Papillon Lexical Database Project: Monolingual Dictionaries and Interlingual Links

Mathieu Mangeot

Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links

MATHIEU MANGEOT
GETA-CLIPS-IMAG BP53, Campus universitaire,
385 avenue de la bibliothèque 38041 Grenoble, France
E-mail: Mathieu.Mangeot@imag.fr

Abstract

This paper presents a new research and development project called Papillon. It started as a French-Japanese cooperation between laboratories GETA/CLIPS (Grenoble, France) and NII (Tokyo, Japan). Its goal is to build a multilingual lexical database and to extract from it digital bilingual dictionaries.

The database is built with monolingual dictionaries, one for each language of the database, linked to an interlingual dictionary. The pivot architecture of the database is based on Gilles Sérasset's Ph.D. thesis. The structure of the monolingual dictionaries is based on the lexical work done by Igor Melc'uk and Alain Polguère.

From the lexical database, it is planned to derive user customized bilingual dictionaries in multiple target formats. It will be possible to generate human usage dictionaries as well as specialized dictionaries for machine translation software. These dictionaries will be available under the terms of an open source license.

This project, initiated by some computational linguists, aims at being useful and open to all those who are interested in Japanese and French. It is also opened to any other language. Moreover, the pivot architecture of the database will facilitate the addition of new languages and save translation efforts.

Keywords: dictionary, lexical database, multilingual, lexicologist, lexicographer, French, Japanese

1. Introduction

There are few French-Japanese usage dictionaries, which are really usable and useful for French speakers. The main problem is that the original Japanese script and the rômaji phonetic transcription are present together only in very small dictionaries. Also, dictionaries never contain numeric specifiers, which are as important in Japanese as gender and number in French. On the other hand, the information available in paper dictionaries does not exist in machine-readable forms, or is not accessible on line.

The lack of bilingual resources is also an obstacle to develop linguistic software applications, for which adapted dictionaries are a need. As an example, Nippon Telegraph and Telephone in Japan or Lexiquest in France have to develop their own dictionaries in a separate and time-consuming effort. In the academic world, this implies that applications that have been created for French and Japanese offer only a reduced scope, while good English-Japanese pieces of software are available.

Nevertheless, it is a true fact that Japan is very interested in the French language. Conversely, a growing number of French individuals invest much energy to learn Japanese. There is a vacuum to be filled.

The leveraging of communication that Internet offers allows one to think that a convenient digital dictionary could be produced by a general cooperation between linguists, translators, computer scientists, etc., working together through Internet.

A similar project between English and Japanese has been active for about a decade. This project has allowed the effective building of a free Japanese-English dictionary, available through an Internet server. This Edict project has been created and supported by Pr. Jim Breen from Monash University, Australia [b2]. The current JMDict dictionary comprises now 70,000 entries of common vocabulary, a specific kanji dictionary, and around twenty specialized dictionaries (biology, law, etc).

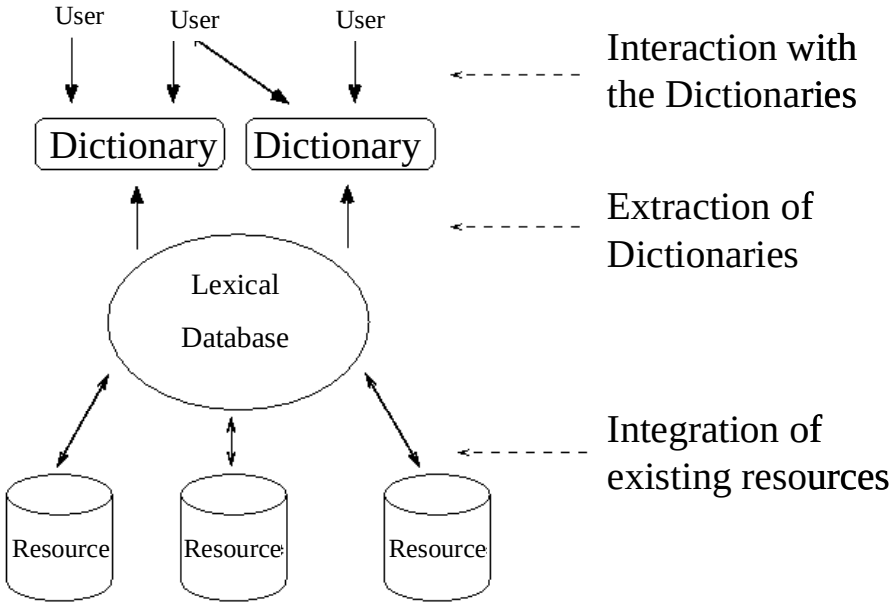
A different project, fed by volunteers, is supported by NEC Corporation. Its aim is to increase the dictionaries used by the NEC translation tool [b3], and to bring in new entries on a constant way.

We should also mention the SAIKAM project [1], [b5] cooperation between NII (Tokyo, Japan) and NECTEC (Bangkok, Thailand) active since about 5 years, where Thai students working or having worked in Japan have built a sizable Japanese-Thai online dictionary through Internet.

