# MOTÀMOT PROJECT: BUILDING A MULTILINGUAL LEXICAL SYSTEM VIA BILINGUAL DICTIONARIES

*Mathieu MANGEOT, Sereysethy TOUCH*

Laboratoire GETALP-LIG 385 rue de la bibliothèque BP 53,
F-38041 GRENOBLE CEDEX 9, France

## ABSTRACT

The MotAMot project aims to develop of a multilingual lexical network focused on languages of Southeast Asia and especially Vietnamese and Khmer. The macrostructure is a pivot structure with a monolingual volume for every language and a pivot one connecting each word sense of each monolingual volume. The microstructure is based on the explanatory and combinatorial lexicography. Contributions will be made online on the Jibiki platform by a community of volunteers constituted around serious games lexical. Each entry will be given a level of quality, as well as for each contributor.

***Index Terms*---** multilingual lexicography, under resourced languages, contributive project, MotÀMot project.

## 1. INTRODUCTION

Economic issues related to technical processing Information is very important. The development of such technology is a key asset for developing countries such as Cambodia and Laos, or emerging ones such as Vietnam, Malaysia and Thailand.

As indicated by V. Berment in his thesis [1], "Development of personal computers and networks make are now necessary to write and communicate in the same way as paper and printing were before. Word processing, emails, or even more advanced systems such as dictation software or speech synthesis are now widespread tools. It is then necessary to consider that computer programs must be added to the traditional tools otherwise the targeted goals can not be achieved any more. Computerization of a language has and an essential place in this broad context."

However, among the 6,000 languages spoken around the world, only a handful of them reach a satisfactory "Level of computerization". To quantitatively assess the degree of computerization of a language, V. Berment proposes the following protocol: to each service or resource, a group of users representative of the language speakers assign a level of criticality Ck and a score Nk. The average weighted scores - called index - reflects their overall satisfaction. A poorly equipped language can be defined as a language whose index is less than 10/20. For example, the Khmer language, spoken in Cambodia obtains 6.5/20, and the Vietnamese language 10/20.

This is mostly because the services related to the treatment of oral (or speech technologies, ie speech synthesis and word recognition) are not yet available for these two languages. It is also the case for a majority of languages in the world some of which are spoken by several tens of million speakers (for example, Bengali: 189 million, Tamil: 63 million), including within Europe countries (Lithuanian, Latvian, Polish ...)!

## 2. PRESENT STATE OF BILINGUAL LEXICOGRAPHY

The main difficulty at present for bilingual lexicography is the prohibitive construction cost for large amounts of data. For example, the Electronic Dictionary Research project (EDR) whose aim was to build a Japanese-English dictionary required more than 1,200 men-years of work. Its selling price of approximately € 84,000 is far below the actual costs of construction, costs that will probably never be reached.

Anyway, these costs are too high for an individual. Thus, only institutions can acquire such a resource. Moreover, data provided at this price is used by some machine translation systems based on specific techniques.

Faced with these costs difficult to manage, publishing houses end up living on their laurels and do mainly propose new editions of existing dictionaries. Few publishers have the courage to embark on the implementation of a new high quality bilingual dictionary from scratch.

Moreover, even in the most complete dictionaries, there is almost always a lack of information especially on collocations. The few resources that take them into account do it not systematically.

Despite the advent of the Internet, there are currently few lexical resources available freely online in a good quality. Most are in fact small bilingual lexicons made by volunteers not specialists in lexicography.

Multilingual lexicography as such is still in its infancy. Indeed, there is no really a way to print a true "multilingual dictionary". However, it is possible to find multilingual terminological databases (like Iate) or of small lexicons or multilingual phrases books.

It has also not been sufficiently proven that reusing a dictionary of a language couple A➔B in order to build two other language B➔language C and language A ➔language C was really advantageous. So this is what we would like to tackle in this project.

## 3. GOALS OF THE PROJECT

With the overall objective to participate in the computerization of under-resourced languages, this project aims to develop a lexical system in multiple languages by simultaneously building several bilingual dictionaries sharing at least one language between them. The construction of the bilingual dictionaries will be online on a Papillon-like site built on the Jibiki platform with a collaborative and volunteer based work like Wikipedia.

The bilingual links created during the edition of the entries are used first to generate bilingual reverse links, and second to create new interlingual links.

