Développement d'une maquette de traduction automatique khmer-français comme modèle pour des langues peu dotées

Guillaume de Malézieux

Mémoire de MASTER

Ingénierie Multilingue

Directeurs de recherche :

Vincent BERMENT

Frédérique SEGOND

Présenté et soutenu le 21 novembre 2014

Remerciements

Je remercie mes directeurs de recherche, Vincent Berment et Frédérique Segond qui m'ont guidé dans l'élaboration de ce mémoire. Une attention particulière à Vincent Berment, qui était aussi mon maître de stage, pour avoir rendu possible ce projet et pour sa disponibilité sans faille.

Je remercie Michel Antelme de m'avoir aidé plusieurs fois sur la partie khmère.

Merci également aux chercheurs du GÉTALP, Sylviane Chappuy, Jean-Philippe Guilbaud et Christian Boitet de m'avoir rencontré, conseillé et aidé.

Un remerciement envers mes camarades de master, et particulièrement à l'attention d'Amélie Bosc avec qui j'ai effectué mon stage.

Enfin, je remercie ma famille et mon chat pour le soutien qu'ils m'ont apporté.

Table des matières

Index des figures

Index des tableaux

1 Introduction

1.1 Problématique

Ariane est un générateur de systèmes de traduction automatique conçu et développé au GÉTA dans les années 70 et 80. Les systèmes y sont programmés dans des langages de haut niveau adaptés à la programmation linguistique. En concevant Ariane, le GÉTA a créé les langages de programmation qui permettaient d'exprimer le plus intuitivement possible un processus linguistique bien défini. Cette méthodologie linguistique, qui s'est affinée au fil des projets du GÉTA, est la connaissance « métier » fondamentale qui permet au développeur de programmer un système de TA de qualité.

Le projet consistera à développer une maquette de traduction automatique khmer-français pour un chapitre du Petit Prince, conte écrit en 1943 par Antoine de Saint-Exupéry, et qui est, selon Wikipédia, l'ouvrage littéraire le plus traduit au monde après la bible (Le Petit Prince a été traduit en 270 langues, y compris en UNL qui est une représentation abstraite sous forme de graphe sémantique).

Les différentes phases du système seront réalisées séquentiellement :

1. Définition des variables pour l'espace linguistique khmer
2. Segmentation du khmer, en transformant les espaces en séparateurs qui seront interprétés comme des virgules
3. Réalisation de l'analyseur morphologique

- Réalisation du dictionnaire correspondant aux entrées du chapitre à traduire
- Étude de l'analyse des affixes des formes stables

4. Spécification syntaxique de type « grammaire statique »
5. Réalisation de l'analyseur structural

- Désambiguïsation
- Analyse avec production d'une m-structure (structure contenant trois niveaux linguistiques : les syntagmes, les fonctions syntaxiques reliant les syntagmes et une représentation logico-sémantique basée sur le modèle prédicat-arguments)

6. Réalisation d'un transfert khmer-français (un transfert khmer-UNL sera aussi envisagé) et branchement sur l'une des générations du français existantes (celle de l'anglais-français ou celle du russe-français)

Outre son intérêt intrinsèque pour la traduction automatique du khmer, le projet montrera aussi l'aptitude de la méthodologie linguistique du GÉTA à être utilisée à la fois en dehors du laboratoire et pour une langue peu dotée informatiquement (le khmer). Cet aspect situe donc d'emblée le projet dans la problématique de l'informatisation des langues peu dotées en témoignant de la possibilité pour des groupes désireux de développer leur propre système de traduction automatique d'utiliser ces outils. De ce fait, le projet utilisera toutes les méthodes applicables aux langues peu dotées pour optimiser les temps de développement et en particulier :

- la réutilisation de code existant (ex. : système FR3-AN3)
- le développement en synergie avec d'autres développements
- l'utilisation d'un site collaboratif dédié aux développeurs Ariane (lingwarium.org)

Les gains de temps obtenus feront l'objet d'une réflexion sur la méthode. Les astuces découvertes pendant le projet seront consignées sur le site collaboratif lingwarium.org et dans un guide pédagogique qui pourra être enrichi au-delà du projet.

1.2 Collaboration avec le GETALP

Les premiers travaux sur la traduction automatique en France ont commencés en 1959 (Léon, 2002) au sein du CETA (CNRS) à Grenoble sous la direction du Professeur Vauquois. Les recherches effectuées sur un système russe-français (400 000 mots) (Boitet, 1988) seront malheureusement abandonnées suite à un changement d'ordinateur (avec changement d'architecture de processeur rendant les programmes incompatibles).

Ce laboratoire devient le GETA en 1971, et produira un système de TA repensé nommé Ariane-78 (Ariane en référence au fil car « il s'agissait de souligner que l'informatique, si elle est essentielle, doit être au service de non-informaticiens et leur permettre de travailler de façon autonome, grâce à des langages spécialisés adéquats (langages symboliques de règles), et à une interface interactive transparente. » et 78 pour l'année de mise en service).

Un des axes phares de l'équipe du GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) est le traitement des langues peu dotées. Il était alors tout naturel de créer un partenariat entre l'INALCO et le GETALP.

2 État de l'art

2.1 La notion de langue peu dotée

Afin de mieux comprendre ce que l'on entend par langue peu dotée (informatiquement), il sera d'abord question de vous présenter l'indice de satisfaction servant à déterminer si une langue est peu dotée, moyennement dotée ou très bien dotée.

Ensuite serons présentés des exemples de difficultés liées au traitement informatique de la langue khmère

2.1.1 Indice de satisfaction (indice-σ)

Qu'est-ce qu'une langue peu dotée informatiquement ? Il n'existe pas de mesure scientifique du degré d'informatisation d'une langue. Aussi, Vincent Berment a proposé dans sa thèse (Berment, 2004) une méthode d'évaluation de l'indice de satisfaction (**indice-σ**) des outils informatiques disponibles pour une langue. Pour ce faire, il est demandé à un groupe d'utilisateurs locuteurs de la langue d'évaluer des services et ressources selon un niveau de criticité C_k et une Note N_k. La moyenne pondérée de ces notes, l'indice-σ, permet de classer les langues en trois catégories :

- **langues-π** : moyenne entre 0 et 9,99 (peu dotées)
- **langues-μ** : moyenne entre 10 et 13,99 (moyennement dotées)
- **langues-τ** : moyenne entre 14 et 20 (très bien dotées)

Une langue-π est ainsi une langue dont l'informatisation est jugée insatisfaisante par ses utilisateurs. Si la subjectivité de cette notation ne permet pas de déterminer une frontière précise entre les différents niveaux d'informatisation, elle permet néanmoins de mettre en évidence l'absence de certains services ou de certaines ressources pour les langues-π, alors qu'ils/elles sont considéré(e)s comme acquis(e)s pour les langues-τ.

2.1.2 Exemple de difficultés pour le khmer

La principale source de difficultés pour le traitement automatique des langues peu dotées, et en particulier le khmer, est liée à l'encodage des caractères et à leur représentation par des polices d'écriture. Si la prise en charge du khmer par Unicode a permis de régler les problèmes inhérents au Legacy Font (Suzuki, 2006), il reste des problèmes non-résolus.

Parmi ces problèmes, celui du multilinguisme des polices de caractères. En effet, si Unicode permet la couverture de 100 écritures dans sa dernière version (7.0), il n'existe aucune police d'écriture capable d'en représenter ne serait-ce que la moitié. Ces polices couvriront généralement la table ASCII avec un design basique, et la table de caractères de la langue

pour laquelle elles ont été conçues. Ainsi lorsqu'un document multilingue est conçu, deux choix s'offrent à l'auteur :

- Utiliser une seule et même police d'écriture si l'on souhaite écrire de l'anglais et une autre langue. Cela ne présente pas de difficulté particulière si la deuxième langue est à écriture latine, sinon auquel cas il n'y a généralement que peu de cohérence entre les deux.
- Utiliser une police différente pour chaque langue du document, laissant place à une cacophonie visuelle.

C'est pour cela qu'Alan Wood a consacré une partie de son site[1] à la présentation de polices d'écriture en fonction de leur couverture linguistique.

D'autres, tentent une harmonisation entre l'anglais et la langue de travail. C'est le cas de la police « Khmer SBBIC Serif Font » publiée sur le site sbbic.org. Le design des caractères pour l'anglais et le khmer n'ont pas été conçus conjointement, ce qui aurait été l'idéal, mais sont en fait extraits de polices existantes (Droid Serif pour l'anglais, KhmerOS Battambang pour le khmer). Le résultat est correct :

Le Petit Prince : ព្រះអង្គម្ចាស់តូច Time + KhmerOS Content

Le Petit Prince : ព្រះអង្គម្ចាស់តូច Khmer SBBIC Serif Font

Si cette police permet une meilleure cohabitation de l'anglais et du khmer dans un même document, le problème persiste si l'on utilise une autre écriture, qu'elle soit dite complexe, ou plus simple comme le français. Le « e dans l'o » est considéré comme un problème résolu, mais cette police ne le représente pas correctement (œ : ◌ٮ).

Récemment Google a lui aussi tenté de répondre à ce besoin en publiant la famille de polices « noto »[2]. Le projet contient actuellement 98 polices et vise à couvrir l'ensemble des écritures disponible dans Unicode tout en les harmonisant. Toutes ces polices ont donc un style commun qui est celui employé au sein du système d'exploitation Android.

Autre problème majeur, celui de l'interlignage. Lorsque l'on souhaite publier un ouvrage de qualité il est conseillé d'avoir un interlignage régulier et cohérent. Vous constaterez beaucoup

[1] http://www.alanwood.net/unicode/fonts.html
[2] http://www.google.com/get/noto/#/

d'irrégularités et d'incohérences tout au long de ce document, cela est malheureusement dû à la présence de plusieurs systèmes d'écritures au sein d'un même paragraphe.

2.2 Traitement automatique du khmer

Les travaux d'informatisation de la langue khmère ont débutés au début des années 1980, au moment où le pays pansait ses plaies. Les premières initiatives ont souvent été individuelles, et proposées par des membres de la diaspora cambodgienne (Khin et Loeurng 1983).

2.2.1 Institutions

Deux organismes sont actuellement en charge du TAL au Cambodge :

- The National Information Communications Technology Development Authority (NiDA)[3] à qui l'on doit la standardisation du clavier khmer.
- L'Institut de Technologie du Cambodge (ITC)[4] dont les axes de recherche du département TAL sont la reconnaissance optique de caractères et la reconnaissance vocale.

2.2.2 Organisations et individus

Plusieurs organisations et individus contribuent activement à l'informatisation du khmer :

- Khmer Software Initiative (Khmer OS)[5] est un projet de l'Open Institute[6] ayant pour objectif de produire des programmes informatiques en khmer, et d'internationaliser (i18n) et de traduire (l10n) des logiciels open source. Son fondateur, Javier Sola[7] a contribué à nombres de ces projets.
- PAN Localization est un projet de collaboration entre les pays d'Asie du Sud et d'Asie du Sud-Est initié par le Centre de Recherches pour le Développement International (Canada).
- Society for Better Books in Cambodia (SBBIC)[8] est une association fondée par Nathan Wells dont la vocation est de reprendre la tradition des moines copistes en la modernisant afin de proposer une version informatisée de la bible en khmer. Ils proposent des articles et des outils visant à améliorer le traitement informatique du khmer.
- Le pôle traitement informatique[9] au sein du Groupe de Travail pour le Développement

[3] http://www.nida.gov.kh/
[4] http://itc.edu.kh/
[5] http://www.khmeros.info/en
[6] http://open.org.kh/en
[7] http://javiersola.users.sourceforge.net/
[8] http://www.sbbic.org/
[9] http://camdevel.free.fr/info/infoprincipal.html

du Cambodge a produit un correcteur d'orthographe ainsi qu'une version informatisée du Dictionnaire Khmer-Français d'Alain Daniel. Cette version rend la consultation plus pratique, mais elle le maintient aussi en vie puisqu'il est aujourd'hui épuisé.

- Danh Hong, a été un des précurseurs de la création de polices d'écriture khmères et n'a jamais arrêté. Ses polices de la famille Khmer OS sont les plus répandues, et les polices utilisées par Google dans le cadre des projets Google Fonts et Google noto en sont des sœurs jumelles. Ses contributions ne s'arrêtent pas là, il publie sur son blog[10] de nombreux articles sur l'informatisation du khmer et crée un clavier khmer sous Android[11] (Le khmer n'est disponible sous Android que depuis un an et seulement pour les derniers modèles. L'iPhone, lui, ne le prend toujours pas en charge.).
- Le blog d'Olivier Berten[12] est entièrement consacré aux polices d'écriture khmères. Anciennes ou récentes, rares ou répandues, leur nombre est considérablement élevé. C'est une mine d'or pour qui a besoin d'une police précise, ou de trouver la police adéquate.
- Le site http://kheng.info/ est apparu en 2014 et semble basé en Californie[13]. Il propose plusieurs outils dont un segmenteur et son but premier est la constitution d'un dictionnaire audio en ligne.

[10] http://www.khmertype.org/
[11] http://www.khmerkeyboard.com/
[12] http://www.selapa.net/khmerfonts/
[13] https://who.is/whois/https://kheng.info

2.3 Traduction Automatique

2.3.1 Les systèmes existants

Le nombre de systèmes de traduction automatique (TA) est très grand. Dans la 16e édition (Hutchins, 2010) de son « Compendium of Translation Software », John Hutchins liste plus de 600 produits liés à la traduction automatique. Le nombre de produits et de couples de langues disponibles a crû considérablement depuis une dizaine d'années, en particulier du fait de l'explosion de la TA statistique. La progression la plus stupéfiante est celle de Google qui offre actuellement des traductions entre 80 langues (voir annexes 1 et 2). Cette progression est illustrée par le schéma ci-dessous (voir http://en.wikipedia.org/wiki/Google_Translate , et le détail de la chronologie en annexe 3).

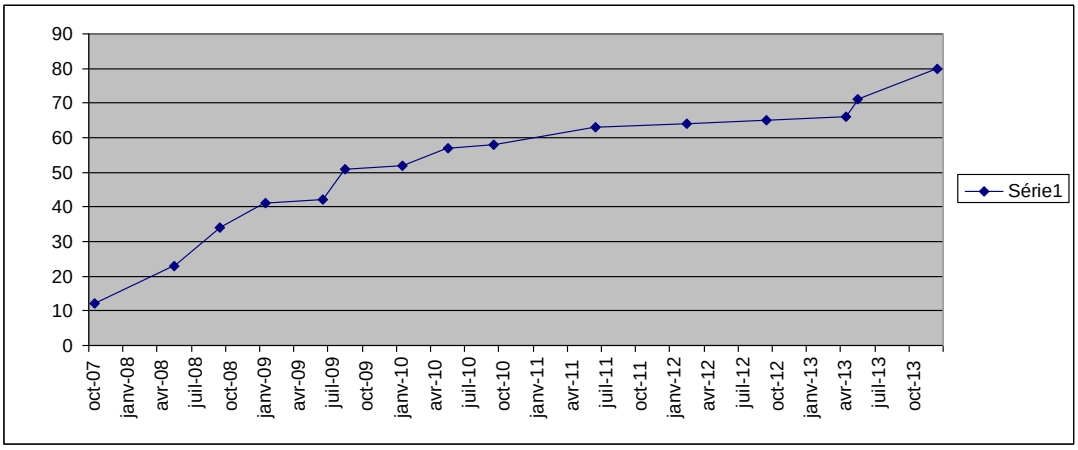


Figure 1 Evolution du nombre de langues couvertes par Google Translate

On voit là toute l'efficacité de l'approche statistique par rapport à l'approche à règles, l'un des leaders de cette technologie, culmine à moins de 30 couples de langues, si l'on ne compte qu'une fois un couple (http://www.mysoft.fr/produit/systran_traduction_automatique.htm), sans tenir compte de la direction pour pouvoir comparer avec Google. La solution Bing de Microsoft (http://www.bing.com/translator/), elle aussi basée sur une approche statistique, offre des traductions entre 34 langues.