In such a context, the GETA/CLIPS laboratory (Grenoble, France) and the National Institute of Informatics (Tokyo, Japan) started a research and development project in order to plan and implement a French-English-Japanese lexical database. Here are described the architecture of the database, the structure of the entries and the methodology adopted for the project.

2. General View of the Database

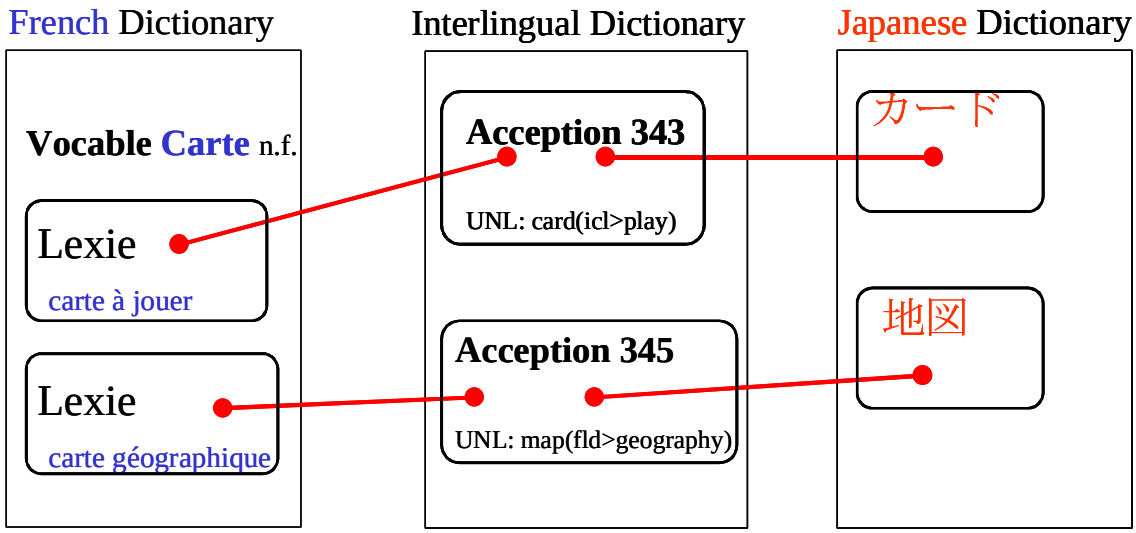
The lexical database is built on the one hand by integrating existing resources and on the other hand by writing and correcting new entries.



Once the database is homogeneous, users will be able to extract their own customized dictionaries dynamically from the database and to interact with them.

3. Internal Architecture of the Database

The database will be built using a pivot architecture based on Dr. G. Sérasset's Ph.D. thesis [6] and experimented by Dr. E. Blanc in PARAX [2]. The monolingual dictionaries will be linked only through a pivot dictionary of interlingual links called acceptions. These acceptions will also be linked together by refinement links. They may also be translated into the UNL language [8], [b6].



Each sense or meaning of each entry of a monolingual dictionary is linked to one or more acceptions of the pivot dictionary. For example, in French " carte " has two meanings: " carte à

jouer (card) ” and “ carte géographique (map) ”. The entry “ carte ” will consequently be linked to two "lexies" (corresponding to 2 word senses) in the French monolingual dictionary, which in turn will be linked to 2 acceptions in the pivot dictionary: in the example, the first has number 343, with the corresponding UNL "UW" (universal word) “card(icl>play)”, and the second one has number 345, with UW “map(fld>geography)”.

```
<axi id="a001">
    <lexies>
           <lexy lang="fra" ressource='papillon-fr.xml'
               idref="meurtre#n.m.@1" />
    </lexies>
    <external_references>
           <UWs ressource="UNL-fr.unl">
                <uw idref="murder" />
           </UWs>
    </external_references>
</axi>
```

Example of an interlingual acception encoded in XML

4. Structure of the monolingual dictionaries

The structure of the entries or microstructure of the monolingual dictionaries is based on the structure used for the formal lexical database DiCo [5] of the OLST laboratory in Université de Montréal. The encoding methodology is directly borrowed from the explanatory and combinatorial lexicology, which is part of the meaning-text theory [4].

1. Name of the lexical unit: MEURTRE
2. Grammatical properties: nom, masc
3. Semantic formula: action de tuer: PAR L'individu X DE L'individu Y
4. Government pattern: X = I = de N, A-poss Y = II = de N, A-poss
5. (Quasi-)synonyms: {QSyn} assassinat, homicide#1; crime
6. Semantic derivations and collocations: {V0} tuer
   {A0} meurtrier-adj
   {S1} auteur [de ART] //meurtrier-n /*Nom pour X*/
7. Examples: La mésentente pourrait être le mobile du meurtre.
8. Full idioms: _appel au meurtre_
   crier au meurtre

The dictionaries will be encoded in XML to facilitate readability and conversion into various target formats.

5. Building methodology

The building methodology of the lexical database builds on one hand on the reuse of existing data, the French-English-Malay dictionary [3], [b1] and the Japanese-English dictionary of Jim Breen [b2], and on the other hand on the contribution of volunteers working through the Internet.

