Projet Mot à mot : élaboration d'un système lexical multilingue par le biais de dictionnaires bilingues

Mathieu Mangeot, Hong Thai Nguyen

Projet Mot à mot : élaboration d'un système lexical multilingue par le biais de dictionnaires bilingues.

Mathieu Mangeot[1,2] et Hong-Thai Nguyen[1]
[1,2]Laboratoire GETALP-LIG 385 rue de la bibliothèque BP 53
F-38041 GRENOBLE CEDEX 9
France
[1]Laboratoire LLS Université de Savoie BP 1104
F-73011 CHAMBÉRY CEDEX
France
Mathieu.Mangeot@imag.fr
Hong-Thai.nguyen@imag.fr

Résumé

Le projet MotAMot d'élaboration d'un système lexical multilingue est ciblé sur des langues d'Asie du sud-est, en particulier le vietnamien et le khmer. La macrostructure est une structure pivot avec un volume pour chaque langue autour d'un volume pivot reliant les sens de mot de chaque langue entre eux. La microstructure est basée sur la lexicographie explicative et combinatoire. Les contributions se feront en ligne, sur la plate-forme Jibiki, par des communautés de bénévoles constituées autour de jeux sérieux lexicaux. Chaque donnée se verra attribuer un niveau de qualité, de même pour chaque contributeur.
Mots-clés : lexicographie multilingue, langues peu dotées, projet contributif, projet MotÀMot

1. Introduction et contexte

1.1. Informatisation des langues peu dotées

Les enjeux économiques liés aux techniques du traitement de l'information sont très importants. Le développement de telles technologies est un atout majeur pour des pays en voie de développement comme le Cambodge et le Laos, ou émergents comme le Vietnam, la Malaisie et la Thaïlande.
Comme l'indique V. Berment dans sa thèse (Berment 2004), « le développement des ordinateurs personnels et celui des réseaux font aujourd'hui de l'informatique un instrument pour écrire et communiquer au même titre que le papier et l'imprimerie l'étaient auparavant. Traitements de texte, courriers électroniques, voire des systèmes plus avancés comme la dictée ou la synthèse vocale sont des outils largement répandus. L'idée s'impose alors qu'aux moyens traditionnels doivent s'ajouter les outils informatiques appropriés sans lesquels les buts visés ne peuvent plus être atteints ». L'informatisation d'une langue occupe ainsi une place essentielle dans ce vaste contexte.
Cependant, parmi les 6000 langues parlées dans le monde, seul un tout petit nombre d'entre elles atteint un « niveau d'informatisation » satisfaisant. Pour évaluer de manière quantitative le degré d'informatisation d'une langue, il propose le protocole suivant : à chaque service ou ressource, un groupe d'utilisateurs représentatifs des locuteurs de la langue attribue un niveau de criticité Ck et une note Nk, la moyenne pondérée des notes — appelée indice — reflétant leur satisfaction globale. Une langue mal ou peu dotée peut ainsi être définie comme une langue dont l'indice n'atteint pas 10/20. A titre d'exemple, on peut présenter dans les deux figures ci-dessous une évaluation du niveau d'informatisation obtenu ainsi pour la langue khmère, parlée au Cambodge, et pour la langue vietnamienne. La figure 1 représente l'évaluation du niveau d'informatisation pour le khmer et la figure 2 pour le vietnamien.

Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte			
Saisie simple	10	16	160
Visualisation / impression	10	14	140
Recherche et remplacement	8	12	48
Sélection du texte	6	12	72
Tri lexicographique	5	0	0
Correction orthographique	2	0	0
Correction grammaticale	0	0	0
Correction stylistique	0	0	0
Traitement de l'oral			
Synthèse vocale	5	0	0
Reconnaissance de la parole	5	0	0
Traduction			
Traduction automatisée	8	4	32
ROC			
Reconnaissance optique de caractères	9	0	0
Ressources			
Dictionnaire bilingue	10	4	40
Dictionnaire d'usage	10	0	0
Total			540 / 1760
Moyenne			6,2 / 20