The three main objectives of this project are the launching of a new contribution dynamic around the construction of each bilingual dictionary involved - the success of Wikipedia shows that it is possible, provided that you have simple and easy to use tools -; moving laboratory experiments such as the DiCo database [2] or the PARAX system[3] to a large-scale and finally developing a testbed for validation of several assumptions made in previous work:

- Bijectivity of bilingual links and transitivity of interlingual ones;
- Massive contribution on the Web;
- Construction of a multilingual lexical system [4].

## 4. PROGRESS IN THE CONSTRUCTION OF ONLINE RESOURCES

### 4.1. On the architecture of multilingual resources: the Papillon project

A perfect solution, the holy grail of lexical resources, would be a multilingual lexical database with a pivot structure, of good quality and wide coverage with rich monolingual entries and interlingual links, used both by humans and machines, editable online and freely available. We launched in 2000 the Papillon project[1] in order to advance in this direction.

The macrostructure consists of a monolingual volume for each language and a pivot volume in the center (see Figure 4.1). When a new entry in a language A is added, it must be connected to the interlingual volume. These links are created either by reusing existing bilingual dictionaries of language A ➔language B, either by adding them manually from a translation. Link language A ➔language B becomes language A ➔pivot ➔language B. If the language B entry is already connected to another entry of language C, then language A entry also benefit from these links.

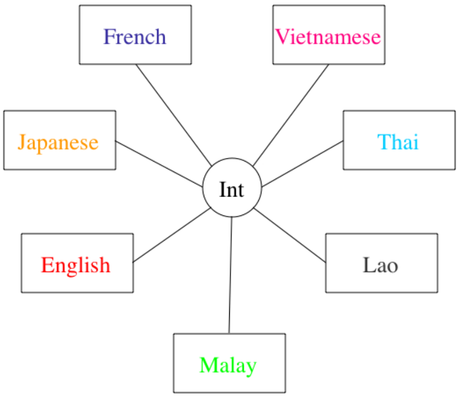


**Fig. 1**. Papillon Pivot Macrostructure

The microstructure of the monolingual entries is rich and very detailed. It is based on the structure used for the DiCo lexical database[5] from OLST, University of Montreal. The encoding method is borrowed directly from the explanatory and combinatorial lexicology part of the Meaning-Text Theory (MTT). This theory gives the information needed to go from a meaning to its realizations in a given language. The microstructure of the dictionaries is independent of the languages and can theoretically be used by humans and machines.

Each entry or lexical unit is based on the word sense or lexie. It consists of a name, grammatical properties, a semantic formula that can be seen as a formal definition - in the case of a predicative word sense, the formula describes the predicate and its arguments - and then a list of lexico-semantic functions - there are 56 basic functions applicable to any language and they can be combined between them - a list of examples and a list of idiomatic expressions.

The Papillon project specifications are directly inspired by this famous holy grail. But like any ambitious project, it cannot be acheived in one shot. Within time, the project Papillon has become a kind of framework or meta-project [6] with several derivatives projects, each one corresponding to a particular aspect of our initial goal. As we will detail, the tools and systems aspects are covered by the Jibiki project and data collection by the JeuxDeMots project.

### 4.2. On the contributing aspects: Wikipedia and Wiktionary projects

The online contributory encyclopedia Wikipedia encountered a large and unquestionable success. We could expect a similar success for its little brother Wiktionary but it is yet to go (1.5 million entries for French and only 44,000 for Japanese). Wiktionaryz, who claimed to solve wiktionary misconception problems has not yet achieved its goal. Wiktionary is anyway

[1] http://www.papillon-dictionary.org

not truly a bilingual dictionary even if there are some translation links (for example. indications of the translations context is missing).

One hypothesis that could explain this problem is the motivation. Indeed, when a person contributes to a Wikipedia article, it is rewarded by the fame. It will then be recognized as an expert in its field. It is not possible with a dictionary. The contributions are located on small parts of information and are therefore anonymous.

On the other hand, there is a technical aspect related to the structure. An encyclopedia article has a more or less free structure while a dictionary entry must follow a very specific one (catchword, grammatical information, semantic blocks, translation block, blocks of examples, etc..). It is not possible to reuse a wiki platform for build a dictionary with a well defined structure.

Once accepted the idea that writing entries dictionary is not as fun as working on a Wikipedia article, we must find solutions to motivate a community of volunteers to contribute to a dictionary. Serious lexical games are a first track. We should also highlight contributors suing for example an array of top contributors of the month. And finally, using community networks such as Facebook should also grist to the mill.

