Back to the roots: building a French-Japanese dictionary

Mathieu MANGEOT

Condillac - LISTIC - Université de Savoie, Campus Scientifique
F-73376 LE BOURGET DU LAC CEDEX France
Tel. +33 4 79 75 87 85, Fax.: +33 4 79 75 88 88
e-mail : Mathieu.Mangeot@univ-savoie.fr

Abstract

After four years of life, the Papillon project is dying. In this paper, I summarise the current situation and then give my opinion of the reasons why we did not achieve what we planned four years ago. Then, taking these reasons as a starting point I propose to go back to the original aim of the project, i.e. to build an online bilingual French-Japanese dictionary free of rights. I describe the macro and microstructures, the new methodology and the existing data we could (re)use in order to build at once a bidirectional bilingual dictionary with links that can be directly integrated afterwards in a multilingual pivot database. This (sub)project serves several goals: first, to build the dictionary itself of course, second, to motivate and federate contributors around a project and a platform, third, to advertise the Papillon project, fourth, to test the platform and the methodologies in real conditions and fifth, to build linked data that can be reused in a more long term for the Papillon Multilingual database. Of course, I cannot achieve this alone, so I am looking for support from your part!

1 Introduction

Let me begin by a warning: the opinions written in this paper come only from my personal point of view and my ideas are probably not shared by everybody. They might shock some sensitive souls. Nevertheless, I have a very good knowledge of the Papillon project (Mangeot, 2001) (Mangeot et al., 2004) (Mangeot and Thevenin, 2004) so I think it is worth to tell them. Furthermore, I also think that it is time to create a new dynamics otherwise, the risk that the project would die is considerable.

I begin by resuming the current situation and by giving what I think are the reasons why we did not achieve the aims previously defined. Then, I take these reasons as a starting point in order to go back to the original aim of the project, that is to build an online bilingual bidirectional French-Japanese dictionary.

2 Current situation

2.1 The Meetings

The first Papillon meeting (Tomokiyo et al., 2000) took place in Tokyo, in August 2000. since then, we regularly organised a meeting per year. The 2001 meeting took place in July in Grenoble. The 2002 meeting took place in July in Tokyo. The 2003 meeting took place in July in Sapporo. The 2004 meeting took place in August in Geneva. The first two years, we discussed a lot about the dictionary structures we would use and the organisation of the project.

2.2 The Data

Concerning the data, we decided to separate it into two layers: the purgatory and the paradise.

The purgatory gathers all the recuperated data in XML but in original structure. In four years, we managed to gather more than 1 million of entries from 14 dictionaries in Chinese, English, French, German, Japanese, Korean, Malay, Thai, and Vietnamese. This data must then be converted into the Papillon structure.

The Paradise gathers the data in Papillon structure. I first converted 613 French entries from the DiCo database (Polguère, 2000). They were reconverted in more details by Guy Lapalme (Lapalme and Sérasset, 2003) in 2003. Then a few contributors edited a small set of

dictionary entries in Papillon XML format. In 2001, Mutsuko Tomokiyo wrote 202 French entries, 183 English entries and 105 Japanese entries; In 2002, Tang Enya Kong and Siti Khaotijah Mohammad wrote 67 Malaysian entries. Unfortunately, few entries are translations the ones from the others. Nevertheless, I could write 41 axies between English, French and Japanese entries and one axie between English, French, Japanese and Malaysian entries. No other entries were created or converted since then. To summarise, today, we have 1,170 lexies in 4 languages and 42 axies.

2.3 The Platform

The best progress in Papillon project were realised on the platform. Gilles Sérasset began to write a prototype in Enhydra/Java in 2000. In 2001, I began to co-operate closely with him. I worked almost full time on the development of the platform during three years from 2002 to 2005. We still work on it when time constraints are not too heavy.

In 2003, I began to work on another dictionary construction project: the GDEF project[1] about a big Estonian-French dictionary. We had a deadline for the official beginning of the project that was July 2005. Thus, I had to solve as many problems and bugs as possible in order to have a usable version at that time.