Figure 1 : niveau d'informatisation du khmer

Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte			
Saisie simple	10	16	160
Visualisation / impression	10	16	160
Recherche et remplacement	8	17	136
Sélection du texte	6	17	102
Tri lexicographique	5	6	30
Correction orthographique	2	6	12
Correction grammaticale	0	0	0
Correction stylistique	0	0	0
Traitement de l'oral			
Synthèse vocale	5	0	0
Reconnaissance de la parole	5	0	0
Traduction			
Traduction automatisée	8	6	48
ROC			
Reconnaissance optique de caractères	9	12	108
Ressources			
Dictionnaire bilingue	10	13	130
Dictionnaire d'usage	10	0	0
Total			886 / 1760
Moyenne			10 / 20

Figure 2 : niveau d'informatisation du vietnamien

De toutes façons, ces coûts sont trop élevés pour un particulier. De ce fait, seules des institutions peuvent l'acquérir. D'autre part, les données fournies à ce prix ne sont utilisables que par certains systèmes de traduction automatique fondés sur des techniques particulières.

Face à ces coûts difficilement gérables, les maisons d'édition finissent par vivre sur leurs acquis et ne proposent principalement que des nouvelles éditions de dictionnaires existants. Rares sont les éditeurs à avoir le courage de se lancer dans la réalisation d'un nouveau dictionnaire bilingue de qualité en partant de zéro.

D'autre part, même dans les dictionnaires les plus complets, on constate quasiment toujours un manque d'informations en particulier concernant les collocations. Les rares ressources qui en tiennent compte ne le font pas de manière systématique.

Malgré l'arrivée d'Internet, il existe à l'heure actuelle peu de ressources lexicales de bonne qualité disponibles gratuitement en ligne. La plupart sont en fait des lexiques bilingues faits par des bénévoles non spécialistes en lexicographie.

La lexicographie multilingue en tant que telle n'en est en fait qu'à ses débuts. En effet, il n'existe pas vraiment de moyen d'imprimer un vrai "dictionnaire multilingue". Il est par contre tout à fait possible de trouver des bases terminologiques multilingues (comme IATE) ou bien, à la rigueur des petits lexiques ou livres de phrases multilingues.

Il n'a par ailleurs pas été encore suffisamment prouvé que la réutilisation d'un dictionnaire d'un couple de langue A➔langue B pour en construire deux autres langue B➔langue C et langue A➔langue C était réellement avantageuse. C'est donc ce à quoi nous voulons nous attacher avec ce présent projet.

2. Objectifs du projet

Avec l'objectif général de participer à l'informatisation des langues peu dotées, ce projet consiste à élaborer un système lexical multilingue en construisant simultanément plusieurs dictionnaires bilingues partageant au moins une langue entre eux. La construction des dictionnaires bilingues se fera en ligne sur un site de type "Papillon" construit sur la plate-forme Jibiki selon une méthodologie de travail collaboratif et bénévole, inspirée du projet Wikipedia.

Les liens bilingues créés lors de la rédaction des articles sont utilisés d'une part pour générer des liens bilingues inverses, et d'autre part pour créer de nouveaux liens interlingues.

1http://www.populationdata.net

Ces dictionnaires seront aussi disponibles sous la forme d'une version multimédia, avec une interface conviviale et ergonomique, dont les résultats seront accessibles par deux média : langue (texte) et parole.
Pour certaines langues (que nous traiterons comme exemples), des modules de synthèse seront rajoutés pour permettre de participer à l'apprentissage de la langue et de donner des exemples sonores de type « mots isolés » ou « livre de courtes phrases » aux apprenants.
Les trois objectifs principaux de ce projet sont donc le lancement d'une dynamique de contribution autour de la construction de chaque dictionnaire bilingue en présence. Le succès de Wikipédia montre que cela est possible, à condition d'avoir des outils simples et faciles à utiliser; le passage à grande échelle d'expériences de laboratoire telles que la base DiCo (Mel'čuk et Polguère 2006) ou le système PARAX (Blanc 1996); et enfin l'élaboration d'un terrain d'expérimentation pour la validation de plusieurs hypothèses formulées dans de précédents travaux :

- Bijectivité des liens bilingues et transitivité des liens interlingues;
- Contribution massive sur le Web;
- Construction d'un système lexical multilingue (Polguère 2006).

