู	Porte:V, *	75/110
จิเยร์	Gier	31/31				

Tableau 8: Formes extraites parmi les 15 unités lexicales les plus spécifiques.

Problèmes de segmentation des noms propres

Le Tableau 8 montre que les formes référençant des noms propres situées dans le volet gauche n'apparaissent que dans la partie thlex. Ceci indique qu'elles ont été découpées différemment dans la partie thmor. C'est à première vue surprenant puisque ce sont des mots empruntés donc impossible à analyser morphologiquement mais il faut garder à l'esprit que le segmenteur n'analyse pas en morphèmes les unités lexicales. En effet, il découpe d'abord le texte en syllabes, puis en morphèmes lexicaux enfin recompose les unités lexicales à partir de ces morphèmes.

Ceci étant dit, on peut avoir affaire à deux problèmes. Le premier est un problème de sous-segmentation c'est-à-dire que des parties de noms propres sont rattachées à d'autres unités lexicales. Le second est un problème de sur-segmentation c'est-à-dire que des bouts de morphèmes de noms propres n'ont pas été rattachés ensemble.

La méthode pour retrouver des occurrences de formes mal segmentées avec ***Lexico3*** consiste à calculer les segments répétés sur le corpus segmenté en trois parties puis à utiliser conjointement l'outil de recherche de groupe de formes et les expressions rationnelles.

Expressions	Exemples de formes		Expressions	Exemples de formes
.*เฟลิซิเต.*	**เฟลิซิเต** **เฟลิซิเต**ก็อด **เฟลิซิเต**พามันออกไป **เฟลิซิเต**ไป **เฟลิซิเต**ไอ **เฟลิซิเต**รักการ . . . **เฟลิซิเต**ออกท่า		.*จิเยร์.*	**จิเยร์** มาดาม กรอง **จิเยร์** เจ้า หนู กรอง **จิเยร์** กรอง **จิเยร์** มอง เมอซิเออร์ กรอง **จิเยร์**
14 formes trouvées, la plupart sont des hapax.			14 segments répétés, distribution variée.	

Tableau 9: Groupe de formes avec segments répétés

Les résultats de la recherche consignées dans le Tableau 9 montrent un problème de sous-segmentation pour la forme เฟลิซิเต /félicité/ puisque nous avons trouvé un certain nombre de formes contenant เฟลิซิเต. On voit tout l'intérêt d'utiliser les segments répétés puisqu'on remarque que la forme จิเยร์ /jier/ n'est pas un nom propre. Le vrai nom propre est กรองจิเยร์ (Grangier) puisqu'en contexte, la séquence est précédée de มาดาม (Madame), เมอซิเออร์ (Monsieur), ou เจ้า หนู (le petit [Grangier]). On vient donc d'identifier un problème de sur-segmentation.

Le problème de sur-segmentation de la forme กรองจิเยร์ s'explique partiellement par le fait que la forme กรอง /krong/ est un mot thaï. On utilise la carte des sections pour trouver des occurrences de กรอง n'apparaissant pas en coocurrence avec la forme จิเยร์.