### 4.3. On the data collection via serious games: the JeuxDeMots project

The JeuxDeMots game[7] aims at building a rich and evolving lexical network, that could be compared to a certain extent to the famous WordNet [8] database.

The principle is the following: a game needs two players. When player A initiates a game, an instruction is displayed concerning a type of competency corresponding to a lexical relation (synonym, antonym, domain, intensifier, etc.) and a word W is chosen randomly in the database. Player A has then a limited amount of time for giving propositions that answer the instruction applied to the word W.

The same word W with the same instruction is proposed to another player B and the process is the same. The two half-games of player A and player B are not simultaneous but asynchronous. For each common answer in A and B propositions, the two players earn a certain amount of points and credits. For the word W, the common answers of A and B players are entered into the database. This process participates to the construction of a lexical network linking terms with typed and weighted relations, validated by pairs of players. The relations are typed by the instructions given to the players and weighted with the number of pair players that proposed them. The first version of the French game was launched in July 2007.

### 4.4. On the technical aspects: the platform Jibiki

Jibiki [9] is a generic online platform for manipulating lexical resources with users and groups management. It is a community website developed initially for the Papillon Project. The platform is programmed entirely in Java, based on the "Enhydra" environment. All the data is stored in XML format in a database (Postgres). This website offers two main services: a unified interface to access simultaneously to many heterogeneous resources (monolingual dictionaries bilingual. multilingual databases, etc..) and an editing interface in order to contribute directly to dictionaries available on the platform.

The editor is based on a HTML template interface instantiated with the entry one wants to edit. The template can be generated automatically from a description of the entry structure using an XML schema. It may then be modified to improve screen rendering. It is possible to edit any type of dictionary provided that it is encoded in XML.

Several construction projects of lexical resources have used or still use this platform with success, like the GDEF project about a Estonian-French bilingual dictionary or the LexALP terminological database. The code for this platform is open source and available for download from the LIG laboratory forge[2].

## 5. DESCRIPTION OF THE RESOURCE TO BUILD

### 5.1. Microstructure of entries based on the Meaning-Text Theory

The microstructure of the entries composing the volumes monolingual is a simplification of the Papillon project one. This time, the entry is based on a whole word. A word is either a combination of lexical items (word meanings) or an idiomatic expression. To cope with different skill levels of contributors, the editing interface can adapt itsel and show an adapted granularity of information. For example, a beginner contributor will be invited to give a simple gloss to in order to characterize a word sense, while an expert linguist will describe the entire semantic formula.

### 5.2. Pivot macrostructure via bilingual interfaces

The macrostructure is also derived from the Papillon Project with a monolingual volume for each language and pivot volume in the center. However, in order not to confuse users, they will contribute via an interface with a classical view of bilingual dictionary.

Each bilingual link language A ➔language B added via this interface will actually be translated in the background by the creating two interlingual links and a pivot entry representing the initial translation link. Finally the following schema will be obtained: language A ➔pivot entry ➔language B.

### 5.3. Creating bilingual and interlingual links

If a contributor wants to add a translation link between a word Wa in language A and a word Wb in language B, s/he can establish this link at different levels.

[2] http://jibiki.ligforge.imag.fr

The ideal solution is to connect a word sense Sa of the word Wa to another word sense Sb of the word Wb. In this case, the link is bijective and Sb is also connected to Sa (see Figure 5.3).

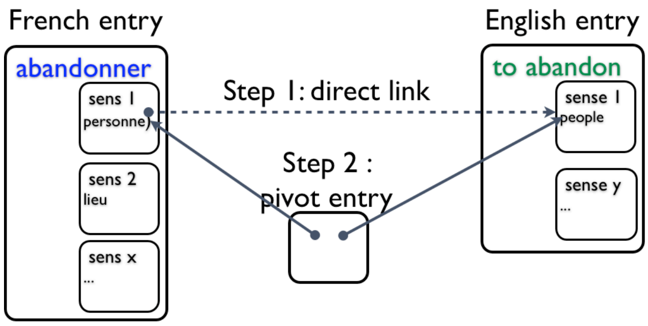


**Fig. 2**. Word Sense Linking Process

If the word Wb does not have any precise word sense or If the contributor is not able to choose the correct one, it can connect directly to the word Wb. In this case, a new word sense Sb is created with a draft quality level and the link and the word sense are labeled as to refine.