In the meantime, more exactly since the beginning of 2005, Gilles began to work on a terminology construction project, the LexAlp project that uses also the platform. He had to show rapidly a usable prototype for June. This is why we both worked hard on the platform in order to deliver a stable version for this summer. Furthermore, we had to take into account the differences of the two projects.

I enriched the platform used by the GDEF project with two important functionalities. In order to facilitate the communication between the platform users, I installed an Open Source online forum in PHP (phpBB2). The second tool I installed is called Uplug[2]. It's a GPL software for building and querying online bilingual corpora.

During the test phase of the GDEF project, I received a good feedback from the lexicographers. They built 1,500 entries from scratch. Now, the platform is used everyday by a group of 10 people. They edit entries online. In 10 days, 250 entries were edited and revised without any problem.

The platform can be used as is for building a new dictionary. All the essential functionalities are implemented and work fine. Of course, we will not stop here the development. We would like in particular to redesign some parts and work on specifications in order to build a generic platform easily adaptable for new projects.

3 Identified problems

3.1 Wrong identification of the bottleneck

When we began the project, we first thought that the tools we needed to build the dictionary would be the bottleneck. In other words, once the tools would be ready to use, we could start to write Papillon entries. Emmanuel Planas, and Magali Drant (an intern) even started right away a prototype in Java that could create inter-lingual links between entries of two different volumes simply by drawing the link with the mouse. Unfortunately, this prototype has never been used.

In the meantime, Gilles Sérasset started the development of the actual Papillon platform. Once again, during this development, we strongly thought that once the basic needed functionalities would be implemented, we could begin to write entries on the platform. Four years later, the platform is not only ready but also used daily in two other lexical resources construction projects with success.

We have to admit and two other projects later, the we have to admit that the key problems that forbid us to continue on the project are from another type. My hypothesis is the following: we are not specialists of lexicography, most of us are computer scientists with some knowledge of linguistic. Thus we underestimated the lexicographic aspects of the project.

3.2 Very few Japanese collaborators

The project was launched by Emmanuel Planas and François Brown de Colstoun who were working in Japan at that time, and Mutsuko Tomokiyo who was working in France. I spent myself more than two years in Japan to

[1] http://www.estfra.ee

[2] http://stp.ling.uu.se/cgi-bin/joerg/Uplug

work on the project. Most of the Papillon partners have strong interests in Japanese: Francis Bond and Yves Lepage are still working in Japan; Jim Breen, Ulrich Apel and Jean-Marc Desperrier are directly working on bilingual Japanese dictionary building projects, etc.

Despite all our efforts, very few Japanese researchers accepted to join Papillon project: Mutsuko Tomokiyo (researcher in France), Kyo Kageura (ass. prof. at NII) and Kyoko Kuroda (High school teacher).

For this particular problem, making an hypothesis is quite difficult. I'll give two directions:

First, it seems that the needs of a French-Japanese dictionary is not that crucial for Japanese. Almost every already existing French-Japanese resources were designed for Japanese. For example, the missing kana or counters are not an issue for them. Furthermore, although they are not free, some of them are relatively complete with a good quality (like Shogakukan-Robert 小学館ロベール with about 120,000 entries and 250,000 examples).

Second, once again, the researchers we know in Japan are not lexicographers or researchers in lexicography. The people interested were generally very busy and wanted to see a more advanced concrete project before joining. We should build some core data in order to demonstrate the viability of the project as well as look into other directions.

3.3 Macrostructure too abstract

The pivot macrostructure that links monolingual volumes with inter-lingual links through a pivot volume seems to be difficult to understand by the majority of the people, even the domain experts. Most people believe that we try to build a semantic inter-lingua. For most of the people, it is impossible to imagine how these links can be built.

We may not have explained clearly this pivot structure and the way we build it. Furthermore, we do not have yet any example set to show (I believe that a good example is better than a long even detailed explanation).