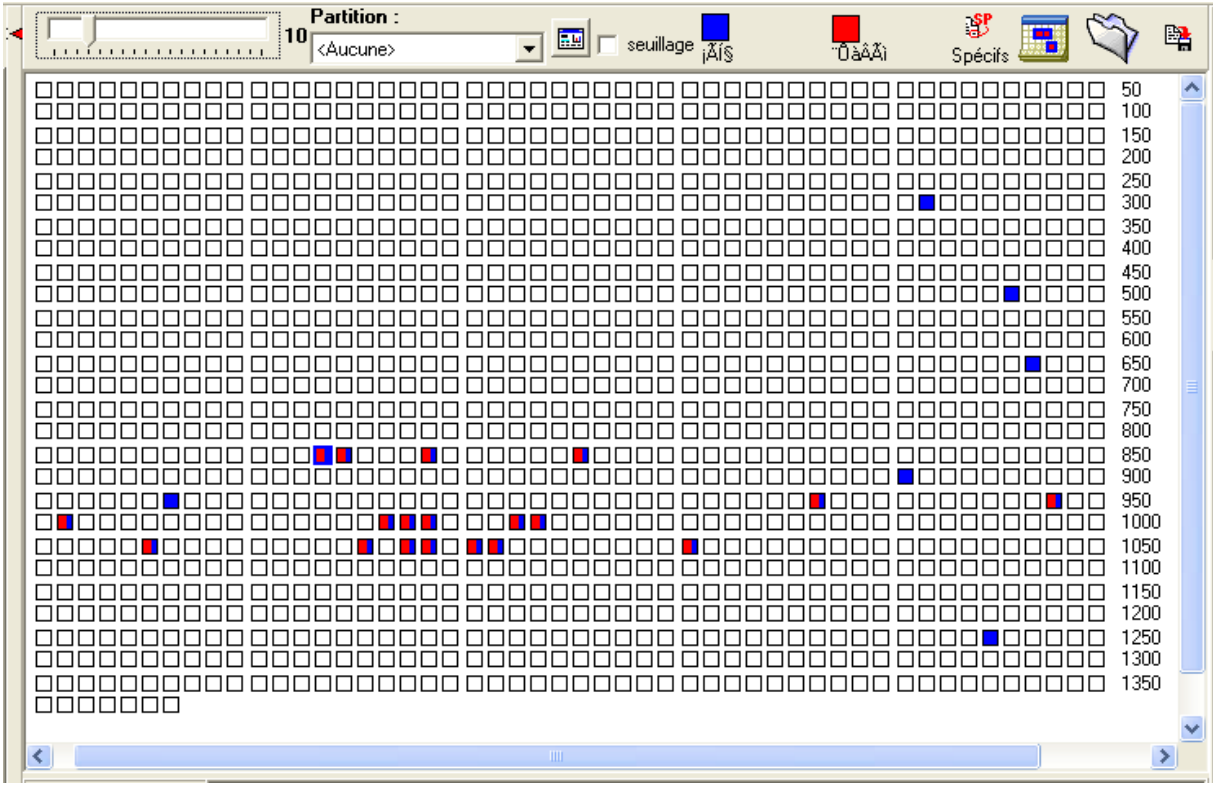


Illustration 5: Carte des sections : en bleu กรอง, en rouge จิเยร์

On remarque que la forme กรอง /krong/ apparaît sans la forme จิเยร์ /jier/ dans six sections. On a pu répartir ces occurrences de กรอง en trois groupes après analyse.

Deux occurrences réfèrent à l'unité lexicale thaïe dont le sens attesté par nos dictionnaires est le verbe filtrer. Dans nos textes, elles sont en relation de traduction avec le nom commun : filtre.

Trois autres occurrences sont des erreurs de segmentation concernant des noms de lieu non reconnus : ถนน **กรอง** ปง | rue Grand – Pont, ที่ **กรอง** วิลล์ | à Granville, ที่ **กรอง** วิลล์ | à Granville. La forme กรอง /krong/ est ici une transcription approximative du son gran qui n'existe pas en thaï.

Enfin, la dernière occurrence n'aurait pas dû exister. En effet nous avons le segment suivant เภสัช **กรอง** ฟร้ว pour Onfroy l'apothicaire, qui est une erreur de segmentation de niveau caractère. En effet la séquence aurait dû être segmentée de la façon suivante : เภสัช**กร** **อง**ฟร้ว | apothicaire Onfroy.

Nous tenons à faire remarquer que nous n'avons pas utilisé le segmenteur dans ses conditions optimales puisque pour résoudre les problèmes de mots inconnus, il utilise une méthode de segmentation basée sur des statistiques globales et locales. On aurait probablement gagné en précision si on avait segmenté nouvelle par nouvelle au lieu du corpus

dans sa globalité. Ainsi les occurrences de กรอง /krong/ dans les autres nouvelles n'auraient peut-être pas interféré avec celles liées à จิเยร์ /gier/.

Cette exploration des noms propres a permis d'expliciter quelques problèmes de segmentation provoqué par le fait que les formes empruntées sont composées de syllabes correspondant à des mots thaï. Ces problèmes ne concernent pas uniquement des textes traduits mais aussi les textes proprement thaïs puisque bien souvent les noms et prénoms thaïs sont des noms venant du sanskrit et du pâli ayant leur propre sens notamment dans le domaine religieux et royal.

Globalement le segmenteur basé sur des méthodes statistiques a réussi à correctement segmenter de nombreuses occurrences de noms propres. Peut-être qu'un post-traitement symbolique de reconnaissance d'entités nommés permettrait d'améliorer cette segmentation.

Problèmes de composition lexicale

Le Tableau 8 montre que les formes du volet droit ont une distribution différente des formes nominales du volet gauche. Par exemple, la forme /ahan/ apparaît 88 fois dans la partie thlex et seulement 43 fois dans la partie thmor. On a déjà expliqué ce phénomène précédemment. De plus, ce qui nous intéresse pour la suite de cette étude est de trouver des formes ou des segments répétés dans la partie thlex contenant la forme /ahan/ afin de déterminer la limite de l'unité lexicale.