3. Avancées dans la construction de ressources en ligne

3.1. Sur l'architecture des ressources multilingues : le projet Papillon

Une solution parfaite, le graal des ressources lexicales, serait une base de données lexicales multilingue à structure pivot, de bonne qualité et large couverture avec des entrées monolingues riches et des liens interlingues, utilisable aussi bien par des humains que par des machines, éditable en ligne et disponible gratuitement. Nous avons lancé en 2000 le projet Papillon[2] de construction d'une base multilingue pour tenter d'avancer dans cette direction.
La macrostructure est constituée d'un volume monolingue pour chaque langue et un volume pivot au centre (voir figure 3). Lorsqu'un nouvel article dans une langue A est ajouté, il doit être relié au volume interlingue. Ces liens sont créés soit en réutilisant des dictionnaires bilingues existants langue A➔langue B, soit en les ajoutant manuellement à partir d'une traduction. Le lien langue A➔langue B devient langue A➔pivot➔langue B. Si l'article langue B est déjà relié à un autre article langue C, alors l'article langue A bénéficiera lui aussi de ces liens.

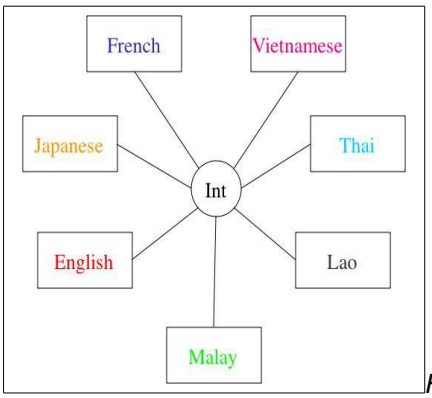

Figure 3 : macrostructure pivot

2 http://www.papillon-dictionary.org

La microstructure des articles monolingues est riche et très détaillée. Elle est basée sur la structure utilisée pour la base lexicale DiCo (Polguère 2000) de l'OLST, Université de Montréal. La méthodologie d'encodage est directement empruntée à la lexicologie explicative et combinatoire, partie de la théorie sens-texte. Cette théorie donne les informations nécessaires pour passer d'un sens à ses réalisations dans une langue donnée. La microstructure des dictionnaires est donc indépendante des langues et peut être théoriquement utilisable par des humains et par des machines.

Chaque article ou unité lexicale est basé sur le sens de mot ou lexie. Il est constitué d'un nom, des propriétés grammaticales, une formule sémantique qui peut être vue comme une définition formelle - dans le cas d'une lexie, prédicative, la formule décrit le prédicat et ses arguments et on trouve aussi le régime qui décrit la réalisation syntaxique des arguments - , puis une liste de fonctions lexico-sémantiques - il y a 56 fonctions de base applicable à toute langue et pouvant se combiner entre elles -, une liste d'exemples et une liste d'expressions idiomatiques.

Les spécifications du projet Papillon sont directement inspirées de ce fameux graal. Mais comme tout projet ambitieux, il ne peut être achevé d'un coup. Avec le temps, le projet Papillon est devenu une sorte de cadre ou méta-projet (Mangeot, 2006) avec plusieurs projets dérivés, chacun correspondant à un aspect particulier de notre but initial. Comme nous le détaillerons par la suite, les aspects outils et systèmes sont couverts par le projet Jibiki et la collecte de données par le projet JeuxDeMots.

3.2. Sur les aspects contributifs : les projets Wikipedia et Wiktionary

L'encyclopédie en ligne contributive Wikipedia[3] a rencontré un gros succès incontestable. On aurait pu s'attendre à une réussite comparable pour son petit frère Wiktionary[4] mais le succès n'est pas encore au rendez-vous (1,5 million d'entrées pour le français et seulement 44 000 pour le japonais). Wiktionaryz[5], qui prétendait parer aux défauts de Wiktionary n'a pas non plus eu l'effet escompté. Wiktionary n'est de toutes façons pas un véritable dictionnaire bilingue même si on trouve quelques liens de traduction (il manque entre autres les indications de contexte de traduction).