Compte tenu de la nécessité de disposer de corpus alignés de grande ampleur, les langues proposées par les systèmes de traduction statistiques sont celles pour lesquelles de tels corpus sont disponibles. Au chapitre 3.1 de son article « Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes » (Boitet, 2008), Christian Boitet compare la taille et le coût des ressources nécessaires pour développer un nouveau système en fonction des architectures computationnelles employées. Le résultat est le suivant.

Type de phrase Architecture	**6.5 mots/phrase** **BTEC, METEO**	**25 mots/phrase** **Informations (news)**
SMT **PSMT** **EBMT** par analogie **Coût :**	0.9 - 3 M mots 3.6 - 12 K pages 0.15 - 0.5 M phrases **2.4—8 h*a**	50 - 200 M mots 200 - 800 K pages 2 - 8 M phrases **100 - 400 h*a** (rarement disponible !)
EBMT avec arbres **SMT** Mastor-1 (IBM) **Coût :**	N/A pour ce type de phrases courtes Apprentissage supervisé 1h/page (par recoupements)	4—12.5 M mots 15—50 K pages 0.15—0.5 M phrases **10 - 40 h*a**
EBMT avec arbres et S-SSTC Banturjah (USM) **Coût :**	N/A pour phrases courtes Apprentissage supervisé 15 h/page (10 h/p espéré) dictionnaire (50 K) souvent disponible	4—12.5 M mots 0.6—1 K pages 0.006—0.01 M phrases **6 - 10 h*a** (travail assez spécialisé)
RBMT **Coût :**	Dictionnaire 3-10 K 0.6—2 h*a **Total 1 - 3 h*a**	Dict. 50-500 K, soit 15—150 h*a Grammaires environ 25 h*a **Total ≈ 40 - 175 h*a**

Figure 2 D'après (Boitet, 2008)

Cette analyse montre qu'en l'absence de corpus aligné, la TA statistique (SMT) revient beaucoup plus chère que la TA à règles (RBMT). Et Christian Boitet de conclure « Créer de très gros corpus parallèles à partir de zéro est 2 à 3 fois plus coûteux que de construire un grand système de TA par approche experte (procédurale et/ou à automates et grammaires). ».

Cela explique que les systèmes réalisés par les grandes entreprises présentes dans le domaine de la traduction automatique soient destinées à des langues déjà relativement bien dotées, même en ce qui concerne la TA statistique. Ainsi, parmi les 80 langues de Google Translate, plus de 70 sont parlées par plus d'un million de personnes, seulement 6 langues vivantes (Google propose le latin et l'esperanto) sont parlées par moins d'un million de personnes : basque, gallois, irlandais, islandais, maltais et maori. Remarquons aussi qu'environ 80% des langues de plus d'un million de locuteurs (il y a environ 320 langues parlées par plus d'un million de locuteurs) ne sont pas dans Google Translate, ce qui montre que le nombre de locuteurs ne suffit pas à l'existence de gros corpus alignés.

Le Cambodge faisant partie d'un ensemble géographique et politique riche en langues, il est intéressant d'étudier la disponibilité de systèmes de TA pour les langues de l'ASEAN[14]. Le

[14] Notons que l'ASEAN arrive au terme d'un processus d'intégration économique de type CEE. Le lancement de l'AEC (*ASEAN Economic Community*) est prévu en 2015 (http://www.asean.org/communities/asean-economic-community).

tableau de la page suivante montre qu'au niveau des langues officielles, des systèmes existent. Sur 9 langues officielles, 7 sont dans Google (filipino, indonésien, khmer, lao, malais, thaï et vietnamien), 4 dans Bing (indonésien, malais, thaï et vietnamien). Il n'en manque donc que deux, le birman (32 millions de locuteurs) et le tetum (450000 locuteurs). Mais la qualité n'est pas toujours au rendez-vous : les indices-σ recueillis pour le lao et pour le khmer lors de la préparation de l'atelier WSSANLP 2014 qui a eu lieu à Dublin étaient respectivement de 4/20 et de 6/20. Remarquons enfin que si 7 langues officielles sur 9 disposent au moins d'un début de service de traduction automatique, il y a au total environ 1500 langues parlées dans la zone (un quart des langues du monde !), dont 27 sont parlées par plus d'un million de personnes.

Notons en particulier l'existence du projet collaboratif ASEAN-MT[15] lancé à l'initiative de NECTEC, organisme thaïlandais très actif en traitement des langues.

[15] http://www.aseanmt.org/

		Multi-languages systems							Language-specific systems		
Language	Country	Google	Microsoft	ASEAN MT	Asia Online	Babylon	Taranis	Apertium	SiSTeC	UCSY	Khmer
Burmese	Myanmar			D			D			D	
Chinese	Singapore	X	X	D	X	X			D		
English	Philippines, Singapore	X	X	D	X	X	X	X	D	D	D
Filipino	Philippines	X (Tagalog)		D							
Indonesian	Indonesia	X	X	D	X			X			
Khmer	Cambodia	X		D			D				D
Lao	Laos	X		D			D				
Malay	Brunei, Singapore			?							
Malaysian	Malaysia	X	X	?	X			X	D		
Portuguese	East Timor	X	X		X	X		X			
Tamil	Singapore	X			D						
Tetum	East Timor										
Thai	Thailand	X	X	D	X	X	D	D			
Vietnamese	Vietnam	X	X	D	D						

Figure 3 Disponibilité de systèmes de TA pour les langues officielles de l'ASEAN

Légende :

X : Disponible

D : En cours de développement ou prototype

2.3.2 La TA pour les langues peu dotées

Peut-on espérer voir un jour les robots qui envahissent aujourd'hui nos vies être capables de communiquer dans n'importe quelle langue du monde, à l'instar du droïde protocolaire C-3PO de *Star Wars* qui affirme « maîtriser plus de six millions de formes de communication » ? Rien n'est moins sûr si l'on considère que les systèmes existants réellement utilisables sont essentiellement les produits d'entreprises qui doivent rentabiliser leurs investissements. Comme le coût des systèmes croît avec la rareté des ressources disponibles (corpus parallèles, dictionnaires...) et que les revenus décroissent, eux, avec le nombre d'utilisateurs, un double effet économique fait que la progression du nombre des langues disposant de TA va connaître une asymptote que Vincent Berment place arbitrairement autour d'un million de locuteurs (Berment, 2014), la qualité des moins rentables pouvant par ailleurs être bien moindre que celle des langues centrales comme les principales langues européennes.

Alors, à quel niveau va se situer cette asymptote, combien de langues peut-on espérer voir dotées de TA par le secteur privé ? Avec la barrière proposée par Vincent Berment à un million de locuteurs, l'asymptote pourrait se situer au mieux autour de 350 langues, 324 langues dépassant ce nombre selon le site Ethnologue[16] et au moins au niveau des langues officielles qui sont au nombre de 172[17] . Nous pouvons donc nous attendre à ce que le secteur privé offre dans un avenir plus ou moins proche des systèmes de TA de plus ou moins bonne qualité pour 200 et 300 langues, soit 3% à 5% des langues du monde.

Si l'on considère la cinétique actuelle des systèmes produits par Google, qui est en moyenne de 10 nouvelles langues par an (voir 2.3.1), il faudra environ 30 ans pour arriver aux 324 langues parlées par plus d'un million de locuteurs. Et compte tenu de l'effet de freinage évoqué précédemment, on peut s'attendre à ce que le rythme annuel décroisse et qu'il faille plus de 30 ans pour y parvenir, malgré les moyens gigantesques de cette entreprise. Alors, que penser du titre d'une interview de Christian Boitet par Pierre Vandeginste paru dans les Dossiers de la Recherche (Boitet, 2013) : « *Les logiciels traduiront 600 langues dans dix ans* » ? Christian Boitet explique : « *Il y aura un effet boule de neige. De plus en plus de ressources linguistiques apparaissent sur Internet, à commencer par les bases de données lexicales comme les WordNet, super-dictionnaires enrichis d'informations de nature sémantique, sous forme de liens, par exemple entre synonymes. ... Il est aussi probable que la*

[16] http://www.ethnologue.com/
[17] http://fr.wikipedia.org/wiki/Liste_des_langues_officielles

traduction automatique de beaucoup de couples de langues peu dotées passera par une représentation sémantique abstraite de l'anglais universalisé, comme le propose notamment UNL. ». Et encore : « *Je m'attends à un retour des solutions dites expertes, qui reposent sur la mise en œuvre de connaissances linguistiques. Les limites des méthodes statistiques sont mieux perçues aujourd'hui. En particulier, on s'aperçoit qu'elles ne réduisent pas forcément les coûts, notamment dans le cas de couples de langues pauvrement dotées (comme le français-thaï), c'est à dire pour lesquels on ne trouve pas d'emblée de gros corpus de textes parallèles de qualité adéquate.* ». L'interview de Vincent Berment dans la même revue va d'ailleurs dans le même sens d'une approche experte utilisant des méthodes adaptées pour les langues et les couples de langues peu dotés.

Il apparaît donc qu'un retournement de la tendance actuelle de la TA dite statistique vers la TA experte, couplée avec des méthodes appropriées comme la réutilisation des ressources de type WordNet, soit à prévoir au moins pour les langues et les couples de langues peu dotés. Il est remarquable à ce titre que des initiatives ouvertes comme NooJ[18] ou Apertium[19] , qui sont des environnements de développement experts (NooJ n'est pas spécialisé sur la TA mais permet de réaliser la plupart des traitements qu'on y trouve) voient de plus en plus d'utilisateurs développer des modules pour des langues peu dotées. Par exemple, Maximiliano Duran, doctorant de INaLCO, a réalisé un analyseur morphologique du quechua d'Ayacucho, sa langue maternelle et langue classée comme en danger par l'UNESCO, à l'aide de l'environnement NooJ de son co-directeur de thèse, Max Silberztein. Ou encore, Vee Satayamas, doctorant à l'université Kasetsart de Bangkok, s'est essayé à l'écriture de modules pour le thaï avec l'environnement Apertium.

Nous constatons qu'il y a une demande pour des outils ouverts et facilement appropriables permettant à des personnes ou à des groupes de personnes de développer eux-mêmes, avec l'aide d'une communauté de bénévoles, des systèmes de TA. C'est cette approche que nous avons défendue dans l'article « RBMT as an alternative to SMT for under-resourced languages » (de Malézieux *et al.* , 2014) que nous avons présenté avec Amélie Bosc et Vincent Berment à l'atelier WSSANLP[20] en août 2014. Notre article montrait en particulier comment utiliser la méthodologie linguistique du GETA et l'environnement de développement Héloïse pour réaliser assez facilement des systèmes de traduction automatique de qualité en marge des systèmes commerciaux du secteur privé.

[18] http://www.nooj4nlp.net
[19] http://www.apertium.org
[20] http://www.sanlp.org/wssanlp2014/

2.4 Méthodologie linguistique du GETA

2.4.1 Principes linguistique

Suivant les principes de la pyramide de Vauquois, le modèle linguistique défini au GETA utilise le concept de « structure multiniveaux » afin de décrire un énoncé avec le plus haut niveau d'abstraction possible. Au dessus de la description lexicale se trouvent :

- Le niveau de surface, composé de deux niveaux d'interprétation :
    - Le parenthésage en termes de classes syntagmatiques,
    - Les fonctions syntaxiques.
- Le niveau profond, composé de deux autres niveaux d'interprétation :
    - Les relations logiques,
    - Les relations sémantiques.

Cette structure profonde est indépendante de la syntaxe (et dans une certaine mesure de la langue) et se veut équivalente pour tout paraphrasage d'un même énoncé.

2.4.2 Phases de traduction

La particularité des étapes d'analyse et de génération des systèmes de TA suivant la méthodologie linguistique du GETA est leur indépendance :

- Analyse : Cette phase est monolingue. Son rôle est de créer la structure multi-niveaux évoquée précédemment.
- Génération : La phase de génération consiste à construire un énoncé grammaticalement correct dans la langue cible, à partir de la structure multi-niveaux traduite lors de la phase précédente.

Quant au transfert, grâce aux niveaux profonds de la structure multiniveaux, il est minimal. C'est la seule phase bilingue, qui consiste en un passage d'un espace lexical à un autre (traduction d'une unité lexicale et de ses variables). Un passage via pivot UNL est possible afin de traduire de/vers une multitude de langues.

La structure de données reste la même tout au long du traitement, à savoir un arbre dont les nœuds portent des décorations indiquant leurs propriétés et leurs relations par rapport aux autres.

Afin que le linguiste puisse travailler avec des concepts familiers (dictionnaires, grammaires) les programmes et les données linguistiques sont séparés.

L'indépendance des phases d'analyse et de génération du processus de traduction permet une approche multilingue. Ainsi les phases d'analyse d'une langue sont réutilisables avec une génération de n'importe quelle autre langue.

3 Développement d'un système khmer-français

3.1 La langue khmère

3.1.1 Origines

Le khmer est une langue du groupe môn-khmer appartenant à la famille des langues austro-asiatiques. Avec 15 millions de locuteurs, elle est la deuxième langue la plus parlée de cette famille après le vietnamien (75 millions de locuteurs dans le monde).

Langue officielle du Cambodge, elle est aussi parlée au Laos, en Thaïlande, au Vietnam, ainsi que dans les diasporas issues de réfugiés fuyant le régime des Khmers Rouges (1975-1979).

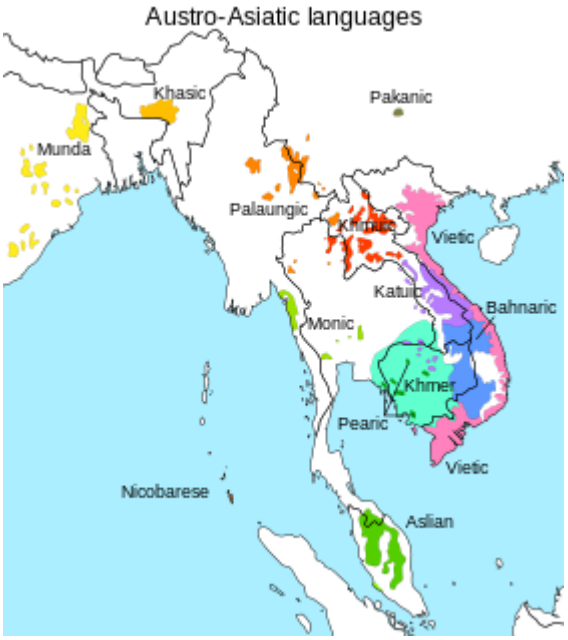


Figure 4 Langues Austro-asiatiques

3.1.2 Écriture

L'écriture khmère est un alpha-syllabaire dérivé du Pallava (Sud de l'Inde). C'est par ailleurs un système d'écriture non segmenté, ce qui veut dire qu'il n'y a pas d'espace entre les mots. Cependant, il existe des espaces entre certains mots ou segments de phrases mais les règles de séparation n'ont pas été clairement définies. Cette absence de séparateur entre les mots est une difficulté majeure de plusieurs langues d'Asie du Sud-Est, dans la mesure où l'indispensable segmentation en mots est une tâche complexe et qui reste imparfaite pour toutes ces langues, du fait de la présence d'ambiguïté de segmentation.