La méthode pour retrouver ces formes avec ***Lexico3*** consiste à calculer les segments répétés sur le corpus aligné puis à utiliser conjointement l'outil de recherche de groupe de formes et les expressions rationnelles comme dans l'Illustration 6.

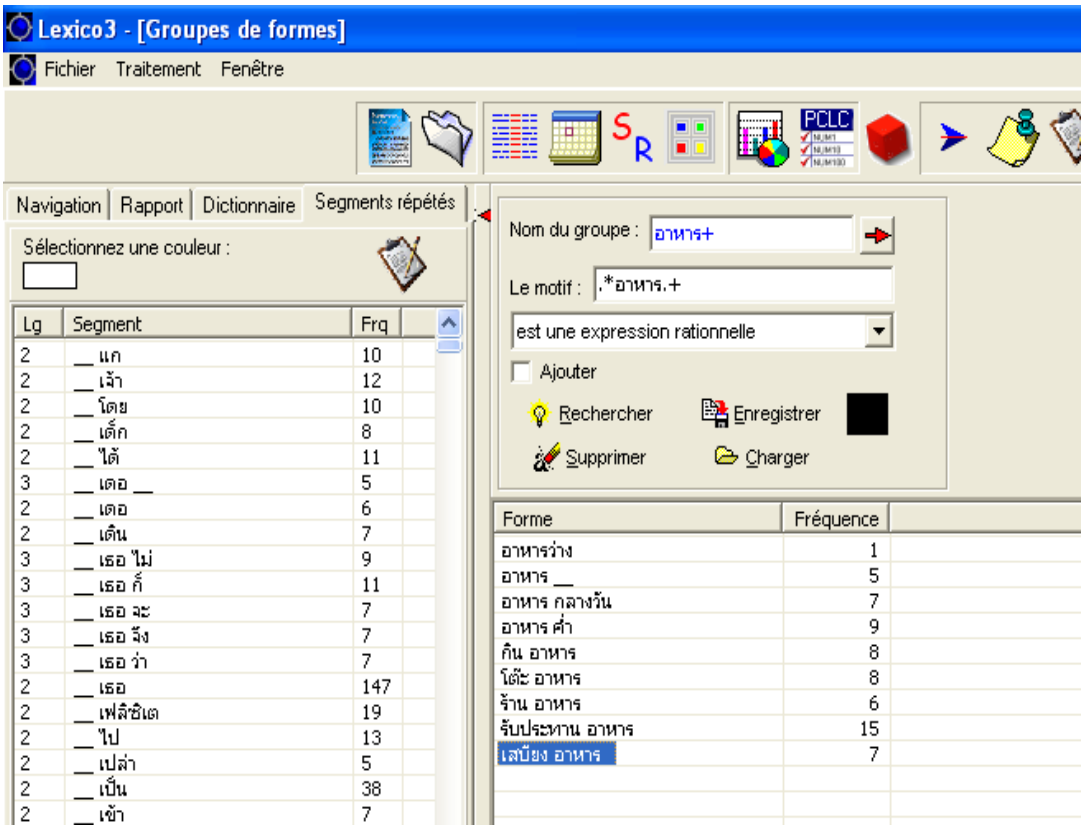


Illustration 6: Recherche groupe de formes, segments répétés, อาหาร

On voit déjà apparaître quelques segments intéressants mais pour compléter la recherche on réalise un inventaire distributionnel sur l'ensemble du groupe. Une fois que nous disposons de ces formes composées, on recherche l'expression correspondante source dans les textes français afin de déterminer le sens en contexte. Pour analyser les résultats, on construit une matrice dite de composition lexicale (cf Tableau 10) où les formes de la première colonne se combinent avec certaines formes de la première ligne pour traduire un mot source.

+	**อาหาร**	**อาหาร** กลางวัน	**อาหาร** เย็น	**อาหาร** ค่ำ
_	nourriture aliment (หุง หา+) faire la cuisine (เหล้า ย่อย+) eau de vie (เสิร์ฟ+) อาหาร servir	(ยก||)déjeuner:N		Souper:N, (เวลา+)(heure [du]) dîner:N, dîner:N
รับประทาน	(ใน ระหว่าง ที่+) [pendant le] repas:N, (ร่วม โต๊ะ+) manger à table, manger,dîner:V	(เวลา+)(heure [du])Déjeuner :N, (นั่ง ร่วม วง+||หลังจาก) Déjeuner :V,	(ร่วม+) Dîner:V (หลัง+)Dîner:N	Dîner:V
กิน	Manger:V,déjeuner:V, (+เลิศ รส) être nourrie de	(นั่ง+) déjeuner:V		Dîner:V
โต๊ะ	Table:N, (นั่ง+) [se mettre à] table			
เสบียง	Réserve:N, provisions:N, (การ รับประทาน+) [manger ses] provisions:N			
ร้าน	Cabaret, Restaurant (+ถูก ๆ) gargote:N,			