In the case of reusing existing data, it is often impossible to relate information to a specific word sense. In this case, we add at the end of the word Wa, the information that one of Wa word senses may be linked to one of the word senses of Wb, but This information will not be added to Wb. It is be of course tagged as to refine as soon as possible!

With the pivot macrostructure, if two links language A ➔language B and language B ➔language C exist, then it will automatically created a link language A ➔language C which quality level will be marked as draft and revise.

### 5.4. Data quality and contributors levels

Each part of information for each entry will be assigned a level of quality. The levels range from 1 star for a draft (when the reused data quality is not known) to 5 stars quality certified by an expert (eg, a link translation validated by a certified translator).

Similarly, contributors will be assigned a proficiency level (1 to 5 stars also). 1 star is the level of a beginner unknown in the community and 5 stars being the level of an acknowledged expert.

Then, when a contributor of level 3 reviews an entry of level 2, the entry level rises automatically to level 3. Similarly, if the work of a contributor without corrections is systematically validated by other contributors of higher level, s/he can pass automatically to the next level after a certain threshold (eg 10 contributions). For example, Figure 5.4 shows an entry with a level of 3 stars.

To go further, we plan to analyze the work of contributors. If a person contributes heavily for example on a particular area, the system can automatically send regular contribution proposals in the domain.

## 6. DATA BUILDING METHODOLOGY

The data building methodology consists in three main steps: retrieving existing data, collecting new data via serious games and finally, online contribution on the Web.

### 6.1. Retrieving existing data

To encourage contributions, it is preferable to propose a skeleton of existing data (even of bad quality), rather than an empty dictionary (writer's block). For each language involved, a list of words will be collected in order to create an initial list of entries. It is always possible to create a new entry, but the creations will be subject to verifications.

According to the sub-projects and languages involved, several dictionaries can be used:

- Fe* dictionary projects (French - English + other language): FeM (Malay) [10], FET (Thai), Feb (Vietnamese);
- The DiCo database for French;
- The VietDict French-Vietnamese bilingual dictionary.
- The French-Khmer phonetic bilingual dictionary[11].

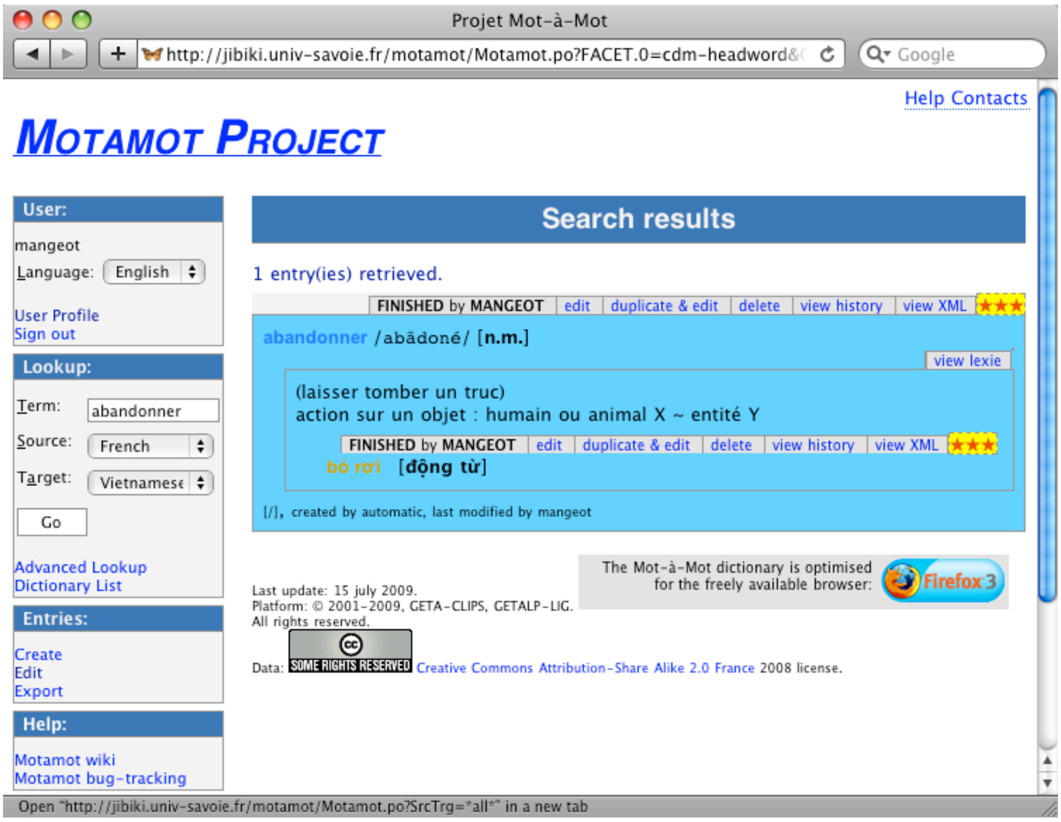
The number of stars of initial items generated from this data set is based on the quality of dictionary and the granularity of data retrieved.