3.1.3 Orthographe

Evénements clés expliquant les problèmes orthographiques actuels :

- Unification de l'orthographe avec le premier dictionnaire monolingue 1938 par Samdech Chuon Nath, chef de l'ordre Mahānikāy 1967 5ème édition (élève de Louis Finot et en contact avec Georges Coedès).

  Tendance : orthographe étymologique

- Années 1960 mise en place de la « khmérisation » de l'enseignement. (post-indépendance) Keng Vannsak.
- Sous les Khmers Rouges l'orthographe n'est pas importante, seul le vocabulaire communiste modifiera quelques peu la langue.
- République populaire du Kampuchéa : reconstruction d'un système d'enseignement avec deux orthographes (aucune n'ayant été officialisée) et par des personnes peu qualifiées.
- De 1979 à 1999, l'orthographe de la khmérisation sera la plus utilisée.
- 2000 : création de l'Académie royale du Cambodge (projet initial de Sihanouk post 1970) retour officiel de l'orthographe du dictionnaire Chuon Nath.

  Le Ministère de l'Éducation mettra dix ans à appliquer ce changement.

- La traduction du Petit Prince ayant été réalisé entre 2000 et 2002, l'orthographe correspond donc à celle de la « khmérisation » qui n'est plus utilisée maintenant.

3.1.4 Morphosyntaxe

La structure des mots est dissyllabique à tendance monosyllabique. Les quelques cas de polysyllabes viennent d'emprunts, principalement au sanskrit et au pāli.

La typologie morphologique du khmer est isolante. Il existe pourtant un procédé d'affixation (préfixes et infixes) qui permet de modifier la fonction syntaxique d'un mot.

3.1.5 Ressources linguistiques

3.1.5.1 Dictionnaires

Les ressources linguistiques sur la langue khmère sont peu nombreuses, le Cambodge n'ayant pas de tradition d'étude de la langue.

Le dictionnaire qui fait référence en khmer est le *Chuon Nath*, du nom de son auteur, un bonze ayant voué sa vie à la sauvegarde de la langue khmère. La dernière édition de ce dictionnaire par l'institut bouddhique date de 2005 (41000 entrées). Il est maintenant disponible informatiquement sous forme d'application sur ordinateurs et smartphones, ou en ligne sur certains sites qui proposent sa consultation.

Un des deux plus importants dictionnaires khmer-français est issu de la traduction du *Chuon Nath* par le père Rondineau, aidé de Mme Thérésa Long Lean et de Mme Thérésa Salay Sangkhum. Ce travail a pris onze années pour produire un ouvrage de 4000 pages en deux volumes. Il a été édité par les missions étrangères de Paris en 2007.

Un précédent ouvrage est à signaler, le *Dictionnaire Cambodgien-Français* de monsieur Alain Daniel (linguiste, ancien responsable de la section khmère de l'INALCO), publié en 1985 par l'Asiathèque. Il est malheureusement épuisé, mais comme pour les autres dictionnaires, des copies circulent autant en version papier que sous forme d'applications pour smartphone. Il est à signaler que M. Daniel est pourtant disposé à diffuser le contenu de son dictionnaire avec son accord. C'est grâce à cette générosité que monsieur Albert Sy a créer une version informatisée de ce dictionnaire sous la forme d'une application téléchargeable gratuitement. Le mérite de monsieur Sy est double, car pour garantir à un maximum d'utilisateurs d'y avoir accès, il a pris soin d'en rendre possible la consultation avec l'encodage Unicode, mais aussi avec l'ancien système des « *legacy font* ».

Dans le sens français-khmer le dictionnaire le plus complet est celui de monsieur Michel Antelme et de madame Hélène Suppya Bru-Nut, L'Asiathèque 2013.

Afin de compléter ces ressources, il est utile de pouvoir consulter les travaux anglo-saxons. Deux dictionnaires Cambogien-Anglais sont remarquables. Celui de Judith M. Jacob édité en 1974 par l'université d'Oxford, ainsi que celui de Robert K. Headley édité une première fois en 1977 puis une seconde fois en 1997 dans une édition augmentée.

Il est possible de consulter le Headley ainsi que le Chuon Nath en ligne sur le site du projet sealang.

3.1.5.2 Grammaires

Les grammaires de référence sont les suivantes :

- Grammaire de la langue khmère (cambodgien). Georges Maspero, 1915
- Introduction to Cambodian. Judith Jacob, 1968
- La grammaire du khmer moderne. Khin Sok 1999

3.2 Méthodologie

Nous avons vu que la problématique du traitement des langues peu dotées était avant tout une question de moyens, de disponibilité des données, de coût, de nombre de participants et de

temps de travail. Il est alors essentiel de factoriser les efforts partout où cela est possible.

3.2.1 Corpus

La première étape de notre projet de système de traduction automatique est de constituer le corpus. Nous avons choisi *Le Petit Prince* pour sa disponibilité dans un grand nombre de langues (270[21]) et son vocabulaire générique. Le livre est relativement court (100 pages), mais nous avons décidé de nous limiter au premier chapitre dans un premier temps, afin qu'il puisse servir de modèle pour les trente-six suivants. Ne disposant pas de version informatique du texte en khmer, nous avons commencé par le saisir dans un fichier Word. Ce format a l'avantage de nous donner la possibilité de sauvegarder la police de caractère utilisée ainsi que sa taille et l'espacement entre les lignes. Tout cela apporte un confort visuel non-négligeable lors de la saisie. En cas de besoin, il nous sera tout à fait possible d'extraire le contenu du fichier au format texte.

3.2.2 Branchement sur la génération d'un système existant

Les phases d'analyse et de génération étant indépendantes, nous avons développé la phase d'analyse en khmer pour la « brancher » sur une génération du français existante via une phase de transfert, seule phase bilingue.

3.3 Analyse Morphologique

3.3.1 Segmentation

Le khmer étant une langue écrite avec un système d'écriture non segmenté (voir 3.1.2), une étape de segmentation est nécessaire avant l'analyse morphologique. Pour ce faire, nous disposons de l'outil en ligne « Motor » qui est intégré à Héloïse (préprocesseur). C'est un programme écrit en C++ par Vincent Berment et qui utilise un algorithme de plus petit nombre de mots avec retour arrière (*maximum matching algorithm*) pour segmenter en mots un texte sans espace écrit dans un système d'écriture non segmenté.

On passe en paramètre de Motor une chaîne de caractères et il va calculer la meilleure segmentation, au sens du plus petit nombre de mots pour la chaîne. Dans le cas où plusieurs segmentations donnent le même nombre de mots, la première solution trouvée est fournie. Pour chaque possibilité de mot (chaîne contenant un nombre de caractères compris entre deux bornes), Motor va tester sa présence dans un dictionnaire sous forme d'une base de données que nous lui avons fournie.

Cette base de données est une simple liste de mots, dont la qualité est primordiale. Comme

[21] http://fr.wikipedia.org/wiki/Le_Petit_Prince

pour les autres ressources, les listes de mots khmers sont rares. La plus complète que nous ayons trouvée est celle intégrée à la bibliothèque ICU[22]. C'est cette liste qui est actuellement utilisée dans la majorité des programmes informatiques prenant en charge le khmer pour permettre les différentes opérations sur un texte, vues dans la partie consacrée à l'indice de satisfaction (chapitre 2.1.1). Après avoir effectué des premiers tests de segmentation, nous avons réalisé que cette liste de mots n'était pas utilisable en l'état, le niveau de granularité n'étant pas suffisant pour une analyse correcte.

La phrase exemple de Motor (អ្នកចេះនិយាយភាសាខ្មែរទេ ?) était par exemple segmentée ainsi :

> អ្នកចេះ/និយាយ/ភាសាខ្មែរ/ទេ/ ?
>
> personne éduquée/parler/khmer/particule interrogative/ ?

La segmentation correcte au niveau du sens serait :

> អ្នក/ចេះនិយាយ/ភាសាខ្មែរ/ទេ/ ?
>
> vous/savoir parler/khmer/particule interrogative/ ?

Nous avons ici une ambiguïté qu'il est impossible de traiter avec la liste ICU. អ្នកចេះ est un mot composé du nom អ្នក et du verbe ចេះ. La présence de ce mot composé dans la liste ICU rend impossible la désambiguïsation avec le cas où អ្នក est sujet du verbe ចេះ.

La segmentation qui nous permettrait d'analyser chaque élément de la phrase est la suivante :

> អ្នក/ចេះ/និយាយ/ភាសា/ខ្មែរ/ទេ/ ?
>
> vous/savoir/parler/langue/khmer/particule interrogative/ ?

Nous avons testé la segmentation du premier chapitre du *Petit Prince* avec cette liste, et constaté plusieurs problèmes :

Le premier, à l'image de l'exemple précédent, est celui de la longueur des mots. Nous l'avons vu, le khmer est une langue dissyllabique à tendance monosyllabique. Or, si l'on s'en tient à la segmentation issue de la liste ICU, il en résulte des mots anormalement longs :

22 International Components for Unicode : projet open source de bibliothèques en C/C++ et Java pour la prise en charge des textes au standard Unicode. ICU est largement utilisé par les plus grandes compagnies et organisations. https://ssl.icu-project.org/trac/ticket/9311

បាក់ទឹកចិត្ត

démoraliser

Cette segmentation coïncide parfaitement avec sa traduction française ou anglaise. Mais ce mot n’est présent dans aucun dictionnaire monolingue, car il s'agit d'une expression composée d'un verbe បាក់ (casser) suivi des mots ទឹក (liquide) et ចិត្ត (sentiment) formant le mot composé ទឹកចិត្ត (moral). Cette segmentation ne respecte pas la typologie de langue isolante du khmer, et semble tirée d'un dictionnaire bilingue.

Il peut être intéressant de noter qu'une telle segmentation pourrait faire ressembler le khmer à une langue agglutinante. En effet, le procédé de ces langues consistant à « coller » des morphèmes invariables les uns aux autres pour former un mot, ressemble fortement aux procédés des langues isolantes consistant à juxtaposer des mots les uns aux autres pour former une idée.

Cette liste de plus de 85000 mots est donc à corriger. Cela représentait un travail titanesque qui n'était ni possible de réaliser dans la durée impartie, ni nécessaire. Nous avons simplement produit le lexique du texte en le segmentant manuellement afin que Motor puisse l'utiliser en prétraitement.

Pour ce faire, on ouvre un tableur Excel et l'on place tous les mots dans la première colonne, le premier sur la première ligne, le deuxième sur la seconde et ainsi de suite, en veillant à ne mettre qu’une fois chaque mot (ceci peut être fait simplement, de manière plus sûre et a posteriori par filtrage dans Excel). Voici les différentes étapes décrites pour la première phrase :

Phrase non-segmentée :

កាលពីខ្ញុំអាយុប្រាំមួយឆ្នាំ មានពេលមួយ ខ្ញុំបានឃើញរូបភាពដ៏អស្ចារ្យនៅក្នុងសៀវភៅមួយមានចំនងជើងថា «កំរងរឿងពិតៗ» ដែលនិយាយពីព្រៃស្តុក ។

Phrase segmentée :

កាល ពី ខ្ញុំ អាយុ ប្រាំ មួយ ឆ្នាំ , មាន ពេល មួយ , ខ្ញុំ បាន ឃើញ

រូប ភាព ដ៏ អស្ចារ្យ នៅ ក្នុង សៀវភៅ មួយ មាន ចំនង ជើង ថា « កំរង រឿង ពិត ៗ » ដែល និយាយ ពី ព្រៃ ស្តុក ។

Mise en tableau :

Numéro	Mot
1	កាល
2	ពី
3	ខ្ញុំ
4	អាយុ
5	ប្រាំ
6	មួយ
7	ឆ្នាំ
8	，
Etc.	Etc.

Tableau 1 Texte "verticalisé"

Nous remarquons l'apparition de virgules dans la phrase segmentée. C'est le résultat d'un des deux prétraitements que nous avons inclus à Motor.

- Nous avons d'abord supprimé les espaces invisibles (zero-width space ZWSP). Ce caractère est couramment inséré manuellement ou automatiquement pour forcer la reconnaissance de la fin d'un mot et faciliter la mise en page et la recherche de mots. Cette étape est donc indispensable pour garantir la qualité de notre segmentation.
- Ensuite, nous avons procédé à un remplacement tactique des espaces séparant des propositions par des virgules. Cela nous permet de conserver cette information

importante.

Nous sauvegardons une copie de ce texte « verticalisé » dans un fichier Excel qui nous servira plus tard pour l'étiquetage morphosyntaxique.

Pour obtenir le lexique complet il suffit de procéder à un tri alphabétique et de supprimer les doublons. Nous sauvegardons ensuite ce fichier au format CSV avec les modifications suivantes :

1. Insertion d'une colonne vide à gauche
2. Insertion d'une ligne avec dans les trois premières cellules : Clé, Article, Fréquence (cette dernière information est facultative et non renseignée dans le cadre de notre étude).

Ensuite, nous convertissons ce fichier en base de données afin qu'il soit utilisable par Motor. Pour cela nous avons utilisé le module Firefox « SQLite Manager[23] » :

1. Nous créons une nouvelle base de données « SegmentationKhmer.db » (nom de fichier attendu par Motor)
2. Nous importons le fichier CSV dans une table « Dictionnaire » en choisissant la première colonne comme clé numérique que l'on incrémentera automatiquement.

Nous avons vu les différentes étapes techniques pour la réalisation de notre liste de mots destinée à la segmentation par Motor. Revenons maintenant sur nos choix linguistiques.

Prenons le résultat de segmentation avec la liste ICU de la première phrase :

កាលពី ខ្ញុំ អាយុ ប្រាំមួយ ឆ្នាំ , មាន ពេល មួយ , ខ្ញុំ បានឃើញ រូបភាព ដ៏ អស្ចារ្យ នៅក្នុង សៀវភៅ មួយ មាន ចំ នង ជើង ថា « កំរង រឿង ពិត ៗ » ដែល និយាយ ពី ព្រៃស្តុក ។

Examinons maintenant un à un les termes qui posent problème :

1. កាលពី, qui se traduira généralement par « quand » est composé de កាល (substantif « temps ») et de ពី (préposition « de, depuis »).
2. ប្រាំមួយ, correspond au nombre « six » mais est composé de ប្រាំ « cinq » et មួយ « un ». C'est un résidu de système de numération à base cinq, mais le système de numération actuel est bien à base dix tout comme le système de notation des chiffres.

[23] https://code.google.com/p/sqlite-manager/

Nous avons choisi de segmenter afin de respecter l'étymologie des nombres.

3. បានឃើញ, qui pourrait se traduire par « avoir vu » est composé des verbes បាន « obtenir », ayant ici la fonction d'auxiliaire, et ឃើញ « voir ». Si nous avons dans le cas présent un auxiliaire suivi d'un verbe, il est fréquent de rencontrer plusieurs verbes en série exprimant une seule idée, chacun d'eux apportant plus de précisions et de nuances.
4. រូបភាព, « image », est composé de រូប « image » et de ភាព « état, forme ». Il existe plusieurs types de noms composés et cet exemple assez délicat sera traité plus en détails dans le chapitre qui leur sera consacré.
5. នៅក្នុង, « dans, dedans », composé du verbe នៅ « se trouver » et de la préposition ក្នុង « dans ».
6. ចំ នង, devrait être segmenté en un seul mot ចំនង. Cela est dû à son absence de la liste ICU qui ne prend pas en compte l'orthographe de la « khmérisation ». Au vu des nombreux changements orthographiques, il est nécessaire de pouvoir traiter les différents types d'orthographes et d'être capable de les distinguer. Nous expliquerons comment y remédier dans le chapitre sur les variables.
7. ព្រៃស្តុក, « Forêt vierge » terme lexicalisé en français qui ne l'est pas en khmer et qui est traduit par le substantif ព្រៃ « forêt » suivi de l'adjectif ស្តុក « dense ».