Tableau 10: Matrice de composition lexicale

Le Tableau 10 est une matrice de composition lexicale. On la lit en combinant les formes de la première colonne avec les formes de la première ligne. Par exemple _ + อาหาร = nourriture. Cela signifie que le sens de la forme à la place du caractère _ était clairement séparé de celui de la forme อาหาร. De plus une séquence entre parenthèse précise le contexte, par exemple (นั่ง+) + โต๊ะ + อาหาร = [se mettre à] table ou ร้าน + อาหาร + (+ถูก ๆ) = gargote.

D'après cette matrice en regardant la première ligne et la première colonne on peut isoler le sens de อาหาร comme étant nourriture ou aliment.

La séquence หุง หา **อาหาร** (hung ha ahan) existe en entrée de dictionnaire avec le sens de cuisiner dans un niveau de langue littéraire. On peut décomposer cette séquence de la manière suivante หุง /hung/ : cuisiner (attesté dans le même dictionnaire), หา /ha/ : Il existe en tant que verbe à multiple sens (chercher), mais je pense qu'ici il a une valeur euphonique (ha) plus que sémantique. อาหาร/ahan/ **:** ici nourriture. Il sert de complément à หุง de la même manière que pour เสิร์ฟ (servir). Il va de soi que si หา /ha/ a une valeur euphonique alors il faut considérer l'expression entière comme une seule unité lexicale.

La séquence เหล้า (liqueur) ย่อย (digérer) อาหาร (nourriture) n'est pas attestée dans les dictionnaires, mais il semble que ce soit un bon équivalent du mot digestif si on calcule le sens global à partir de chaque unité.

La forme ร้าน /ran/va permettre d'illustrer en corpus la notion de termes génériques bien connues des étudiants de thaï. En effet, à opposer à la séquence ร้าน + อาหาร (nourriture) = restaurant nous avons dans le corpus les séquences :

- ร้าน + กาแฟ (café (boisson))=café (le lieu), ร้าน /ran/+ ขาย /khaj/ (cf Tableau 11)
- ร้านค้า (boutiques, commerces, …) ค้า (V: commercer, marchander, ...)
- ร้าน : On trouve quelques occurrences isolées mais toujours en coocurrence dans un paragraphe avec une autre des formes composées. La seule séquence isolée dans une nouvelle est งาน (fêtes) ออก (sortir) ร้าน qui est utilisé pour traduire fête foraine.

+	_	เหล้า (alcool)	เนื้อ (viande)	ยา (médicament)	ของเก่า (vieille chose)	...
ร้าน + ขาย(vendre)	boutique	estaminet	boucherie	pharmacie	brocante	...

Tableau 11: Composition lexicale ร้าน + ขาย + X

On peut déterminer le sens de ร้าน /ran/ à partir de ces exemples. C'est un terme générique désignant un local dans lequel s'exerce une activité commerciale. Il peut s'utiliser avec une certaine autonomie, ce n'est donc pas un préfixe au sens de l'analyse morphologique mais la plupart du temps il est spécialisé par un ou plusieurs autres morphèmes lexicaux.

Le fait que le segmenteur a traité différemment ร้านค้า/rankha/ des autres formes composées de ร้าน s'explique certainement par le modèle statistique utilisé, basé sur le score d'information mutuelle. Toutes ces formes devraient être traitées de la même manière qu'on les considère comme une seule unité lexicale ou comme plusieurs. Si on se réfère à l'article [Asa2002], il est possible que les auteurs du segmenteur considèrent une séquence débutant par ร้าน comme un syntagme nominal et non comme une unité lexicale puisque la séquence ne fait que préciser le sens de ร้าน sans changer le concept fondamental auquel il réfère.

On a dénombré six occurrences de เสบียง /sabieng/ อาหาร répartie dans deux nouvelles dont cinq comme traductions de provisions au sens de produits alimentaires et une comme traduction de réserves au sens de réserves alimentaires en coocurrence dans le même paragraphe avec provisions. On a aussi dénombré trois occurrences de เสบียง sans อาหาร comme traduction de provisions. De plus, nous avons aussi relevé une occurrence de la séquence suivante การ/kan/ รับประทาน /rappratan/ เสบียง /sabieng/ อาหาร /ahan/ comme traduction de provisions. On a aussi relevé la séquence ตู้/tou/ เสบียง/sabieng/ comme traduction de wagon-restaurant. On en déduit donc que la présence de อาหาร à la suite de เสบียง n'est pas obligatoire à la construction du sens mais servirait plutôt un but littéraire.