Une hypothèse pour expliquer ce problème est celui de la motivation. En effet, lorsqu'une personne contribue à un article de Wikipedia, elle est récompensée par la renommée. Elle sera ensuite reconnue comme un expert dans son domaine. Cela n'est pas possible avec un dictionnaire. Les contributions portent sur des petites parties d'informations très ciblées et sont de ce fait anonymes. D'autre part, il y a un aspect technique lié à la structure. Un article d'encyclopédie a une structure plus ou moins libre tandis qu'une entrée de dictionnaire doit suivre une structure très précise (mot-vedette, informations grammaticales, blocs sémantiques, bloc de traduction, blocs d'exemples, etc.). Il n'est donc pas possible de réutiliser une plate-forme wiki pour construire un dictionnaire avec une structure bien définie.

Une fois acceptée l'idée que rédiger des entrées de dictionnaire n'est pas aussi plaisant que travailler sur un article de Wikipédia, il faut trouver des solutions pour motiver une communauté de bénévoles à contribuer à un dictionnaire. Les jeux sérieux lexicaux constituent une première piste. Il faut aussi mettre en valeur les contributeurs à travers par exemple un tableau des meilleurs contributeurs du mois. Et enfin, l'exploitation de réseaux communautaires tels que FaceBook devraient aussi apporter de l'eau au moulin.

---

3 http://fr.wikipedia.org

4 http://fr.wiktionary.org

5 http://www.wiktionaryz.com

3.3. Sur la collecte de données via des jeux sérieux : le projet JeuxDeMots

JeuxDeMots (Lafourcade & Joubert 2008) est une tentative de réponse au problème précédent. Ce projet a pour but de construire un réseau lexical riche et évolutif, qui peut être comparé à un certain degré à la fameuse base WordNet (Miller *et al.* 1990). Le principe est le suivant : une partie nécessite 2 joueurs. Lorsqu'un joueur A débute une partie, une consigne concernant un type de compétence (synonymes, contraires, domaines) est affichée, ainsi qu'un mot M tiré aléatoirement dans une base de mots. Le joueur A a alors un temps limité pour répondre en donnant des propositions répondant, selon lui, à la consigne appliquée au mot M. Ce même mot, avec cette même consigne, est proposé à un autre joueur B ; le processus est identique.
Les deux demi-parties, celle du joueur A et celle du joueur B, ne sont pas simultanées, mais asynchrones. Pour toute réponse commune dans les propositions de A et B, ces deux joueurs gagnent un certain nombre de points. La structure du réseau lexical que nous cherchons ainsi à obtenir s'appuie sur les notions de nœuds et de relations entre nœuds pour construire un réseau du type de (Polguère, 2006). Chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies et les relations entre nœuds traduisent des fonctions lexicales. La première version du jeu pour le français a été lancée en juillet 2007. Il existe aussi des versions anglaises, arabes, japonaises, vietnamiennes, et thaï. Elles sont disponibles sur le Web[6].

3.4. Sur les aspects techniques : la plate-forme Jibiki

Jibiki (Mangeot et al., 2004) est une plate-forme générique en ligne pour manipuler des ressources lexicales avec gestion d'utilisateurs et groupes, consultation de ressources hétérogènes et édition générique d'articles de dictionnaires. C'est un site Web communautaire développé au départ pour le projet Papillon. La plate-forme est programmée entièrement en Java, basée sur l'environnement "Enhydra". Toutes les données sont stockées au format XML dans une base de données (Postgres). Ce site Web propose principalement deux services : une interface unifiée permettant d'accéder simultanément à de nombreuses ressources hétérogènes (monolingues, dictionnaires bilingues. bases multilingues, etc.) et une interface d'édition spécifique pour contribuer directement aux dictionnaires disponibles sur la plate-forme.
L'éditeur est basé sur un modèle d'interface HTML instancié avec l'article que l'on veut éditer. Le modèle peut être généré automatiquement depuis une description de la structure de l'entrée à l'aide d'un schéma XML. Il peut être modifié ensuite pour améliorer le rendu à l'écran. La seule information nécessaire à l'édition d'un article de dictionnaire est donc le schéma XML représentant la structure de cette entrée. De plus, il est possible d'éditer n'importe quel type de dictionnaire s'il est encodé en XML.
Plusieurs projets de construction de ressources lexicales ont utilisé ou utilisent toujours cette plate-forme avec succès. C'est le cas par exemple du projet GDEF de dictionnaire bilingue estonien-français[7]. Le code de cette plate-forme est disponible gratuitement en source ouverte en téléchargement depuis la forge du laboratoire LIG[8].