Les mots polysyllabiques sont des emprunts et ne sont pas décomposables. Voici un tableau des mots polysyllabiques présents dans le premier chapitre du Petit Prince :

Lemme	Origine	Translittération	Sens
ប្រយោជន៍	sanskrit	pra-yojana	cause
សៀវភៅ	cantonais	shü + pò	livre + volume
កីឡា	pâli	kīlā	sport

ជីវិត	sanskrit, pali	jīvita	vie
ទស្សនៈ	pâli	dassana	perception
នយោបាយ	sanskrit, pâli	naya	sagesse
ប្រវត្តិ	sanskrit, pâli	pavitta	événement
នព្វន្ត	sanskrit, pâli	nava + anta	chiffres
វេយ្យាករណ៍	pâli	veyyākaraṇa	grammaire
អារម្មណ៍	pâli	ārammaṇa	sentiment
អារីសូណា	anglais	arizona	Arizona
អាហារ	sanskrit, pâli	āhāra	nourriture
ប្រទេស	sanskrit	pra-deśa	pays
ពិចារណា	pâli	vicāraṇā	examiner

Tableau 2 Emprunts

3.3.2 Morphologie

Une fois la segmentation réalisée, nous pouvons passer à l'analyse morphologique. Langue isolante, le khmer a une morphologie relativement pauvre. Pourtant, il existe des cas de dérivation par affixation et contraction qu'il est possible de traiter grâce à ATEF. Malheureusement, s'il existe des travaux sur ces procédés de dérivation, aucun ne fait pour l'instant consensus. Nous avons donc décidé de ne pas les traiter pour l'instant, mais de les lister pour une étude ultérieure. Les mots infixés seront donc traités comme des mots autonomes, ce qui ne change pas le résultat de l'analyse.

Mot racine	Mot dérivé
គូរ	គំនូរ
ដើរ	ដំនើរ
រលាយ	រំលាយ
ក្រង	កំរង
ក្រើក	កំរើក
ឆី	ចំណី
ស្រាប់	សំរាប់
យល់	ពន្យល់
ក្លែង	កន្លែង

Tableau 3 Affixes

Avec l'infixe « ម », le mot « គូរ » (dessiner) à la première ligne du tableau devient « គំនូរ » (dessin). Ces deux termes devraient ainsi disposer d'une unité lexicale (UL) commune, le nom étant reconnu comme tel grâce à la variable DRV qui, affectée de la valeur VN, indique que le mot est une dérivation verbe-nom.

3.3.3 Choix des variables

Une fois la segmentation réalisée et la question des phénomènes morphologiques reportée à plus tard pour les raisons expliquées ci-dessus, nous pouvons passer à l'étiquetage morphosyntaxique. Pour ce faire, nous reprenons le tableau utilisé en segmentation, et nous ajoutons simplement une colonne pour la catégorie morphosyntaxique :

Numéro	Mot	Catégorie morphosyntaxique
1	កាល	
2	ពី	
3	ខ្ញុំ	
4	អាយុ	
5	ប្រាំ	
6	មួយ	
7	ឆ្នាំ	
8	’	
Etc.	Etc.	

Tableau 4 Choix des variables

3.3.3.1 Catégorie morphosyntaxique

Nous somme parti du jeu d’étiquettes utilisé pour le français pour le modèle CALLIOPE AERO[24] que nous avons adapté aux besoins spécifiques du khmer. Chaque catégorie peut avoir une sous-catégorie (par exemple pour distinguer les adjectifs des adverbes dans la catégorie des adjoints), et certaines sont associées à une classe syntagmatique (K), ce que nous verrons dans la partie suivante.

24 « Le système de traduction français-anglais CALLIOPE AERO développé dans le cadre du Projet National TAO par la société B'VITAL en collaboration avec le GETA, la société SG2 (maître d’œuvre du projet) et la société SONOVISION chargée des dictionnaires et de la terminologie aéronautique. Son développement a commencé en 1984 et s'est poursuivi jusqu'en 1990. Il est dédié à la traduction de la documentation pour la maintenance d'appareils. Spécifié et développé à partit des documents du Mirage 2000 et testé sur les documents du Falcon 900. » (Chappuy, 2013)

CAT	CATégories morphosyntaxiques
V	Verbe
N	Nom
A	Adjoint (adjectifs ou adverbes)
D	Déterminant (articles et adjectifs possessifs, démonstratifs)
R	Représentant (pronoms)
S	Subordonnant (prépositions et conjonctions de subordination, locutions du même type)
C	Coordonnant (conjonctions simples ou « à balance »)
P	Signes de ponctuation
PREF	PREFixe (ex : post- anti- non-)
INC	Catégorie des mots inconnus i.e. ceux qui ne sont pas dans le dictionnaire.
EDIT	Signes d'édition
NA	Non alphabétique

Tableau 5 Catégories morphosyntaxiques du français

Nous avons pour l'essentiel repris les catégories morphosyntaxiques du français (sauf PREF, EDIT et NA dont nous ne sommes pas servis).

Nous avons ajouté la catégorie particule (PART) pour certains modificateurs se plaçant devant un mot. Il aurait été possible de conserver la catégorie PREF, cela mériterait une étude plus poussée. Exemple :

របស់

dont la fonction première est substantif avec pour signification « chose », sera employé entre un sujet et un objet pour définir une relation d'appartenance :

គំនូររបស់ខ្ញុំ

mon dessin

De même, ទី est un substantif « lieu » mais aussi « rang » placé devant un cardinal aura pour fonction de le transformer en ordinal :

មួយ / un → ទីមួយ / premier

Le jeu de variables retenu pour la partie khmère de notre système est résumé dans le tableau ci-dessous.

CAT	CATégories morphosyntaxiques
V	Verbe
N	Nom
A	Adjoint (adjectifs ou adverbes)
D	Déterminant
R	Représentant
S	Subordonnant
C	Coordonnant
P	Signes de Ponctuation
PART	PARTicules
INC	Catégorie des mots inconnus i.e. ceux qui ne sont pas dans le dictionnaire.

Tableau 6 Catégories morphosyntaxiques du khmer

Le cas de la particule de duplication (ៗ) a aussi soulevé la question de la création d'une catégorie. Ce signe est essentiellement une marque de lecture. Ce n'est ni une ponctuation, ni un caractère alphabétique, ni un chiffre (bien qu'il soit dérivé du chiffre 2 : ២). Créer une catégorie pour un élément seul ne serait pas pertinent, et n'est pas nécessaire. En effet, nous pouvons nous servir directement de son UL, qui est une information utilisable comme une variable, pour l'identifier au sein de grammaire d'analyse.

Les variantes orthographiques que nous avons rencontrées peuvent aussi être signalées par une variable.

Khmérisation	Chuon Nath
ដំនើរ	ដំណើរ
ចំនង	ចំណង
ចំនី	ចំណី
ទស្សន:	ទស្សន

Khmérisation	Chuon Nath
រឺ	ឬ
អោយ	ឲ្យ
ម៉ត់	ហ្មត់

Tableau 7 Orthographes

3.3.3.2 Unité Lexicale

Gardant à l'esprit l'idée que cette maquette servira de modèle pour d'autres langues, nous avons choisi d'utiliser temporairement les UL(lemme dérivationnel) du français afin de travailler en parallèle avec d'autres langues. Une version des UL en anglais sera ultérieurement créée pour des utilisateurs non-francophones.

Numéro	Mot	Catégorie morphosyntaxique	UL
1	កាល		
2	ពី		
3	ខ្ញុំ		
4	អាយុ		
5	ប្រាំ		
6	មួយ		
7	ឆ្នាំ		
8	’		
Etc.	Etc.		

Tableau 8 Unité lexicale

3.3.4 Génération du dictionnaire ATEF

3.3.4.1 Moulinette de génération automatique de dictionnaires au format ATEF : DB2ATEF

L'écriture d'un dictionnaire en ATEF peut s'avérer fastidieuse. Afin de remédier à cela, Vincent Berment a écrit un programme en C++ capable de traiter un fichier de base de données et de produire automatiquement le dictionnaire. Les étapes sont les suivantes :

- Reprendre notre tableau 8 intitulé Unité Lexicale ou réaliser un tableau exportable au format csv (Excel...), nommé par exemple VocabulairePPChap1.xls, avec quatre colonnes :
    - Numéro de ligne,
    - Lemme,
    - Catégorie morphosyntaxique (suivant la liste du tableau 2),
    - Unité lexicale (UL, basée sur la traduction en français),
- Exporter le tableau au format csv avec une tabulation comme séparateur et le passer en UTF-8 si nécessaire sous le nom VocabulairePPChap1.csv,
- Lancer une version en ligne de commande de sqlite :
    - sqlite3 VocabulairePPChap1.db
- Taper sous sqlite :
    - CREATE TABLE Dictionnaire(Cle,Article,UL,Cat);
    - .separator "\t"
    - .import DicoAMChap1UTF8.txt Dictionnaire

La base de données ainsi créée est facilement manipulable et la moulinette « DB2ATEF » génère automatiquement le dictionnaire ATEF correspondant.

Exemple d'entrée du dictionnaire (le nombre d'espaces a été réduit pour tenir sur une ligne) :

កន្លែង ==STEM (NOM ,lieu).

Les autres fichiers de l'analyseur morphologique sont créés manuellement, mais sont rapides à réaliser. Chaque classe morphosyntaxique (représentée par « NOM » dans l'entrée de dictionnaire ci-dessus) fait l'objet d'un « format syntaxique » ATEF :

```
NOM     01== Créé manuellement le 29/03/2014.
NOM     02 ** CAT-E-N.
```

Ce format permettra d'affecter N à la variable CAT lorsque ce mot sera analysé par la grammaire ATEF. Plus précisément, lorsque le mot កន្លែង sera analysé par ATEF, la règle RFULL de la grammaire sera déclenchée via le « format morphologique » STEM trouvé dans l'entrée.

```
RFULL:   STEM ==
        VARS(C):=VARS(A),VAREM(C):=VAREM(A),VARNM(C):=VARNM(C)-U-VARNM(A)/
        SCHAINE(A,0,1)-E-' '.
```

L'ensemble des variables contenu dans les formats de l'entrée de dictionnaire sera affecté au nœud créé (VARS(C):=VARS(A),VAREM(C):=VAREM(A)...), de même que l'UL (affectation par défaut). Le test « SCHAINE(A,0,1)-E-' ' » est une condition d'exécution de la règle qui vérifie qu'on est bien sur une fin de mot, et non sur un découpage intermédiaire.

Les différents fichiers, créés soit manuellement (variables, formats, grammaire), soit automatiquement (dictionnaire), sont ensuite placés dans l'environnement ASEAN du serveur Héloïse[25] où ils sont compilés pour produire une première AM du khmer permettant de produire un arbre d'analyse morphologique dont le début (« quand j'avais six ans ») est reproduit dans la copie d'écran ci-dessous. Cet arbre sera l'entrée de la phase d'analyse suivante : l'analyse structurale.

[25] http://www.taranis-software.com/Heloise/ASEAN

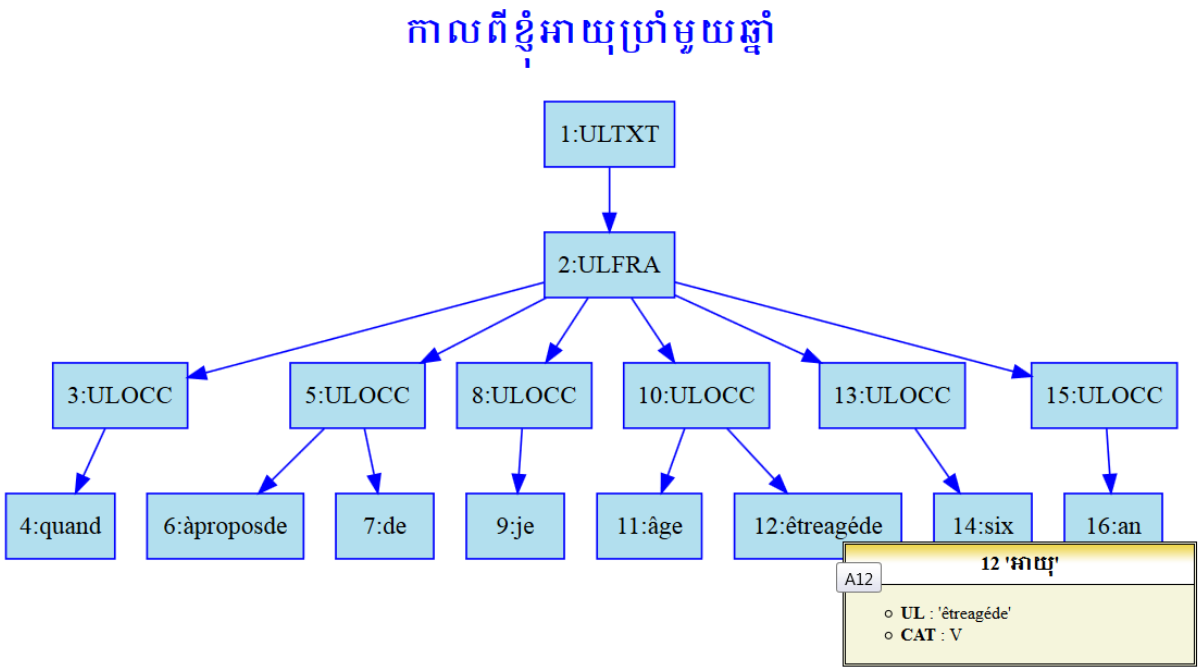


Figure 5 Affichage graphique du résultat d'analyse morphologique

3.3.4.2 Gestionnaire en ligne de dictionnaire monolingue : GDML

Nous avons vu qu'en regroupant selon leurs besoins en décoration nous pouvions utiliser DB2ATEF pour générer un dictionnaire ATEF à partir d'une base de données. Cette base de données purement morphologique que nous avons créée doit être enrichie d'informations d'ordre syntaxique, logique pour les mots prédicatifs... L'outil en ligne GDML[26] permet de gérer cet enrichissement de manière ergonomique et avec l'approche dite de « dictionnaire mille-feuilles » de Sylviane Chappuy (Chappuy, 2013) qui consiste à séparer les informations lexicales selon leur nature : morphologique, dérivationnelle, syntaxique, sémantique et argumentaire. L'outil GDML permet d'ajouter, de modifier et de supprimer des mots, et de leur attribuer un ensemble de variables cohérent (toutes les variables ne sont pas compatibles, par exemple, la sous-catégorie « nom commun » ne peut cohabiter avec la catégorie « adjoint ».

Les figures suivantes montrent les différentes « couches » lexicales du verbe « aller », provenant de l'analyseur du système FR3-AN3.