I would like to go a step farther by saying that the reason why we did not explain the pivot structure clearly is it properly because even ourselves have not a clear idea of this structure. Another illustration of this is that in four years, despite very promising ideas (like semantic vectors), we were not able to build any example set (even very small).

Something we might consider in order to build an example set would be to recuperate data of the PARAX pivot database built by Étienne Blanc (Blanc et al., 1994) (Blanc, 1995) if it is still technically possible.

3.4 Wrong entry unit

I think that the idea to directly use the lexies as the basic entry unit is wrong. I list here four reasons why:

Some data is shared between lexies of the same vocable (transcriptions, pronunciation, grammatical features, etc.), thus if we manage the lexies separately, it increases the risk of inconsistencies.

The ranking (mostly by frequency) between the lexies of the same vocable is very important when someone wants to translate a word into another language. For the moment, we have no way to indicate this ranking.

Another problem is the difficulty to distinguish between lexies of a same vocable and lexies of homograph vocables. For the moment, the lexies are displayed alphabetically but not gathered in vocables.

The latest reason is that the majority of the people are used to the traditional "vocable" entry unit. We should follow their entry unit representation in order to facilitate the adhesion of the contributors.

3.5 Microstructure too complex

The Papillon microstructure is based on the Explanatory and Combinatorial Lexicography (ECL)(Mel'čuk et al., 1995), part of the Meaning Text Theory. It is an XML'ized version of the DiCo database designed by Alain Polguère, Igor Mel'čuk and their colleagues of Université de Montréal in Canada.

This structure may be very interesting and convenient scientifically but the learning curve is by far too long to think that any contributor will be able to understand it (even parts of it) in a glance.

Furthermore, I think that the idea that the users will contribute at their level of knowledge and that ECL specialists will complete the structure is wrong because in order to contribute correctly on even the most easily understandable part of the entry structure, peo-

ple need to understand at least the main lines of the ECL and this is already too much to ask.

3.6 Eyes bigger than stomach

Instead of building a first core of data by hands, we imagined that we could first build a skeleton of data from existing one by using various techniques in order to build (semi)-automatically inter-lingual links. We tried to use for example the angular distance between semantic vectors in order to evaluate the semantic distance between two lexies.

Of course, this direction must be followed, but we should not put all our eggs in one basket. While we are waiting for this core, people are forgetting the project. Thus it is slowly dying.

In the meantime, we should also start to build data "by hands" from what we have now. This experiment is necessary to know if the chosen structures are adequate. Furthermore, The data can then be used for various experiments and bootstrap for automatic linking mechanisms. It would also help to establish a community of contributors around the project, and thus to increase the visibility of the project.

I think that instead of willing to build automatically all at once, we should increment the complexity step by step: building data by hands and then automatically; building first a bilingual dictionary and then a multilingual pivot database, etc. The next section tries to follow this principle.

4 Building a French-Japanese dictionary

In this section, I use the previous analysis of what I identified as problems in Papillon project in order to propose a more realistic direction to the project that would be a first step in the building of a multilingual pivot database. It consists in building "by hands" a relatively simple bilingual bidirectional French-Japanese dictionary. This was also in fact the very first aim of Papillon project.

4.1 Structures

4.1.1 Macrostructure

I propose to keep a pivot macrostructure with inter-lingual links (axies) between lexies. But the inter-lingual linking mechanism is hidden. The linking interface shows only two languages, so that the contributors build only bilingual links between French and Japanese.

4.1.2 Entry unit

The entry unit is a traditional "vocable" made of one grammar block per part-of-speech. Each grammar block contains one or more lexie blocks sorted by frequency order.

4.1.3 Microstructure

The microstructure should tend to the ECL in order to reuse this data in a future Papillon database but must be much more simple:

- For every vocable: headword, homograph number, variants, pronunciation in IPA.
- For French vocables: kana transcription, inhaled h (h aspiré).
- For Japanese vocables: yomigana, romaji transcription, quantifiers.
- For every grammar blocks: the part-of-speech.
- For French grammar blocks: gender, number, irregular plural and feminine for substantives; conjugation table and "être" auxiliary for verbs.
- For every lexie : the government pattern for predicative lexies, a free text definition, the domain, the language levels, a list of examples, a list of idioms, a translation link to the other language.