Le Tableau 11 laisse clairement apparaître deux autres oppositions que nous ne détaillerons pas. Il s'agit de l'opposition entre รับประทาน /rappratan/ et กิน /kin/ qui est normalement une opposition de registre de langue, l'emploi de รับประทาน étant plus soutenu que กิน. La seconde opposition concerne อาหาร กลางวัน(milieu du jour), อาหาร เย็น (soirée:N/frais:ADJ), อาหาร ค่ำ(N:nuit) où les trois formes viennent préciser อาหาร en ajoutant une information temporelle.

Cette partie a montré quelques problèmes de composition lexicale puisque même si อาหาร /ahan/ est décrit par les spécificités et donc par le segmenteur comme une des formes les plus

représentative de l'unité lexicale, on a vu bien des cas où elle rentre en composition avec d'autres unités lexicales pour être en relation de traduction avec un seul mot français.

Toute cette analyse doit nous permettre de réinterpréter les courbes d'accroissement de vocabulaire Illustration 2 page 7. En effet, les morphèmes lexicaux, sans parler de l'instabilité de cette segmentation, n'est qu'une étape intermédiaire de la syllabe vers l'unité lexicale. L'unité lexicale regroupe les morphèmes lexicaux parmi lesquels certains on été recomposés en noms propres et en mots composés de certains types. Toutefois de nombreuses séquences pouvant être considérées comme unité lexicale vis-à-vis du référentiel sémantique français telle celle commençant par la forme ร้าน /ran/ n'ont pas été recomposées. Sous l'hypothèse que ces séquences s'apparentent à des syntagmes nominaux, ce segmenteur thaï imite les segmenteurs pour les langues à écriture segmentée en ne les recomposant pas, laissant, si besoin est, le soin à un analyseur morpho-syntaxique de les reconstituer. Mais quelle est la différence réelle entre syntagme nominal et unité lexicale dans une langue dite isolante qui n'isole rien à l'écrit ?

5 Conclusion

Cette première étude a illustré en corpus l'intrication entre les syllabes, les morphèmes lexicaux et l'unité lexicale en thaï et par conséquent certains problèmes de segmentation qui en découlent. La méthode originale d'utilisation des outils de ***Lexico3*** tel que le calcul des spécificités par partie, segmentée selon un niveau, pour faire émerger des formes spécifiques ainsi que l'utilisation des segments répétés associée à la recherche par expression rationnelle à permis de trouver des exemples pertinents.

L'analyse des formes ainsi repérées et de leurs contextes à permis de préciser la manière dont travaille le segmenteur. Ainsi les syllabes semblent correctement segmentées. La segmentation en morphèmes lexicaux ne constitue pas véritablement une analyse morphologique mais une étape intermédiaire vers la construction des unités lexicales. Enfin, il semble que la segmentation en unités lexicales ne corresponde pas à la plus grande composition lexicale possible au point de ne plus distinguer l'unité lexicale du syntagme nominal mais à la composition de morphèmes lexicaux en une unité dont le sens n'est pas vraiment calculable à partir de ceux-ci. C'est aussi l'étape de reconstitution des mots inconnus tels les noms de personnes qui sont imparfaitement mais assez bien reconstitués.

Cette étude a donc montré qu'il était possible en utilisant ce segmenteur de réaliser une étude textométrique avec ***Lexico3*** mais qu'il fallait prendre quelques précautions quant à la définition de l'unité lexicale notamment dans le cas d'étude comparative.

6 Références

[Tha1978] Kobkool THAWARANON, 1978.

[Asa2002] Nattakan Pengphon, Asanee Kawtrakul, Mukda Suktarachan, : Word Formation Approach to Noun Phrase Analysis for Thai

[Asa2003] S.P et Kawtrakul Asanee : Thai Word Segmentation based on Global and Local Unsupervised Learning.

[Kos2003] Krit Kosowat : Méthodes de segmentation et d'analyse automatique de textes thaï, thèse de doctorat Université Marne-La-Vallée.

[Ber2004] Vincent Berment : Méthodes Pour Informatiser Des Langues Et Des Groupes De Langues « Peu Dotées », thèse de doctorat Université Joseph Fourier.

Dictionnaires : HAAS, Stanford 1964, Thai-English Students dictionnary. SE-ED'S, Bangkok 2001, Modern Thai-Enlish dictionnary. , พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๔๒ (Dictionnaire en ligne de l'institut royal 2542 : http://rirs3.royin.go.th/dictionary.asp)