#### *6.1.1. Special handling for Khmer*

For the Khmer language, we plan to computerize an existing French-Khmer phonetic dictionary. Its building began in the late 90s, and was completed in 2006 by a small group of researchers and computer scientists gathered in the non profit organisation "Pays perdu" created by Denis Richer, a French ethnolinguist, established in Siem Reap (Cambodia). The first version of dictionary was published in spring 2007 and includes 20,000 entries. Table 1 shows an example of what the dictionary looks like.

**Table 1**. Excerpt of the French-Khmer dictionary

| French | Khmer |
|---|---|
| jarret | kɔ̄nlēak-cə̥̄ɳ |
| jars | kəɳān-chmō̥l |
| jasmin | mlih |
| jauge | □cɑmnoh-rɔ̄ɳvuəl |
| — (techn.) | māet-stuəɳ |



Projet Mot-à-Mot

http://jibiki.univ-savoie.fr/motamot/Motamot.po?FACET.0=cdm-headword&

Help Contacts

MOTAMOT PROJECT

**User:**

mangeot

Language: English

User Profile
Sign out

**Lookup:**

Term: abandonner

Source: French

Target: Vietnamese

Go

Advanced Lookup
Dictionary List

**Entries:**

Create
Edit
Export

**Help:**

Motamot wiki
Motamot bug-tracking

**Search results**

1 entry(ies) retrieved.

FINISHED by MANGEOT | edit | duplicate & edit | delete | view history | view XML ★★★

abandonner /abɑ̃dɔne/ **[n.m.]**

view lexie

(laisser tomber un truc)
action sur un objet : humain ou animal X ~ entité Y

FINISHED by MANGEOT | edit | duplicate & edit | delete | view history | view XML ★★★

bỏ rơi **[động từ]**

[/], created by automatic, last modified by mangeot

Last update: 15 july 2009.
Platform: © 2001-2009, GETA-CLIPS, GETALP-LIG.
All rights reserved.

The Mot-à-Mot dictionary is optimised for the freely available browser: Firefox 3

Data: SOME RIGHTS RESERVED Creative Commons Attribution-Share Alike 2.0 France 2008 license.

Open "http://jibiki.univ-savoie.fr/motamot/Motamot.po?SrcTrg=*all*" in a new tab

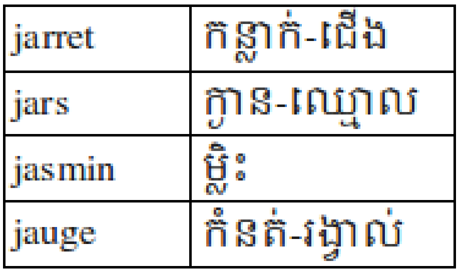
**Fig. 3**. Example entry in MotÀMot dictionary

The dictionary is in Word format and the Khmer part consists only in a phonetic transcription of the entry written in a special IPA script (SIL Sophia IPA 93) set up by the Summer Institute of Linguistics. Therefore, in order to obtain a completely XML Unicode data, we have to perform the following steps:

- convert the words in SIL IPA script into an Unicode IPA script
- obtain the equivalent words in khmer script. This step might be done semi-automatically, but we will probably need a post-edition step.
- tag the entries Most of the French entries consist in a simple word, but some of them have additional information that is not tagged, eg: a gloss `jambose (fruit)`, a feminine `jardinier, ère`, a domain `jauge - (techn.)`, etc.

Figure 6.1.1 shows the same example in khmer script.