6 http://jeuxdemots.liglab.fr/

7 http://www.estfra.ee/

8 http://jibiki.ligforge.imag.fr/

4. Description de la ressource à construire

4.1. Microstructure des articles basée sur la théorie sens-texte

La microstructure des articles composant les volumes monolingues est une simplification de celle du projet Papillon. Chaque article est cette fois basé sur le vocable. Un vocable étant soit un regroupement de lexies (sens de mot), soit une locution.

Pour faire face aux niveaux de compétences différents selon les contributeurs, l'interface d'édition pourra s'adapter et afficher une granularité d'information adaptée. Par exemple, un contributeur débutant sera invité à renseigner une simple glose pour caractériser une lexie, alors qu'un linguiste expert devra décrire une formule sémantique complète. De même, certains contributeurs seulement auront accès à la liste des fonctions lexicales à remplir. La figure 4 décrit l'article correspondant au vocable "abandonner". Celui-ci n'a pour l'instant qu'un seul sens, caractérisé par une glose (laisser tomber un truc) et une formule sémantique (action sur un objet : humain ou animal X ~ entité Y).

Figure 4 : article "abandonner"

4.2. Macrostructure pivot via des interfaces bilingues

La macrostructure est également tirée du projet Papillon avec un volume monolingue pour chaque langue et un volume pivot au centre. Cependant, afin de ne pas dérouter les utilisateurs, ceux-ci contribueront via une interface présentant une vue classique de dictionnaire bilingue. Chaque lien bilingue langue A➔langue B ajouté via cette interface sera en fait traduit en arrière plan par la création de deux liens interlingues ainsi que d'une axie représentant le lien de traduction d'origine pour obtenir finalement : langue A➔axie pivot➔langue B (voir figure 5).

4.3. Établissement des liens bilingues et interlingues

Lorsqu'un contributeur veut ajouter un lien de traduction entre un vocable Va de langue A et un vocable Vb

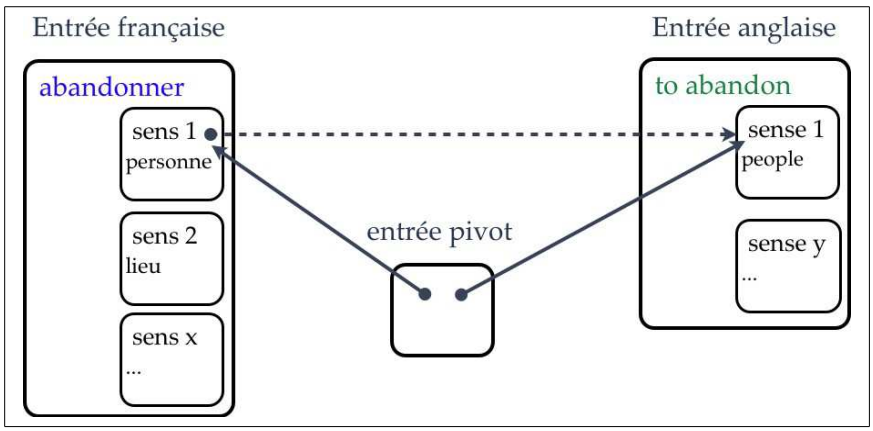

Figure 5 : établissement de liens bilingues

de langue B, il peut établir ce lien à différents niveaux.

La solution idéale est de relier un sens de mot Sa du vocable Va à un autre sens de mot Sb du vocable Vb. Dans ce cas, le lien est bijectif et Sb est donc aussi relié à Sa.

Si le vocable Vb n'a pas encore de sens de mots précis ou si le contributeur n'arrive pas à choisir de sens de mot, il peut relier Sa directement au vocable Vb. Dans ce cas, un nouveau sens de mot Sb' est créé avec un niveau de qualité brouillon et le lien ainsi que les sens de mots sont marqués comme étant à raffiner.

Dans le cas de la récupération de données existantes, il est bien souvent impossible de rattacher une information à un sens de mot précis. Dans ce cas, on ajoute à la fin du vocable Va l'information selon laquelle un des sens de mots de Va peut être relié à un sens de mot de Vb, mais cette information ne sera pas ajoutée à Vb. Elle sera bien sûr marquée comme étant à raffiner d'urgence !