[26] http://gmsware.org/Taranis/GestionDicosML/index.php#

Figure 6 GDML, couche morphologique

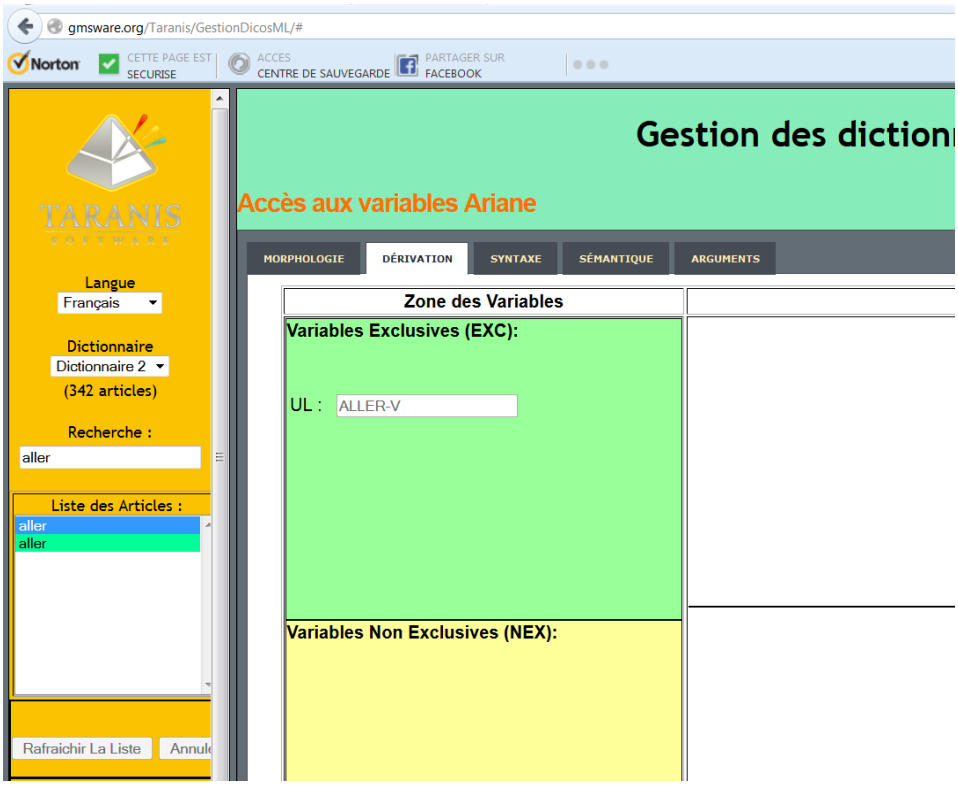


Figure 7 GDML, couche dérivationnelle

Figure 8 GDML, couche syntaxique

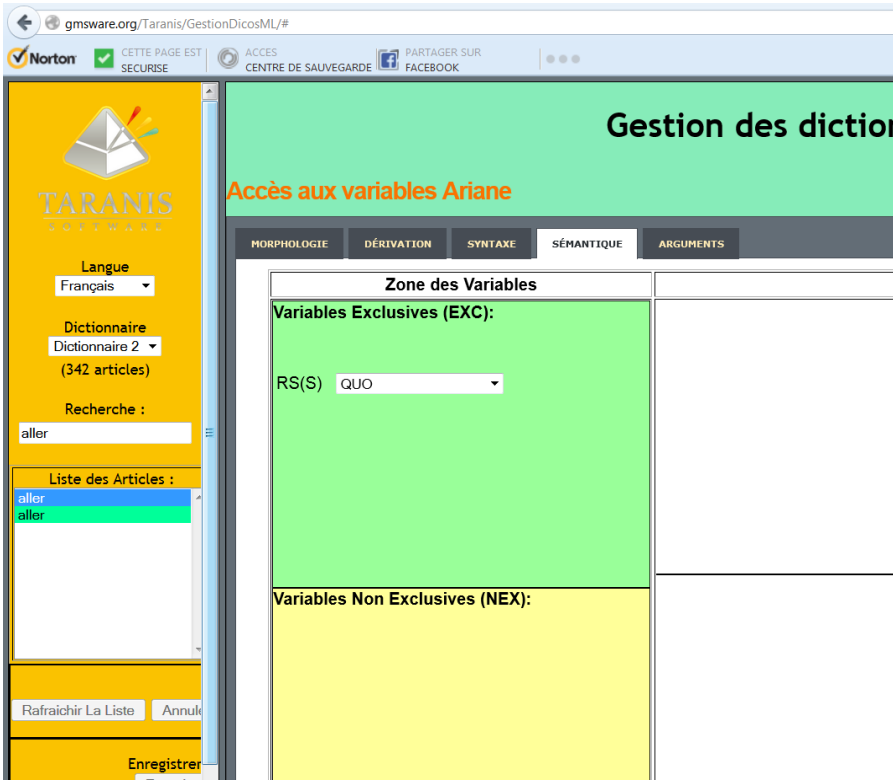


Figure 9 GDML, couche sémantique

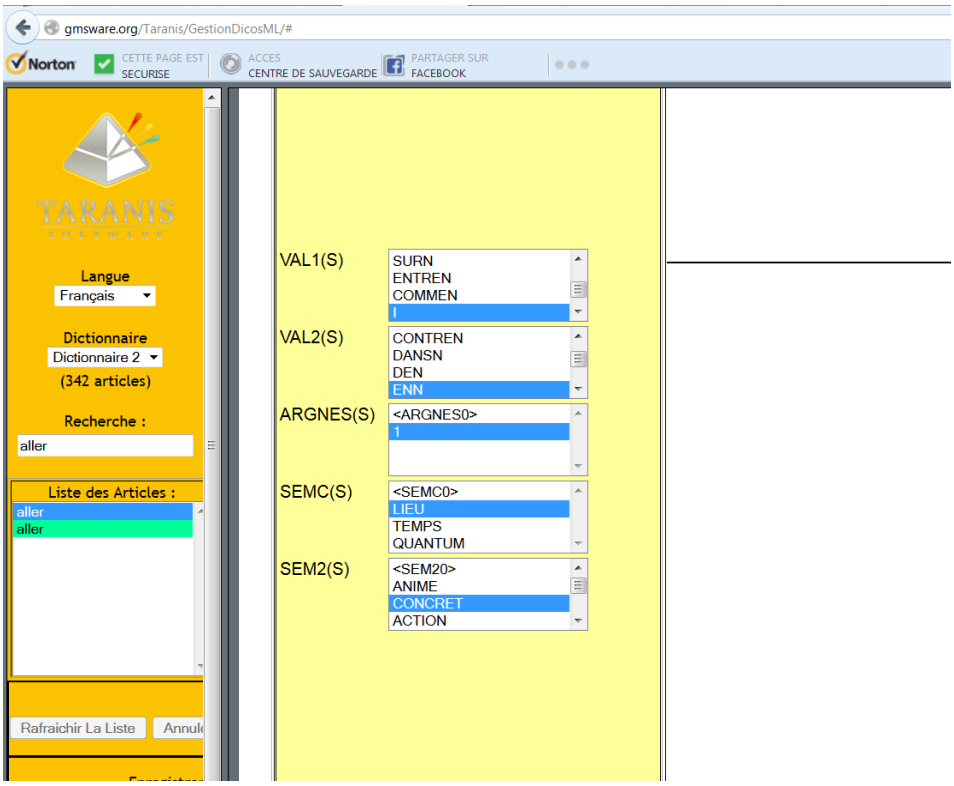


Figure 10 GDML, couche argumentaire

Nous voyons, par exemple, sur cette dernière copie d'écran, que le linguiste a choisi « I » comme rection, c'est à dire un infinitif, pour l'argument 1 de « aller » (« aller **faire**... »), et ENN, c'est à dire « en + nom » pour l'argument 2 (« aller **passer** ses vacances **en** France ». Ce choix correspond à la typologie du corpus à traiter par ce système FR3-AN3. Nous observons par ailleurs que la variable ARGNES est à 1 (un argument est nécessaire), que la sémantique attendu pour l'argument privilégié est un lieu et pour l'argument 2 un objet concret.

GDML est actuellement adapté au dictionnaire d'AM du système FR3. Nous avons commencé à étudier comment le généraliser à n'importe quel type de langue. Hormis des aménagements techniques comme le passage sous Unicode (la base du français, développée dans les années 1980, est restée en Latin 1), l'extension à des langues isolantes comme le khmer est trivial, ce type de langues pouvant être vu comme une restriction des langues flexionnelles comme le français dans lequel il n'existe qu'un paradigme morphologique consistant en la classe des mots invariables. Par contre, nous n'avons pas encore de solution satisfaisante pour étendre GDML aux langues agglutinantes comme le quéchua, qui s'est rallié récemment au projet « Petit Prince ».

3.4 Analyse structurale

Après avoir analysé la morphosyntaxe de notre corpus en attribuant pour chaque mot une valeur à la variable CAT, nous procédons à l'analyse structurale pour déterminer la constitution des classes syntagmatiques identifiées par la variable K.

3.4.1 Analyse manuelle

Comme pour l'analyse morphologique, nous reprenons le jeu de variables des classes syntagmatiques du français afin de l'adapter au khmer :

PVB	Proposition verbale (conjuguée)
PPART	Proposition participiale
PINF	Proposition infinitive
PREL	Proposition relative
PSUB	Proposition subordonnée ou complétive
NV	Noyaux verbal (choix stratégique de regrouper autour du verbe ce qui le modifie, négation, modalité, adjoints)
GN	Groupe nominal
GADJ	Groupe adjectival
GADV	Groupe adverbial
GORD	Groupe ordinal
GCARD	Groupe cardinal

Tableau 9 Classes syntagmatiques

Définir les classes syntagmatiques d'une langue demande une étude encore plus poussée que de définir les parties du discours. Nous n'avons pas trouvé de description linguistique suffisamment couvrante pour déterminer l'ensemble des classes syntagmatiques du khmer, alors nous avons décidé de procéder de façon hiérarchique toujours en nous aidant du modèle français.

Nous avons tout d'abord conservé les classes syntagmatiques qui sont les extensions directes de nos classes morphosyntaxiques :

Classes syntagmatiques	Catégories morphosyntaxique / sous-catégories
NV	V
GN	N
GADV	A / ADV
GADJ	A / ADJ
GORD	A / ORD
GCARD	A / CARD

Tableau 10 Classes syntagmatiques et leur catégories morphosyntaxiques associées

3.4.1.1 Les cardinaux et les ordinaux

Il existe aussi une hiérarchie entre ces classes syntagmatiques. Les classes GORD et GCARD étant les plus élémentaires, nous avons commencé par elles.

Pour ce faire, nous avons analysé chaque phrase du corpus en les plaçant dans un tableau, à l'horizontale cette fois :

Lemme	កាល	ពី	ខ្ញុំ	អាយុ	ប្រាំ	មួយ	ឆ្នាំ	,	Etc.
CAT	S	C	R	V	A / CARD	A / CARD	N	P	Etc.
K			GN	NV	GCARD	GCARD	GN		Etc.

Tableau 11 Cardinaux et ordinaux

Nous plaçons systématiquement chaque cardinal sous un groupe cardinal. Nous constatons que nous pouvons placer sous un même groupe cardinal les groupes cardinaux apparaissant côte à côte. Ici ប្រាំ « cinq » et មួយ « un » forment ប្រាំមួយ « six ». Notre exemple est un peu particulier, mais il reflète le fonctionnement de la langue qui consiste à former une idée en accolant deux mots. Ce que nous définirons dans le prochain chapitre comme une règle de formation des groupes cardinaux est valable pour tous les cardinaux comme par exemple mille vingt-quatre : មួយ« un » ពាន់ « mille » ម្ភៃ« vingt » បួន« quatre ». Nous avons bien quatre groupes cardinaux côte à côte que l'on peut regrouper sous un seul groupe.

Une fois les cardinaux identifiés, nous passons à l'étude des ordinaux. En khmer, un groupe ordinal est formé d'un cardinal que l'on préfixe avec la particule ទី que l'on appellera particule ordinale de par son rôle.

3.4.1.2 Les pronoms personnels

Les pronoms personnels que nous avons rencontrés sont de deux natures.

- Simples :
    - Nous les avons précédemment classés dans la catégorie morphosyntaxique des représentants. Nous mettons un groupe nominal au dessus de chacun d'eux.

	ខ្ញុំ	វា	គេ
PERS	1	3	3
NB	SING	SING	SING

Tableau 12 Pronoms personnels

- Complexes :
    - Il existe plusieurs types de pronoms personnels complexes, composés d'un pronom personnel simple auquel on appose un ou plusieurs termes pour le préciser. Nous n'avons rencontré qu'un seul de ces pronoms personnels complexes, apparaissant plusieurs fois dans notre corpus : ពួកគេ.
        - Nous retrouvons le pronom គេ que l'on a préfixé par le substantif ពួក « groupe » qui prend ici le rôle de « collectif » afin d'apporter l'information du pluriel.

3.4.1.3 Les noms composés

Nous avons identifié clairement deux types de mots composés, tous deux formants des noms :

- Mots composés d'un collectif suivi d'un nom :

  Ce procédé n'est pas sans rappeler celui que nous avons vu précédemment avec les pronoms personnels complexes. Nous n'avons rencontré qu'un seul cas dans notre corpus : កំរងរឿង « histoires » កំរង « collection » រឿង « histoire ». Si cette règle semble cohérente, il sera nécessaire de la confirmer avec un nombre plus grand d'exemples.

- Mots composés d'un hyperonyme précisé par un hyponyme :

  Ce deuxième type de mot composé est beaucoup plus présent dans notre corpus. Chacun des deux termes peut apparaître séparément, mais avec des sens différents (éventuellement dans un autre mot composé).

Mot composé	Hyperonyme	Hyponyme
ពស់ថ្លាន់ « python »	ពស់ « serpent »	ថ្លាន់ « python »
សត្វម្រឹគ « animal sauvage »	សត្វ « animal »	ម្រឹគ « animal sauvage »
រយៈពេល « moment »	រយៈ « période »	ពេល « moment »
រូបគំនូរ « dessin »	រូប « image »	គំនូរ « dessin »
ប្រទេសចិន « Chine »	ប្រទេស « pays »	ចិន « Chine »
រដ្ឋអារីសូណា « Arizona »	រដ្ឋ « État »	អារីសូណា « Arizona »
ពេលយប់ « nuit »	ពេល « temps »	យប់ « nuit »

Tableau 13 Mots composés

3.4.1.4 Les groupes adjectivaux

Nous avons rencontrés trois types de groupes adjectivaux :

- Les adjectifs seuls :

  ស្តុក « dense »

- Les adjectifs suivis du signe de répétition :

  ពិត « vrai » ៗ « signe de répétition »

  À l'oral, l'adjectif sera lu deux fois et cela aura pour effet d'intensifier d'une façon assez proche de l'ajout de l'adverbe « très » en français ou de la répétition elle-même dans certains cas comme « vite vite ! ».

- Les adjectifs préfixés par la particule adverbiale ដ៏, qui peut avoir deux rôles :

  - Devant un adjectif, elle aura un effet similaire à celui du signe de répétition :

    ដ៏ « particule adverbiale » អស្ចារ្យ « spectaculaire » signifiera « très spectaculaire »,

- Devant un nom, elle le transforme en adjectif :

  ដ៏ « particule adverbiale » មហស្ចារ្យ « spectacle » signifiera « spectaculaire »,

- Nous remarquons que ដ៏អស្ចារ្យ et មហស្ចារ្យ sont équivalents et tous deux employés dans notre corpus. Cela est certainement dû à un effet de style, l'usage de la particule ដ៏ étant plutôt littéraire.

3.4.1.5 Les groupes nominaux

Identifier les groupes nominaux a été beaucoup plus complexe.