Different steps are planned: The first step is the integration of existing resources. It consists in preparing a "lexical soup" by merging the two dictionaries thanks to the presence of English. This merging operation will produce correct as well as incorrect acceptions (interlingual links). These wrong acceptions will be corrected or deleted by lexicologists.

Then the voluntary contributors will index new entries and the lexicologists will correct and integrate them into the database. It will create a cycle of edition/correction/modification of the entries between the lexicographers/contributors and the lexicologists. Different kind of contributors can work on the database:

- specialists of one language will write the monolingual entries;
- people with good knowledge of French and Japanese like translators will work on the links between the monolingual entries and the acceptions;
- people with good knowledge of UNL will translate the acceptions into UNL [8].

6. Dictionaries produced

Several monolingual or bilingual dictionaries can then be extracted from the database. Different types are needed: for human use, via database and plug in functionalities or via usual dictionary formats, and for machine use.

6.1 For human use, via database and plug in functionalities

Persons that interact in foreigner languages often can access computers. One of the aims of this dictionary is then to provide them with a direct help, within their editor, browser, or their daily used personal digital assistant.

6.2 For human use, via usual dictionary formats

We plan to automatically derive from the database digital presentations for web consultation and paper edition. The FeM [b1] and JMDict [b2] formats are the first targeted formats.

6.3 For machine use

The terminology resources available for building lingware (linguistic software) are almost null between Japanese and French. The rare available ones have to be radically restructured and augmented. The orientation of the Papillon lexical database towards possible use by machines will encourage the realization of lingware including both languages, by providing a first support for such projects.

7. Conclusion

The pivot architecture allows an easy integration of new languages because the reuse of existing links will save a lot of time consuming efforts. The Thai language is already about to integrate the project through a cooperation with Kasetsart University (KU/Thailand), and National Electronics and Computer Technology Center (NECTEC/Thailand).

The open source license makes all the data available to anyone. Furthermore, we will be able to generate multiple formats from the lexical database.

Finally, it should be stressed that such an endeavor will not only need the dedication of as many volunteer contributors as possible, but some stable support, in the form of a server and, more difficult, of a central team of experts charged of "refining the raw ore" of individual contributions.

That team does not have to be in a single place, but convenient groupware tools should be developed for it.

References

[1] **Vuthichai Ampornaramveth, Akiko Aizawa, Keizo Oyama, Tasanee Methapisit (2000)** *Implementation of an Internet-Based Dictionary Development Environment: SAIKAM* (in Japanese) Research Bulletin of the National Center for Science Information Systems vol.12, p.101-109 (2000)

[2] **Etienne Blanc (1999)** *PARAX-UNL: A large scale hypertextual multilingual lexical database*. Proceedings 5th Natural Language Processing Pacific Rim Symposium 1999, Tsinghua University Press, Beijing, 1999, p.507-510.

[3] **Yvan Gut, Puteri Rashida Megat Ramli, Zaharin Yusoff, Chuah Choy Kim, Salina A. Samat, Christian Boitet, Nicolas Nédobejkine, Mathieu Lafourcade et al. (1996)** *Kamus Perancis-Melayu Dewan, dictionnaire français-malais.* Dewan Bahasa Dan Pustaka, Kuala Lumpur, 667 p.

[4] **Igor A. Mel'cuk (1997)** Vers une linguistique Sens-Texte. Leçon inaugurale, Collège de France, Chaire internationale, 43 pages. http://www.fas.umontreal.ca/LING/olst/FrEng/melcukColldeFr.pdf

[5] **Alain Polguère (1998)** La théorie Sens-Texte. Dialangue, Vol. 8-9, Université du Québec à Chicoutimi, pp. 9-30. http://www.fas.umontreal.ca/LING/olst/FrEng/PolgIntroTST.pdf

[6] **Gilles Sérasset (1994)** *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA*, COLING-94, 5-9 August 1994, vol. 1/2 : pp. 278-282.

[7] **Mutsuko Tomokiyo, Mathieu Mangeot & Emmanuel Planas (2000)** *Papillon : a Project of Lexical Database for English, French and Japanese, using Interlingual Links.* Journées Science et Technologie de l'ambassade de France au Japon, 13 Novembre 2001, Tokyo, Japon, 3 p.

[8] **UNL (1996)** *Universal Networking Language.* UNL center, Institute of Advanced Studies, The UN University, 1996, 74 p.

Bookmarks

[b1] **FeM Dictionary:** http://www-clips.imag.fr/geta/services/fem

[b2] **JMDict Japanese->English:** http://meshplus.mesh.ne.jp/CRV2/dic/club/down.html

[b3] **NEC project:** http://meshplus.mesh.ne.jp/CRV2/dic/club/down.html

[b4] **Papillon Project:** http://vulab.ias.unu.edu/papillon/index.html

[b5] **SAIKAM Project:** http://saikam.nii.ac.jp

[b6] **UNL Project:** http://www.unl.ias.unu.edu