Grâce à la macrostructure pivot, si deux liens langue A➔langue B et langue B➔langue C existent, alors il sera automatiquement créé un lien langue A➔langue C dont le niveau sera de qualité brouillon et marqué comme à réviser (voir figure 6).

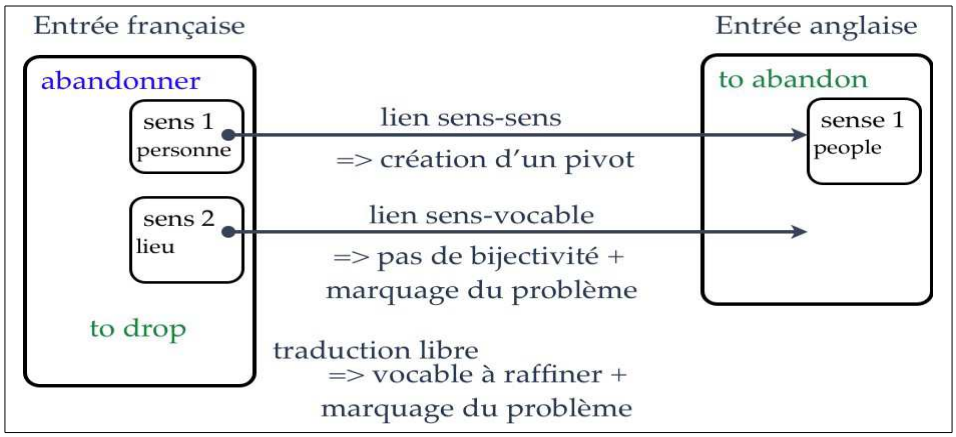

Figure 6 : les différents types de lien

4.4. Niveaux de qualité des données et des contributeurs

Chaque partie d'information de chaque article se verra attribuer un niveau de qualité. Les niveaux s'échelonnent de 1 étoile pour un brouillon (données récupérées dont la qualité n'est pas connue) à 5 étoiles, qualité certifiée par un expert (par exemple, un lien de traduction validé par un traducteur assermenté). La figure 4 montre par exemple que l'article du vocable "abandonner" est de niveau trois étoiles.

De la même manière, les contributeurs se verront assigner un niveau de compétence (1 à 5 étoiles également). 1 étoile étant le niveau d'un débutant inconnu dans la communauté et 5 étoiles étant le niveau d'un expert reconnu.

Ensuite, lorsqu'un contributeur de niveau 3 révise un article de niveau 2, l'article monte automatiquement au niveau 3. De même, si le travail d'un contributeur est systématiquement validé sans corrections par d'autres contributeurs de niveau supérieur, celui-ci peut passer automatiquement au niveau supérieur au bout d'un certain seuil (par exemple 10 contributions).

Pour aller plus loin, nous envisageons d'analyser le travail des contributeurs. Si une personne contribue massivement par exemple sur un domaine particulier, le système pourra de manière automatique lui envoyer régulièrement des propositions de contribution dans son domaine.

5. Méthodologie d'élaboration des données

La méthodologie d'élaboration des données est constituée de trois étapes principales : la récupération de données existantes, la collecte de données via des jeux sérieux et enfin la contribution en ligne sur le Web.

5.1. Récupération des données existantes.

Afin d'encourager les contributions, il est préférable de proposer un squelette de dictionnaire à modifier plutôt qu'un dictionnaire vide (syndrome de la page blanche). Pour chaque langue en présence, une liste de mots de cette langue sera récupérée pour créer une première liste d'articles. Il sera toujours possible de créer un nouvel article, mais les créations seront soumises à vérification.

Selon les sous-projets et les langues en présence, plusieurs dictionnaires peuvent être utilisés :

- Les dictionnaires des projets Fe* (français - anglais + autre langue) : FeM (malais) (Gut et al. 1996), FeT (thaï), FeV (vietnamien) ;
- La base DiCo pour le français ;
- Le dictionnaire bilingue français-vietnamien VietDict.

La figure 7 montre un exemple d'article du dictionnaire Fe* avec des traductions en anglais (vert), vietnamien (marron clair), malais (vert clair) et thai (violet).