Nous avons décrit quatre groupes nominaux supplémentaires, tous rencontrés dans notre corpus :

- Cardinal + unité de temps

  Normalement, le nombre se place après un nom pour le compter, avec la présence éventuelle d'un classificateur[27]. Mais dans le cas d'une unité de temps[28] le nombre se place avant et sans avoir besoin d'un classificateur[29] :

  ប្រាំមួយ « six » ឆ្នាំ « ans ».

- GN + GADJ

  Suivant l'ordre déterminé-déterminant, un groupe nominal sera associé au groupe adjectival qui le suit :

  ព្រៃ/ស្ដុក

  Forêt/dense

- Relation d'appartenance

  - La relation d'appartenance se créer en plaçant un pronom personnel après un nom avec la présence optionnelle de la particule របស់.

    គំនូរ/ទីមួយ/របស់/ខ្ញុំ : mon premier dessin

    dessin/premier/particule/je

- Subordonnant introduisant un groupe nominal (comme dans l'analyseur du français qui nous sert de guide, les groupes prépositionnels sont regroupés dans la classe syntagmatique des GN) :

[27] Présence d'un classificateur beaucoup moins régulière et indispensable qu'en chinois
[28] Années, mois, jours, heures, minutes, secondes, etc.
[29] En mandarin les unités de temps portent la classification et ne sont pas doublées lorsqu'elles sont associées à un nombre.

ក្នុង/សៀ/ភៅ/មួយ : dans un livre

dans/livre/un

ពី/ព្រៃ/ស្តុក : (à propos) de la forêt vierge

de/forêt/dense

3.4.2 Code ROBRA

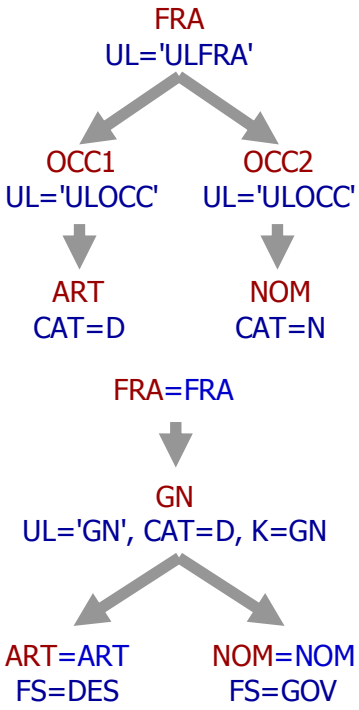
ROBRA est l'un des langages spécialisés pour la programmation linguistique d'Ariane. Il réalise des transformations d'arborescence à la manière des transformations de chaînes que peuvent réaliser les expressions régulières. C'est le même principe mais le fait que les transformations soient réalisées sur des arborescences rend son utilisation sensiblement plus complexe que les expressions régulières.

La transformation de base est réalisée par une règle dans laquelle un « schéma » (forme arborescente recherchée) associé à des conditions constitue la partie gauche, et une « image » associée à des affectations (éventuellement conditionnelles) constitue la partie droite. Les deux parties sont séparées par le symbole « == ».

Par exemple (les exemples de ce chapitre sont issus d'une présentation faite par Vincent Berment à Systran), la règle :

```
ARTNOM01 : (*FRA, &NIV=1)
        FRA (OCC1(ART),*,OCC2(NOM)) /
        FRA:UL-E-'ULFRA'; OCC1:UL-E-'ULOCC';
        OCC2:$ULOCC; ART:CAT-E-D;
        NOM: CAT-E-N
        ==
        FRA(GN(ART,NOM))  /
        *<--OCC1;*<--OCC2;GN<--* /
        GN:NOM, K:=GN, UL:='GN';
        ART:ART, FS:=DES;
        NOM:NOM, FS:=GOV.
```

traite l'analyse des GN du français constitués d'un article et d'un nom. Sous forme graphique, on voit plus facilement l'opération réalisée.

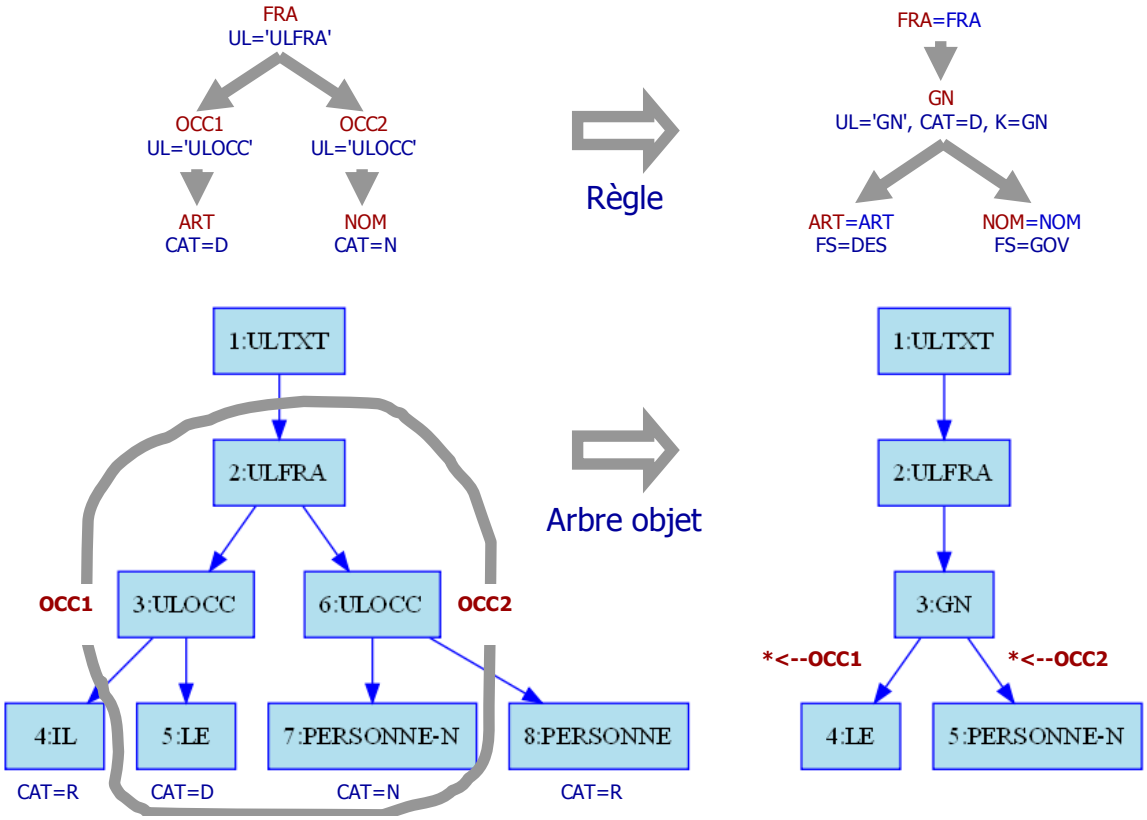


Schéma

Image

```
ARTNOM01 : (*FRA, &NIV=1)
    FRA (OCC1(ART),*,OCC2(NOM)) /
    FRA:UL-E-'ULFRA'; OCC1:UL-E-'ULOCC';
    OCC2:$ULOCC; ART:CAT-E-D;
    NOM: CAT-E-N
    ==
    FRA(GN(ART,NOM))  /
    *<--OCC1;*<--OCC2;GN<--* /
    GN:NOM, K:=GN, UL:='GN';
    ART:ART, FS:=DES;
    NOM:NOM, FS:=GOV.
```

Nous remarquons au passage que cette règle a un rôle de désambiguïsation.

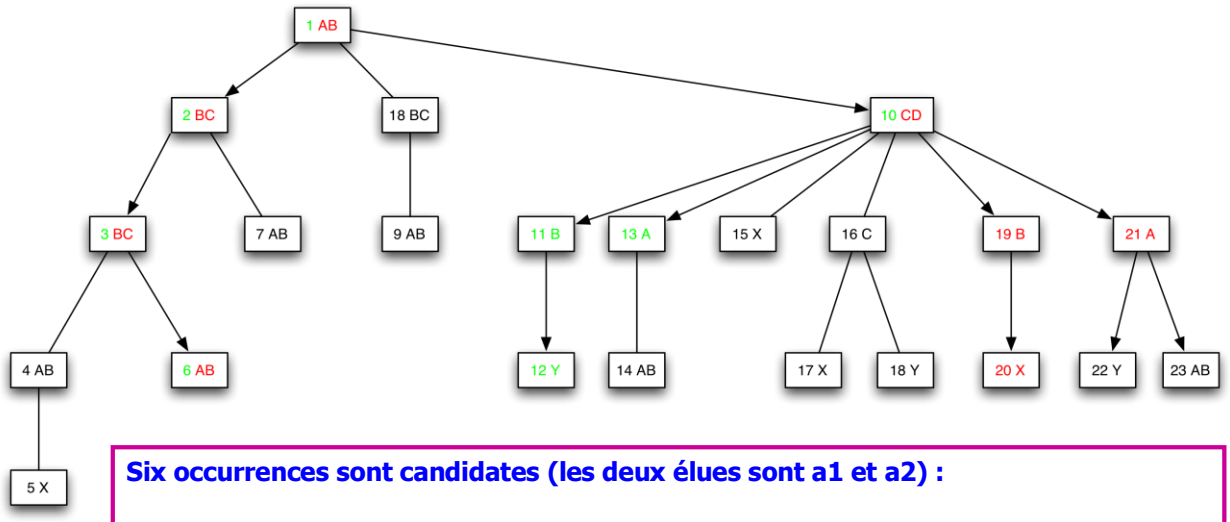


Le texte analysé est ici « la personne ». Le mot « la » peut être article ou pronom, et « personne » peut être nom ou pronom. La règle contient donc l'information de nature syntaxique qui est que parmi les combinaisons possibles [article | pronom] x [nom | pronom], seule la séquence article x nom est possible.

Les règles peuvent être plus complexes, comme la règle suivante.

```
R1 : (*3, &NIV < 3) 1(?2&B(*),3&ND(A,*,B($L)))    /
        1:UL -E- 'AB';
        2:UL -E- 'AB'; 2?: UL -E- 'BC';
        3:UL -E- 'CD' -OU- UL -E- 'X'; A:UL -E- 'A'; B:UL -E- 'B';
        $L: UL -NE- 'AB' /
==   3(C,$L,4(A),1)
        /*<--B; 1<--A; C<--*; 4<--*; A<--*/
        4:2; 1:3; C:*FTNUL, UL:='C'.
```

Appliquée à un arbre objet comportant plusieurs occurrences possibles du schéma, il peut y avoir des conflits et un ordre a donc été défini pour n'en retenir qu'une partie.

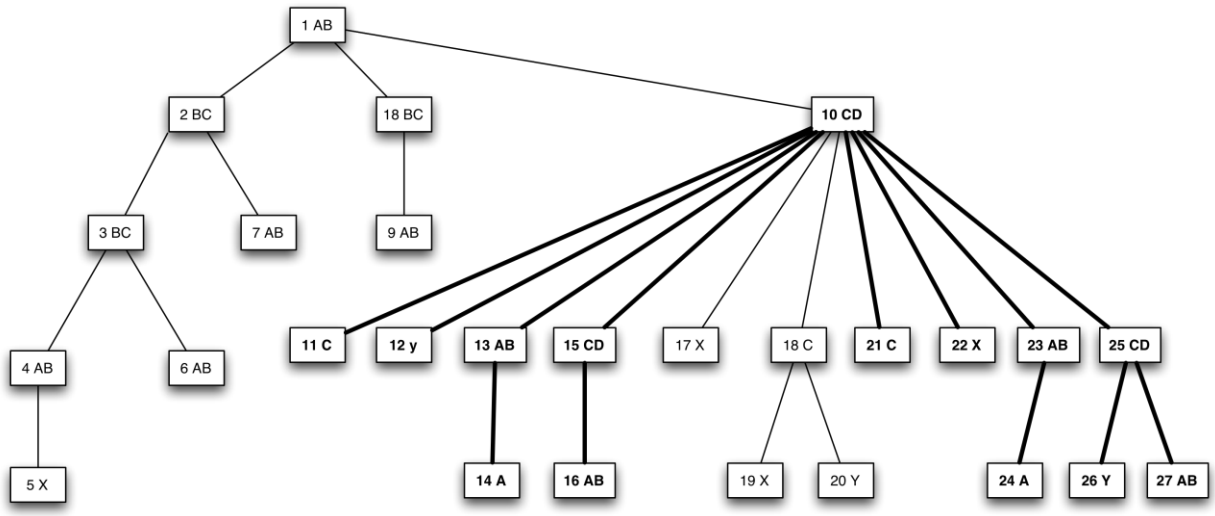


Six occurrences sont candidates (les deux élues sont a1 et a2) :

(a1) 1(?3(6),10(11(12),13))
(b1) 1(?2(7),10(11(12),13))
(c1) 1(?8(9),10(11(12),13))

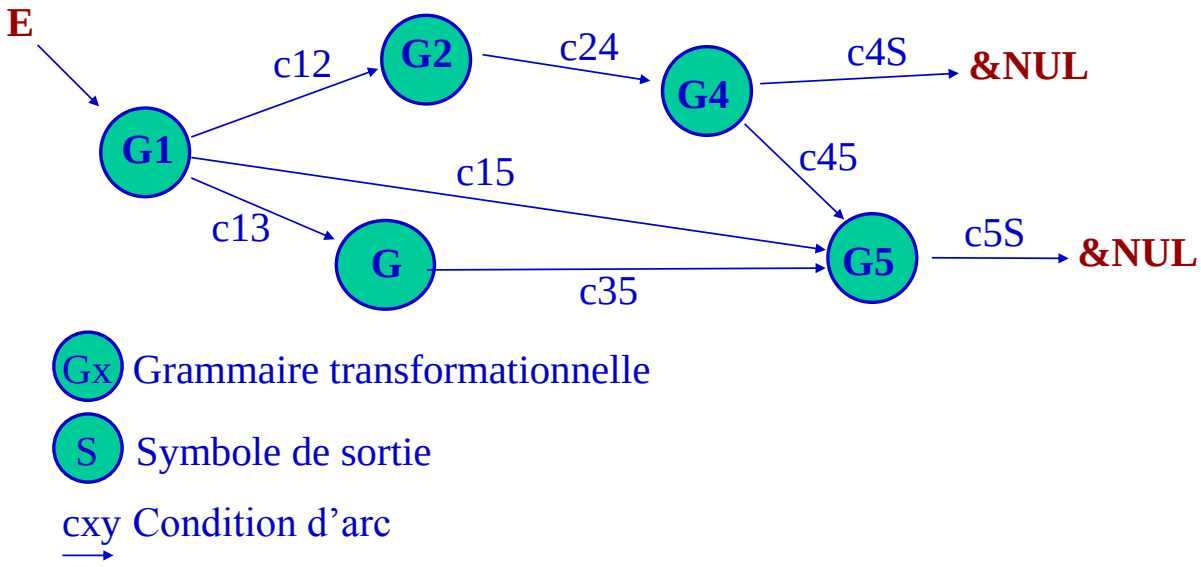
(a2) 1(?3(6),10(19(20),21))
(b2) 1(?2(7),10(19(20),21))
(c2) 1(?8(9),10(19(20),21))

Le résultat de la transformation est donné ci-dessous (en gras, les forêts substituées).



Les règles sont regroupées dans des « grammaires » dans lesquelles elles s'exécutent « en parallèle », c'est à dire que les occurrences éventuellement trouvées pour les schémas de ces différentes règles sont transformées dans le même temps, l'ordre sur les règles induisant l'ordre sur les occurrences et le mécanisme de priorité évoqué précédemment pour une règle seule.