Unfortunately, the Khmer encoded in Unicode cannot yet be read correctly on all the current platforms. On the Apple MacOs, some characters are not rendered correctly. For example, the figure 6.1.1 shows the rendering of the first khmer word. Therefore, we may need to automatically generate an image for each khmer word in order to fix the problem temporarily. A better solution would be to discuss with Apple in order to fix the problem definitely, but this is another story.



| | |
|---|---|
| jarret | កន្លាក់-ជើង |
| jars | ក្ងាន-ឈ្មោល |
| jasmin | ម្លិះ |
| jauge | កំនត់-រង្វាល់ |

**Fig. 4**. Dictionary in Khmer script

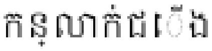


**Fig. 5**. Rendering problems for Khmer

### 6.2. Data Collection via serious games

The idea is to launch a JeuxDeMots for each language project[3]. The French JeuxDeMots was launched two years ago. The Vietnamese version was launched in autumn 2009. The khmer version is being translated. We hope to find a similar success than the French JeuxDeMots. Furthermore, we should should think about games allowing bilingual data collection. People interested can contact us if they want to launch a JeuxDeMots game in their language.

### 6.3. Online contribution on the web

The retrieved data is collected and then merged in order to give birth to a skeleton dictionary. It is then put online for correction and enrichment.

## 7. CONCLUSION

The project is already fairly well advanced. Most technical aspects concerning the platform and online serious games are solved. It remains to gather and convert existing resources. The major challenge of the project is actually our ability to motivate communities of contributors. We hope that our experience and the attraction of such a project will allow us to make a significant step in these sociological aspects.

The benefits of such a project are numerous and will help to revive the interest in the francophonie in the Southeast Asian countries. Data generated can be used by learners of French in these countries, or francophones wishing to learn a Southeast Asian language. The dictionaries may be used by tourists or businessmen directly online or via PDAs.

The communities of contributors should launch a new dynamics of cooperation around a common humanistic purpose. Moreover, it can arouse interest for expanding the project to other languages of the region.

## 8. ACKNOWLEDGMENTS

The MotAMot project has been partly funded by the Action de Recherche en Réseau program of the Agence Universitaire de La Francophonie[4].

## 9. REFERENCES

[1] Vincent Berment, *Méthodes pour informatiser des langues et des groupes de langues "peu dotées"*, Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France, 18 mai, 277 p. 2004.

[2] Igor Mel'čuk and Alain Polguère, ``Dérivations sémantiques et collocations dans le dico/laf.,'' in *Langue française, numéro spécial sur la collocation « Collocations, corpus, dictionnaires »*, vol. 150, pp. 66--83. P. Blumenthal and F. J. Hausmann, June 2006.

[3] Étienne Blanc, ``Parax-unl: a large scale hypertextual multilingual lexical database,'' in *NLPRS'99: the 5th Natural Language Processing Pacific Rim Symposium*, Beijing, China, 1999, p. 4.

[4] Alain Polguère, ``Structural properties of lexical systems: Monolingual and multilingual perspectives,'' in *Workshop on Multilingual Language Resources and Interoperability (COLING/ACL 2006)*, Sydney, 17-21 July 2006, pp. 50--59.

[5] Alain Polguère, ``Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for french,'' in *Proceedings of EURALEX'2000*, Stuttgart, Germany, 2000, pp. 517--527.

[6] Mathieu Mangeot, ``Papillon project: Retrospective and perspectives.,'' in *Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine, LREC workshop*, Pierre Zweigenbaum, Ed., Genoa, Italy, 22 May 2006, p. 6.

[7] Mathieu Lafourcade and Alain Joubert, ``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, 12-14 mars 2008, pp. 657--666.

[8] G. A. Miller, R. Beckwith, C . Fellbaum, D. Gross, and K. J. Miller, ``Introduction to wordnet: an on-line lexical database,'' *International Journal of Lexicography*, vol. 3, no. 4, pp. 235--244, 1990.

[9] Mathieu Mangeot and David Thevenin, ``Online generic editing of heterogeneous dictionary entries in papillon project,'' in *Proc. of the COLING 2004 conference*, Geneva, Switzerland, 26 August 2004, vol. 2, pp. 1029--1035.

[10] Yvan Gut, Puteri Rashida Megat Ramli, Zaharin Yusoff, Kim Choy Chuah, Salina A. Samat, Christian Boitet, Nicolai Nedobejkine, Mathieu Lafourcade, Jean Gaschler, and Dorian Levenbach, *Kamus Perancis-Melayu Dewan, Dictionnaire francais-malais*, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1996.

[11] Denis Richer, Keo T, and Vanra Ieng, *Dictionnaire Français-Khmer (en phonétique)*, D.R. Edition, 2007.

[3] `http://www.jeuxdemots.org`

[4] `http://www.ltt.auf.org`