Le nombre d'étoiles initial des articles générés à partir de ces données est fixé en fonction de la qualité du dictionnaire et de la granularité des données récupérées.

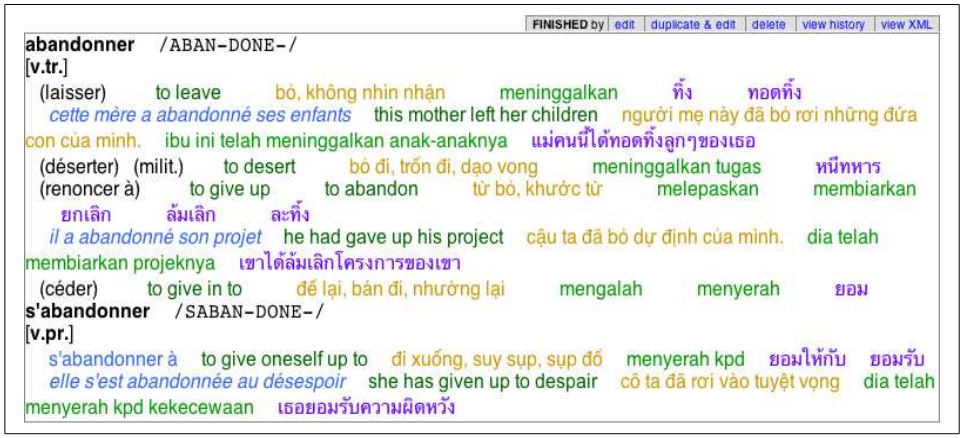

*Figure 7 : article "abandonner" du dictionnaire Fe**

5.1.1 Traitement spécial pour le khmer

Pour la langue khmère, il existe à l'heure actuelle un dictionnaire français-khmer informatisé (Richer *et al.*2007) qui, commencé à la fin des années 90, a été achevé en 2006 par un petit groupe de chercheurs informaticiens réunis dans l'association « Pays Perdu » créée par Denis Richer, ethnolinguiste français établi à Siem Reap (Cambodge). Cette première version du dictionnaire a été publiée au printemps 2007 et comporte 20 000 articles. Le dictionnaire est au format Word. La partie khmer consiste en une simple transcription phonétique khmer de la traduction du mot-vedette écrite dans une police spéciale API (SILSophia IPA93) créée par le Summer Institute of Linguistics. La figure 8 montre à gauche un extrait de ce dictionnaire et à droite ce que l'on voudrait obtenir (écriture khmer au lieu de la transcription phonétique).

French	Khmer
jarret	kɔ̄nlēak-cə̄ŋ
jars	kɔŋān-chmə̄l
jasmin	mlih
jauge	□camnoh-rɔ̄ŋvusl
— (techn.)	māet-stuəŋ

jarret	កន្លាក់-ជើង
jars	ក្រាន-ឈ្មោល
jasmin	ម្លិះ
jauge	កំនត់-ដង្ហាល់

Figure 8 : dictionnaire français-khmer en phonétique et en alphabet khmer

5.2. Collecte de données via les jeux sérieux

Il s'agira ici de lancer un JeuxDeMots pour chaque langue du projet. Le JeuxDeMots français est déjà lancé depuis maintenant 2 ans. Le JeuxDeMots vietnamien est déjà traduit et vient d'être lancé officiellement[9]. Le JeuxDeMots khmer est en cours de traduction. Nous espérons trouver un succès comparable au JeuxDeMots français. Pour aller plus loin, il faudrait réfléchir à des jeux permettant la collecte de données bilingues.

5.3. Contribution en ligne sur le Web

Les données récupérées et collectées sont ensuite fusionnées pour donner naissance à un squelette de dictionnaire. Celui-ci est ensuite mis en ligne pour correction et enrichissement.

6. Conclusion

Le projet est déjà relativement bien avancé. La plupart des aspects techniques concernant la plate-forme et les jeux sérieux en ligne sont réglés. Il reste maintenant à récupérer et convertir les ressources existantes. L'enjeu majeur du projet se trouve en fait dans notre capacité à motiver les communautés de contributeurs. Nous espérons que notre expérience et l'attrait d'un tel projet nous permettra d'avancer sur ces aspects sociologiques.