Ces ensembles ordonnés de règles, les « grammaires », s'enchaînent dans ce qui est appelé un « graphe de contrôle », et que l'on peut voir comme un programme en C dans lequel les instructions seraient les grammaires et les arcs du graphe des conditions « si-alors-sinon ».



Gx : R1, R2, ..., Rn
Rx : <schéma + condition> == <image remplaçant le schéma>

Un mécanisme de retour arrière permet de revenir d'une branche qui échoue.

Donnons un exemple de passage à ROBRA de nos planches structurales avec la règle R-GN3a. Cette règle dit qu'un nœud ayant pour variables K=GN et CAT=N, suivi d'un nœud ayant pour variables K=GADJ, CAT=A et SUBA=ADJ est un groupe nominal constitué d'un nœud ayant pour variable K=GN qui domine les deux précédent. La règle ROBRA suivante en découle directement (règle très proche, dans sa forme, de celle donnée plus haut).

```
RGN3a : (*FRA, &NIV=1)
        FRA (OCC1(GN),*,OCC2(GADJ)) /
        FRA:UL-E-'ULFRA'; OCC1:UL-E-'ULOCC';
        OCC2:$ULOCC; GN:K-E-GN, CAT-E-N;
        GADJ: K-E-GADJ, CAT-E-A, SUBA=ADJ
        ==
        FRA(GN1(GN,GADJ))  /
        *<--OCC1;*<--OCC2;GN<--* /
        GN1:GN, K:=GN, UL:='GN';
        GN:GN;
        GADJ:GADJ.
```

3.5Transfert et générations

3.5.1 Transfert

L'étape appelée « transfert » est constituée au minimum de deux phases :

- Une phase de transfert lexical dans laquelle on passe de l'espace lexical de la langue source dans celui de la langue cible,
- Une phase de transfert structural dans laquelle on réalise des modifications de structures qui sont nécessaires, en complément du changement d'espace lexical, pour présenter à la génération structurale un arbre conforme à sa spécification d'entrée, cette phase pouvant aussi avoir un rôle de désambiguisation contrastive.

L'environnement Ariane permet d'ajouter d'autres phases mais nous avons conservé le principe en deux phases de transfert du système anglais-français duquel nous sommes partis.

Transfert lexical

Le changement d'espace lexical consiste en une traduction des unités lexicales de la langue source dans la langue cible et une transposition des informations lexicales portées par les variables. Pour cela, l'arbre objet[30] est parcouru de manière préfixée, et les transformations sont réalisées sur chaque nœud lors du parcours.

Concernant la traduction des UL, celle-ci peut être simple comme (exemples en fr-en) :

MARTEAU-N ➔ HAMMER-N.

Mais elle peut être complexe, comme par exemple pour traduire « mondialiser » en « spread

[30] C'est ainsi que sont nommées les structures arborescentes manipulées en interne par le système de traduction.

worldwide » qui se traduit par la création d'un petit arbre en substitution du nœud de l'arbre en langue source :

MONDIALISER-V ➔ 0(1:SPREAD-V,2:*ADVP(3:WORLD-WIDE)).

En cas de polysémie, la même technique de création d'un petit arbre de substitution est employée :

MOYEN-N ➔ 0:WAY-N(1:MEANS-N).

Lorsqu'il y a plusieurs solutions, un test peut être effectué sur le voisinage du nœud (les éventuelles conditions portent sur le nœud transformé, son père et ses frères immédiats), pour choisir l'UL et les variables les plus pertinentes, par exemple :

NAVIGUER-V ➔ si condition NAVIGATE-V sinon FLY-V

Lorsqu'un test dans le voisinage immédiat du nœud n'est pas suffisant, il est aussi possible de simplement préparer les données nécessaires et de laisser le transfert structural réaliser la désambiguïsation. Dans l'exemple ci-dessous correspondant à la traduction du verbe « prendre », selon l'argument 0, le transfert structural choisira *break*, *set* ou sinon *take* :

PRENDRE-V ➔ 0:GARG0(1:FIRE-N(2:BREAK-V),3:NOT-UL(4:SET-V),5:TAKE-V)

Transfert structural

Dans le transfert structural, on prépare l'arbre objet pour la génération structurale. Pour cela :

- On traite les conditions contextuelles codées dans des sous-arbres représentant des traductions multiples, pour réduire les polysémies
- On traite les traductions de groupes complexes
- On adapte la structure d'interface cible, en y codant des conseils ou des ordres au générateur en vue de produire telle ou telle forme syntaxique :
    - variables d'actualisation des verbes (aspect, temps, voix, mode),
    - détermination des groupes nominaux.

Nous n'avons pas pu réaliser entièrement cette phase de transfert. Cependant, nous avons pu vérifier le principe sur des exemples. Notre corpus et donc notre dictionnaire en langue source étant très petits, les ambiguïtés sont rares et la plupart des transferts se réduisent à un transfert lexical simple du type « marteau ➔ hammer » évoqué ci-dessus, la transposition des variables étant immédiate (nous avons choisi un système de variables extrait de celui du français, qui est notre langue cible). Le transfert structural, quant à lui, a été réduit au minimum : il ne réalise aucune transformation sur l'arbre objet.

3.5.2 Adaptation de la génération réutilisée

Notre corpus, s'il est petit, est cependant assez différent de celui qui avait servi à la réalisation du système anglais-français qui nous sert de base. Ce système était destiné en particulier à la traduction de textes techniques dans le domaine informatique. De nombreux mots de notre corpus étaient ainsi absents du système dont nous avons réutilisé la génération. Par exemple, les mots suivants sont absents : *Forêt Vierge, serpent boa, fauve, mâcher, digestion, jungle*... Pour réaliser un système khmer-français capable de traduire correctement notre corpus, il est donc nécessaire d'ajouter les entrées manquantes dans l'un des dictionnaires de génération morphologique du français.

Il est aussi possible que certains mécanismes syntaxiques soient absents de la génération structurale du français de laquelle nous sommes partis. Cependant, faute de temps, nous n'avons pas pu aller jusque là et l'essentiel de ce travail sur les générations structurale et morphologique reste à réaliser.

4 Expérience de réplication sur plusieurs langues peu dotées

Afin de mettre à l'épreuve nos expérimentations, nous les avons transposées à un petit groupe de langues.

4.1 Le site et la communauté Lingwarium

Lingwarium est le nom donné à l'ENT (Espace Numérique de Travail) mis en place pour aider les développeurs-linguistes réalisant des systèmes de TA sous Héloïse. Le site est accessible à l'adresse http://www.lingwarium.org. Il est basé sur le collecticiel[31] libre Agora-project[32] qui permet :

- De partager :
    - Fichiers
    - FAQ
- D'organiser :
    - Agendas
    - Tâches
- De discuter :
    - Forums
    - Messagerie instantanée

[31] Groupware, en anglais.
[32] http://www.agora-project.net/

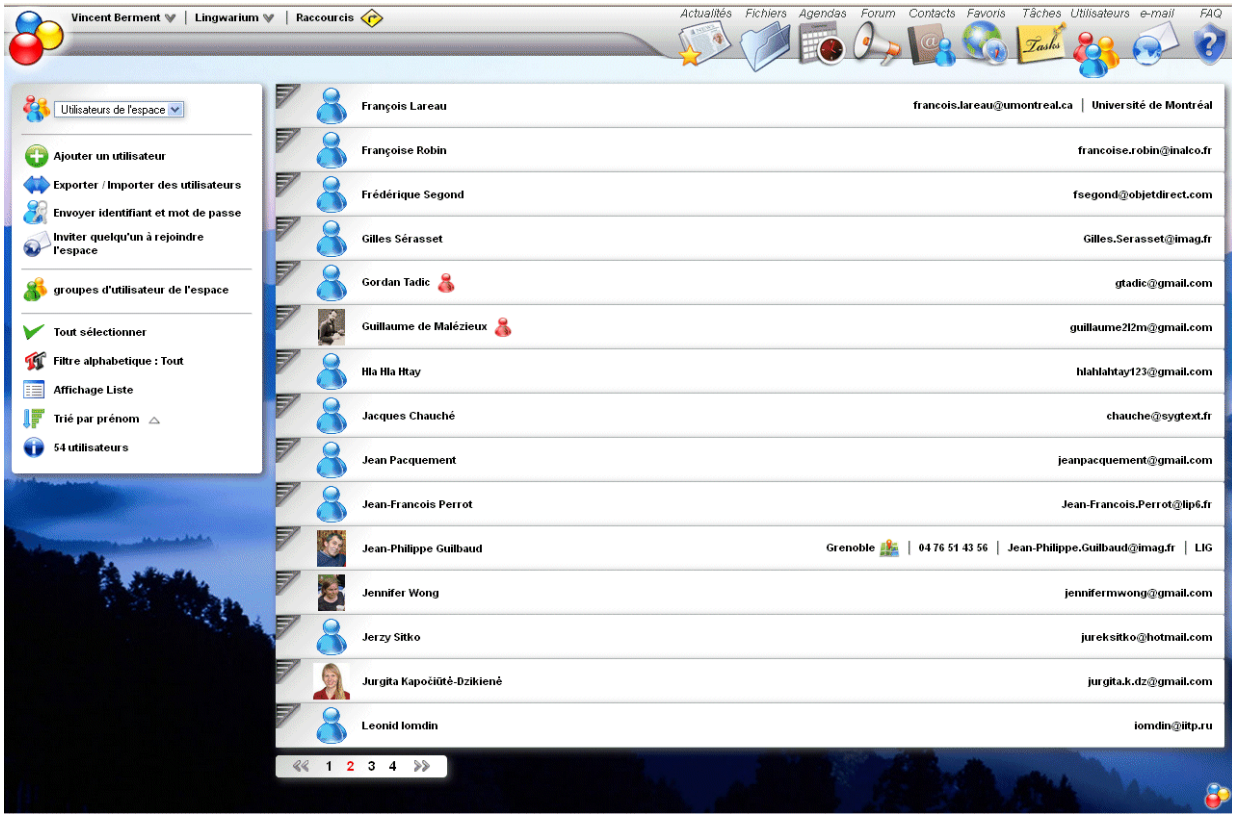


Figure 11 Accès à la liste des membres sur Lingwarium

Lingwarium est d'abord un lieu où on peut trouver de très nombreux documents sur Ariane, ces documents étant souvent introuvables sur Internet (articles, thèses, rapports…).

Lingwarium est aussi un lieu facilitant les échanges entre ses membres, dont la liste est sur le site et accessible à tous. À sa création, Lingwarium était constitué de quelques utilisateurs ayant précédemment échangé à propos d'Héloïse. Les travaux que nous avons réalisés nous ont menés à de nombreux échanges et rencontres, et le nombre d'utilisateurs atteint aujourd'hui 53 personnes. Chercheurs ou étudiants pour la plupart, ils viennent de lieux aussi différents que la Russie, le Japon, la Malaisie, l'Angleterre, la Thaïlande, le Pérou, etc.

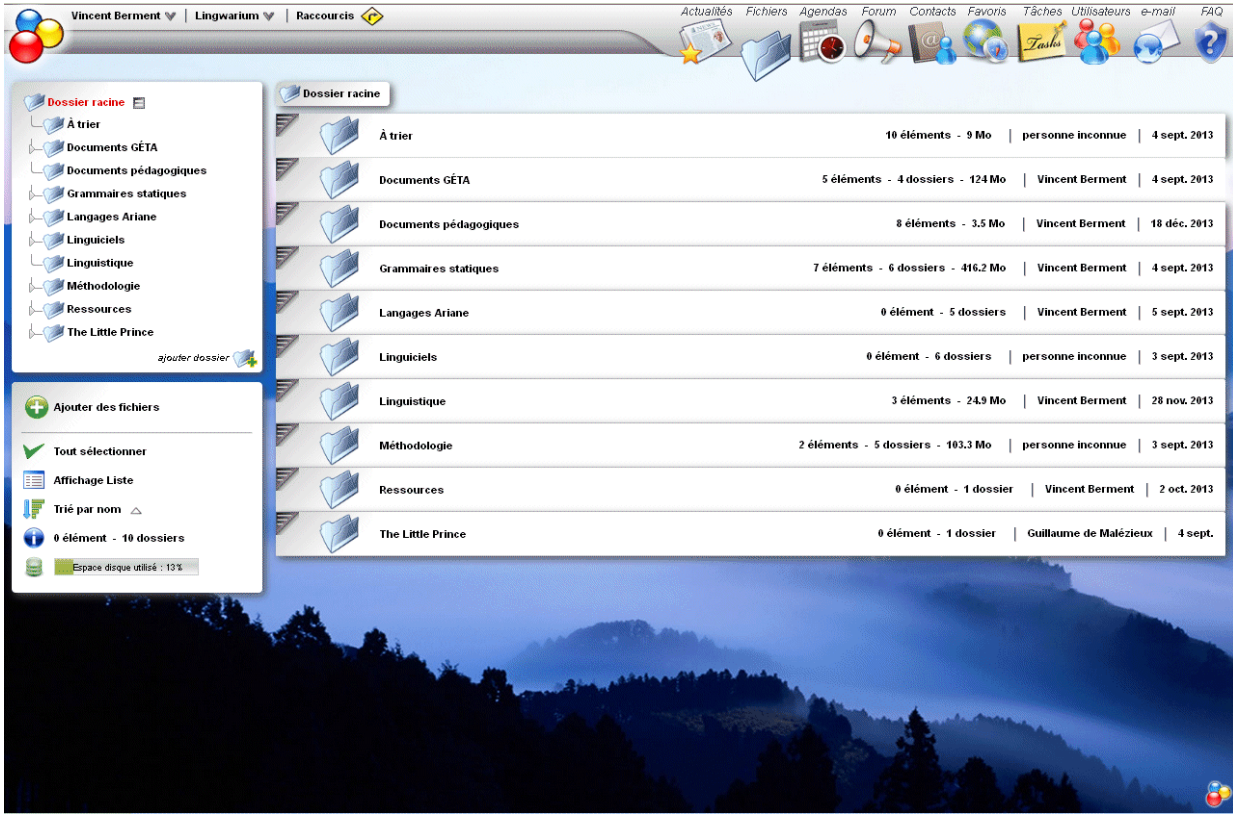


Figure 12 Accès aux documents « Ariane » sur Lingwarium

4.2 Création de « guides de démarrage »

En plus des documents universitaires, souvent difficiles d'accès pour des débutants, nous avons entrepris de mettre à la disposition des candidats au développement de systèmes des « guides de démarrage » en nous basant sur l'expérience que nous avons acquis sur le khmer. Ce travail est actuellement sous forme de notes qui seront intégrées dans la FAQ de Lingwarium quand elles seront prêtes. Le premier guide sur lequel nous avons travaillé est destiné au développement d'une analyse morphologique, la première phase à réaliser pour développer une analyse, comme nous le faisons pour le Petit Prince. Là, nous nous sommes aperçus qu'il fallait distinguer trois cas selon que la langue a une morphologie :

- Isolante,
- Flexionnelle,
- Agglutinante,

À quoi il faut ajouter un guide supplémentaire dans le cas des langues dont le système d'écriture est non segmenté.

Pour les langues isolantes, la méthode est celle que nous avons présenté au chapitre 3.3.2 pour le khmer, une simple liste de mots (qui sont donc des lemmes) accompagnée de leur classe morphosyntaxique.

Pour les langues flexionnelles, on peut gérer la morphologie, soit directement pour chaque

forme d'un même lemme (une entrée de dictionnaire par lemme fléchi), soit complètement à partir d'un paradigme associé à chaque lemme, avec description des paradigmes. C'est cette dernière option que nous avons retenue. La procédure, un peu plus complexe, est détaillée ci-dessous.

- Définir les paramètres morphologiques déterminant les flexions à appliquer, ainsi que les valeurs possibles pour ces paramètres :
  - ⇨ Exemple pour les adjectifs du français : genre(M,S), nombre(S,P)
- Lister les paradigmes flexionnels, a priori séparément pour chaque catégorie morphosyntaxique (noms, adjectifs, verbes...). Pour chaque paradigme :
  - ⇨ lui donner un nom
  - ⇨ indiquer un exemple
  - ⇨ lister les différentes désinences en fonction des paramètres morphologiques
  - ⇨ indiquer le nombre de caractères à ôter au lemme pour obtenir le radical

  Exemple pour les adjectifs du français :

  - ⇨ ParadigmeAdjectif1 ("froid",0) = MS(-), FS(e), MP(s), FP(es).
  - ⇨ ParadigmeAdjectif2 ("rapide",0) = MS(-), FS(-), MP(s), FP(s).
  - ⇨ ParadigmeAdjectif3 ("jumeau",2) = MS(au), FS(lle), MP(aux), FP(lles).
  - ⇨ ParadigmeAdjectif4...

  NB : Il peut y avoir plusieurs formes pour certains paradigmes (ex. : scénario/scénarii)
- Créer le lexique avec ses informations morphologiques en constituant la liste de lemmes et en associant à chaque lemme :
  - ⇨ son paradigme
  - ⇨ les informations lexicales nécessaires au système (ex. : catégorie morphosyntaxique, genre, nombre, syntaxe, sémantique...)
- NB : Il peut y avoir plusieurs paradigmes pour une forme donnée (ex. : « manger » peut être nom ou verbe), ce qui fera deux lignes dans le lexique.

Pour les langues agglutinantes, nous avons eu la chance d'avoir Maximiliano Duran qui s'est joint au projet « Petit Prince », ce qui a permis d'investiguer ce type de morphologie avec le quéchua. À partir de son analyse manuelle d'une partie du premier chapitre du Petit Prince, Vincent Berment a construit un tableau dont voici les premières lignes et colonnes, la phrase en quéchua étant : « Suqta watayuq kaspay huk librupi qawarqani suma sumaq qillqata... ».

Mots	Racine	Traduction	yuq	spa	y	pi	rqa	ni	ta
Suqta	suqta	six							
watayuq	wata	année	x						
kaspay	ka	être / avoir		x	x				
huk	huk	un (chiffre)							
librupi	libru	livre				x			
qawarqani	qawa	voir					x	x	
suma sumaq	sumaq	magnifique							
qillqata	qillqa	dessin							x

Tableau 14 Langues agglutiantes

Dans les trois cas, une « moulinette » de type DB2ATEF a permis de générer les dictionnaires ATEF correspondants de manière automatique. Les langues suivantes ont été ainsi traitées (par Vincent Berment) :

- Isolantes : birman, khmer, lao, tibétain,
- Flexionnelles : anglais, français, lituanien,
- Agglutinantes : quéchua.

5 Conclusion

Nos travaux nous ont permis de dessiner les contours d'une méthodologie pour la production de système de TA pour une langue peu dotée. En nous essayant à cet exercice de construction de maquette pour le khmer, nous nous sommes confrontés à des obstacles que nous avons parfois sous-estimés.

Grâce à ces difficultés, nous sommes maintenant en mesure de déterminer des axes de développement en fonction des besoins d'une langue. Dans cette optique, il serait utile de proposer une campagne d'évaluation en ligne d'indice-σ à destination des participants au projet afin des les guider vers une consolidation de leurs ressources avant de pouvoir se consacrer à l'analyse du *Petit Prince.*

Le nombre toujours croissant de personnes souhaitant se joindre à notre projet nous conforte dans l'idée que nous avons fait les bons choix quant au corpus et au système Héloïse. Maintenir leur enthousiasme tiendra certainement dans l'obtention de résultats. Pour cela il serait intéressant de déterminer des paliers d'avancement, afin de donner des objectifs réalisables, et d'éviter d'en faire une quête du Graal.

Travailler en synergie avec d'autres langues est passionnant et motivant. Lors des phases monolingues tout d'abord : s'ouvrir aux travaux et descriptions linguistiques d'autres langues plus ou moins proches de sa langue de travail permet un enrichissement mutuel inestimable. La phase de transfert, bilingue avec la possibilité d'un pivot UNL pourrait être vue comme une phase multilingue. Plutôt que de travailler sur un couple de langues, chacun apporterait sa pierre à l'édifice pour la construction d'une route plurilingue. Ainsi, une adaptation de GDML permettrait d'éditer non plus un dictionnaire de transfert d'une langue source vers une seule langue cible, mais vers une multitude de langues cibles.

Enfin, la diversité des connaissances nécessaires à une telle démarche et la répartition géographique des nombreux participants nous laisse à penser qu'un MOOC serait un support de formation idéal (à l'image des formations au système Apertium) dont la mise en œuvre pourrait faire l'objet d'une thèse.

Bibliographie

Antelme, Michel Rethy et Bru-Nut, Hélène Suppya : *Dictionnaire français-khmer,* L'Asiathèque, 2013.

Berment, Vincent : *Méthodes pour informatiser des langues et des groupes de langues « peu dotés »*. Thèse de doctorat, Université Joseph Fourier, Grenoble, 18 mai 2004.

Berment, Vincent : *Some thoughts on how to adress commercially unprofitable languages and language pairs,* Invited talk, WSSANLP 2014, Dublin August 23rd 2014

Boitet, Christian : *Corpus pour la TA : types, tailles et problèmes associés, selon leur usage et le type de système.* Revue française de linguistique appliquée, XII, pp. 25-38., Janvier 2007.

Boitet, Christian : . *Bernard VAUQUOIS' contribution to the theory and practice of building MT systems : a historical perspective,* 1988.

Boitet, Christian : *Les architectures linguistiques et computationnelles en traduction automatique sont indépendantes*, TALN 2008, 9-13 juin 2008.

Boitet, Christian : *Les logiciels traduiront 600 langues dans 10 ans.* Les Dossiers de la Recherche, n° 4, Dossier sur la traduction automatique. 2013, p. 97, juin-juillet 2013

Chappuy, Sylviane : *Le développement des systèmes de traduction.* 2013

Delavennat, Estelle (2013) Comparaison des systèmes de décoration des linguiciels traitant les langues FRA, ENG, ALD, RUS.

Hutchins, John.: *Compendium of Translation Software.* Édition 16, IAMT, 19 mars 2010

Khin, Sok : *La langue khmère et l'informatique,* 1983.

Khin, Sok : *La grammaire du khmer moderne,* édition You-Feng, 1999.

Ladmiral Jean-René. « Le traducteur et l'ordinateur ». In : Langages, 28e année, n°116, 1994. pp.5-19.

Lafourcade, Mathieu : *Génie Logiciel pour le Génie Linguiciel*. Thèse de doctorat, Université Joseph Fourier, Grenoble, 1er décembre 1994.

Léon, Jacqueline : *Le CNRS et les débuts de la traduction automatique en France. 2002*.

Malézieux *et al.* : RBMT as an alternative to SMT for under-resourced languages, WSSANLP 2014, Dublin 2014.

Toshiya Suzuki *et al.*, « Encodings in Legacy Khmer TrueType Fonts » Investigation and Propose of Auto-Detection Algorithm, *Document numérique*, 2006/3 Vol.9, p45-68.

Annexes

Annexe 1 : Langues disponibles dans Google Traduction classées par nombre de locuteurs

Language	L1 speaker	L1 speaker	L2 speaker
Maori	148,660	148660	
Icelandic	243,840	243840	
Irish	276,310	276310	1,000,000
Maltese	522,000	522000	
Welsh	536,890	536890	
Basque	545,872	545872	
Estonian	1,165,400	1165400	
Macedonian	1,407,810	1407810	
Yiddish	1,510,430	1510430	
Latvian	1,752,260	1752260	
Slovenian	2,085,160	2085160	
Bosnian	2,225,290	2225290	
Galician	2,355,000	2355000	
Lithuanian	3,001,860	3001860	
Lao	3,273,180	3273180	
Belarusian	3,312,610	3312610	
Catalan	4,079,420	4079420	
Georgian	4,237,710	4237710	
Norwegian	4,741,780	4741780	
Slovak	5,187,740	5187740	
Hebrew	5,302,770	5302770	
Finnish	5,392,180	5392180	
Albanian	5,414,300	5414300	
Danish	5,522,490	5522490	
Croatian	5,752,090	5752090	
Mongolian	5,756,590	5756590	
Armenian	5,902,970	5902970	
Afrikaans	7,096,810	7096810	
Hmong	7,708,420	7708420	
Haitian Creole	7,731,240	7731240	
Bulgarian	8,157,770	8157770	
Serbian	8,957,906	8957906	
Swedish	9,197,090	9197090	
Czech	10,619,340	10619340	
Zulu	11,969,100	11969100	
Hungarian	12,606,130	12606130	
Greek	13,432,940	13432940	
Khmer	14,224,500	14224500	
Somali	14,679,300	14679300	
Nepali	15,360,100	15360100	
Swahili	15,457,390	15457390	
Cebuano	15,810,000	15810000	
Malay	15,848,500	15848500	
Igbo	18,000,000	18000000	

Language	L1 speaker	L1 speaker	L2 speaker
Yoruba	19,380,800	19380800	
Thai	20,396,930	20396930	40,000,000
Dutch	21,944,690	21944690	
Indonesian	23,200,480	23200480	
Romanian	23,782,990	23782990	
Azerbaijani	24,237,340	24237340	
Filipino	24,245,200	24245200	45,000,000
Hausa	24,988,000	24988000	
Ukrainian	36,048,890	36048890	
Kannada	37,739,040	37739040	
Polish	38,663,780	38663780	
Gujarati	46,636,510	46636510	
Persian	56,645,100	56645100	
Italian	63,655,047	63655047	
Urdu	63,948,800	63948800	94,000,000
Vietnamese	67,778,030	67778030	
Tamil	68,763,360	68763360	
Turkish	70,805,930	70805930	
Marathi	71,780,660	71780660	
Telugu	74,049,000	74049000	
French	74,980,460	74980460	
Korean	77,166,230	77166230	
German	78,245,280	78245280	
Javanese	84,308,740	84308740	
Punjabi	93,167,770	93167770	
Japanese	122,056,940	122056940	
Russian	167,332,230	167332230	
Bengali	193,261,200	193261200	
Portuguese	203,349,200	203349200	
Hindi	260,333,620	260333620	
Arabic	290,000,000	290000000	
English	335,148,868	335148868	
Spanish	414,170,030	414170030	
Chinese	1,197,294,060	1197294060	
Esperanto			2,000,000
Latin			

Annexe 2 : Langues disponibles dans Google Traduction classées par ordre lexicographique

Language	L1 speaker	L1 speaker	L2 speaker
Afrikaans	7,096,810	7096810	
Albanian	5,414,300	5414300	
Arabic	290,000,000	290000000	
Armenian	5,902,970	5902970	
Azerbaijani	24,237,340	24237340	
Basque	545,872	545872	
Belarusian	3,312,610	3312610	
Bengali	193,261,200	193261200	
Bosnian	2,225,290	2225290	
Bulgarian	8,157,770	8157770	
Catalan	4,079,420	4079420	
Cebuano	15,810,000	15810000	
Chinese	1,197,294,060	1197294060	
Croatian	5,752,090	5752090	
Czech	10,619,340	10619340	
Danish	5,522,490	5522490	
Dutch	21,944,690	21944690	
English	335,148,868	335148868	
Esperanto			2,000,000
Estonian	1,165,400	1165400	
Filipino	24,245,200	24245200	45,000,000
Finnish	5,392,180	5392180	
French	74,980,460	74980460	
Galician	2,355,000	2355000	
Georgian	4,237,710	4237710	
German	78,245,280	78245280	
Greek	13,432,940	13432940	
Gujarati	46,636,510	46636510	
Haitian Creole	7,731,240	7731240	
Hausa	24,988,000	24988000	
Hebrew	5,302,770	5302770	
Hindi	260,333,620	260333620	
Hmong	7,708,420	7708420	
Hungarian	12,606,130	12606130	
Icelandic	243,840	243840	
Igbo	18,000,000	18000000	
Indonesian	23,200,480	23200480	
Irish	276,310	276310	1,000,000
Italian	63,655,047	63655047	
Japanese	122,056,940	122056940	
Javanese	84,308,740	84308740	
Kannada	37,739,040	37739040	
Khmer	14,224,500	14224500	
Korean	77,166,230	77166230	

Language	L1 speaker	L1 speaker	L2 speaker
Lao	3,273,180	3273180	
Latin			
Latvian	1,752,260	1752260	
Lithuanian	3,001,860	3001860	
Macedonian	1,407,810	1407810	
Malay	15,848,500	15848500	
Maltese	522,000	522000	
Maori	148,660	148660	
Marathi	71,780,660	71780660	
Mongolian	5,756,590	5756590	
Nepali	15,360,100	15360100	
Norwegian	4,741,780	4741780	
Persian	56,645,100	56645100	
Polish	38,663,780	38663780	
Portuguese	203,349,200	203349200	
Punjabi	93,167,770	93167770	
Romanian	23,782,990	23782990	
Russian	167,332,230	167332230	
Serbian	8,957,906	8957906	
Slovak	5,187,740	5187740	
Slovenian	2,085,160	2085160	
Somali	14,679,300	14679300	
Spanish	414,170,030	414170030	
Swahili	15,457,390	15457390	
Swedish	9,197,090	9197090	
Tamil	68,763,360	68763360	
Telugu	74,049,000	74049000	
Thai	20,396,930	20396930	40,000,000
Turkish	70,805,930	70805930	
Ukrainian	36,048,890	36048890	
Urdu	63,948,800	63948800	94,000,000
Vietnamese	67,778,030	67778030	
Welsh	536,890	536890	
Yiddish	1,510,430	1510430	
Yoruba	19,380,800	19380800	
Zulu	11,969,100	11969100	

Annexe 3 : Chronologie des langues disponibles dans Google Traduction

oct-07	mai-08	sept-08	janv-09	juin-09	août-09	janv-10	mai-10	sept-10	juin-11	févr-12	sept-12	avr-13	mai-13	déc-13
English	Hindi	Catalan	Albanian	Persian	Afrikaans	Haitian Creole	Armenian	Latin	Bengali	Esperanto	Lao	Khmer	Bosnian	Hausa
German	Bulgarian	Filipino	Estonian		Belarusian		Azerbaijani		Gujarati				Cebuano	Igbo
Spanish	Croatian	Hebrew	Galician		Icelandic		Basque		Kannada				Hmong	Maori
French	Czech	Indonesian	Hungarian		Irish		Georgian		Tamil				Javanese	Mongolian
Portuguese	Danish	Latvian	Maltese		Macedonian		Urdu		Telugu				Marathi	Nepali
Dutch	Finnish	Lithuanian	Thai		Malay									Punjabi
Italien	Greek	Serbian	Turkish		Swahili									Somali
Chinese S	Norwegian	Slovak			Welsh									Yoruba
Chinese T	Polish	Slovene			Yiddish									Zulu
Japanese	Romanian	Ukrainian												
Korean	Swedish	Vietnamese												
Arabic														
Russian														