Les retombées d'un tel projet seront nombreuses et contribueront à relancer l'intérêt de la francophonie dans les pays de l'Asie du sud-est. Les données produites pourront être utilisées par les apprenants du français dans ces pays, ou par les francophones souhaitant s'initier à une langue d'Asie du sud-est. Les dictionnaires pourront être consultés directement en ligne ou sur PDA par les touristes ou les hommes d'affaires.

Les communautés de contributeurs permettront de lancer une dynamique de coopération autour d'un but humaniste voire humanitaire. De plus, cela peut susciter l'intérêt pour l'apprentissage des langues traitées dans le projet et l'élargissement à d'autres langues de la région.

7. Remerciements

Le projet MotAMot est financé en partie par une Action de Recherche en Réseau du réseau Lexicologie, Terminologie, Traduction (ARR-LTT) de l'Agence Universitaire de la Francophonie (AUF).

8. Bibliographie

Berment (V.), 2004 : *Méthodes pour informatiser des langues et des groupes de langues peu dotées.* Thèse de nouveau doctorat, Université Joseph Fourier, Grenoble, France.

9 http://jeuxdemots.liglab.fr/vie/

Blanc (E.), 1996 : Une maquette de base lexicale multilingue à pivot lexical : PARAX. *Lexicomatique et Dictionnairique*, Actes du colloque LTT, Lyon, septembre 1995, ed. AUPELF-UREF, Montréal, Canada, pp. 43-58.

Caelen-Haumont (G.), 2009 : *Prosodie et sens : une approche expérimentale*, édition l'Harmattan-Marges Linguistiques.

Gut (Y.) , Megat Ramli (P.R.), Zaharin (Y.), Chuah Choy (K.), Samat (S.A.), Boitet (C.), Nédobejkine (N.), Mathieu Lafourcade (M.) et al., 1996 : *Kamus Perancis-Melayu Dewan, dictionnaire français-malais.* Dewan Bahasa Dan Pustaka, Kuala Lumpur, 667 p.

Lafourcade (M.) et Joubert (A.), 2008 : JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, France, pp. 657–666.

Mangeot (M.), 2001 : *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue.* Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier (Grenoble 1), septembre 2001, 280 p.

Mangeot (M.), Sérasset (G.) et Lafourcade (M.), 2003 : Construction collaborative de données lexicales multilingues, le projet Papillon. *Revue TAL, édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ?* (Electronic dictionaries: for humans, machines or both?) Ed. Michael Zock & John Carroll, Vol. 44:2/2003, pp. 151-176.

Mel'čuk (I.) et Alain Polguère (A.), 2006 Dérivations sémantiques et collocations dans le DiCo/LAF. *Langue française, numéro spécial sur la collocation « Collocations, corpus, dictionnaires »*, sous la direction de P. Blumenthal et F. J. Hausmann, 150, juin 2006, 66-83.

Polguère (A.), 2006 : Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives. Proceedings of the *Workshop on Multilingual Language Resources and Interoperability* (COLING/ACL 2006), Sydney, 50-59.

Richer (D.), T. Keo (T.), Vania (I.), 2007 : *Dictionnaire Français-Khmer (en phonétique), D.R. Edition, ISBN-13:890-0-9.*

Sam (S.), 2006 : *Analyse de la langue khmère en vue de la synthèse de la parole*, M2R (DEA) de l'Université Joseph Fourier, Grenoble, France.

Sam (S.), Eric Castelli (E.), L. Protin (L.), Pham Thi (N.Y.), 2007 : *Traitement automatique de la langue khmère : rapport scientifique final* - Projet AUF TALK ITC - MICA - Cambodge, mai 2007.

Seng (S.), 2007 : *Collecte et traitement de données en vue de la reconnaissance automatique de la parole en langue khmère*, Rjcp 2007, Paris, France, à paraître.

Sérasset (S.), Brunet-Manquat (F.) et al., 2006 : Multilingual legal terminology on the Jibiki platform: the LexALP project. Coling-ACL 2006: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney.

Sérasset (G.), 1994 : *SUBLIM : un Système Universel de Bases Lexicales Multilingues, et NADIA, sa spécialisation aux bases lexicales interlingues par acceptions.* Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier (Grenoble 1), 194 p.