4.2 Methodology

4.2.1 Linking process

My hypothesis that drives the linking process is the following: every translation from one language to another is bidirectional, i.e. if there is a non null frequency for a vocable V1 to be translated into another language by a vocable V2, there is also a non null frequency for the vocable V2 to be translated into the vocable V1. The only difference between the two translations is located in the frequencies.

For example, the most frequent Japanese translation of the French vocable "tabouret" is the vocable 椅子【いす】(isu); but the most frequent French translation of the Japanese vocable 椅子 is "chaise". Nevertheless, there are

less frequent cases when 椅子 is translated into "tabouret".

The translation links can be only established in a lexie block. Furthermore, a lexie block cannot contain more than one translation link. Thus, when a user wants to create a translation link for the vocable V1 of language lang1, s/he must add a new lexie L1 in the vocable V1.

Then, when a translation link from vocable V1 of lang1 to vocable V2 of lang2 is established, a new lexie L2 is added to the vocable V2 and another link from lang2 to lang1 is also established. The lexie L2 is added at the end of the lexies block list and is labelled "to be reviewed".

When the vocable V2 is reviewed, the lexie L2 cannot be deleted. Only the order of the lexies can be modified in order to reflect the frequency of the translations of V2 into lang1 (Note to the developers: the interface have to be modified to allow the modification of a list order).

Following our example, the Japanese vocable 椅子 would have at least two lexies: the first one with a translation link to the French vocable "chaise" and the second one to the French vocable "tabouret".

When a vocable cannot be translated directly into another vocable, the translation link is done with the head of the translating expression and the whole expression is taken as a definition for the newly created lexie L2. For example, the Japanese vocable 節分 【せつぶん】(setsubun) is translated in French by the following expression: "jour précédent le printemps où l'on jette des graines de soja pour éloigner la malchance". In that case, a new lexie block is created into the French vocable "jour" and the definition of the lexie is the expression itself.

4.2.2 Writing protocol

The writing protocol involves four types of users:

- Users not logged can look-up existing validated data
- Logged users can create contributions
- Couples of 2 users (A and B): 1 French and 1 Japanese, pertaining to the group "specialist";
- A small group of validators (user V) who validate the reviewed contributions, pertaining to the group "validator";

It follows four steps:

1. creation of a contribution by user A, status "not finished"
2. end of contribution writing by user A, status "finished"
3. revision of the contribution by user B, status "reviewed"
4. validation of the contribution by validator V, status "validated"

Once the contribution is validated, it is added into the volume and can be queried by all the users.

4.3 Available Data

4.3.1 Monolingual Data

We think that the best way is to build a first word list for both languages by reusing existing monolingual data as a starting point.

- **WaDokuJiTen** is a big Japanese-German dictionary construction project led by Ulrich Apel. The data is available for Papillon project. The In our opinion, the Japanese part of the WaDokuJiTen is the most complete Japanese data available freely. It has more than 214,000 Japanese entries.
- **Morphalou**[3] is a French morphological lexicon derived from the Tresor de la Langue Française. It is available with an open-source-like licence. For the French part, we propose to proceed as for the GDEF project and to use Morphalou. It consists of a list of 67,376 lemmas with inflected forms.

4.3.2 Bilingual Data

We propose to reuse only French-Japanese bilingual data and not other bilingual data like Japanese-English one because what people do is to translate the English into French without understanding the Japanese.

[3] http://actarus.atilf.fr/morphalou

- **Dico-F-J** Project led by Jean-Marc Desperrier (Desperrier, 2002); 10,000 entries. Structure common with the JMDict project.
- **FJocean** Terminological French-Japanese lexicon of 3,328 French entries translated into Japanese about the ocean vocabulary available online[4] . The Japanese part have to be parsed in order to identify the translations.
- **Maniette** French translations of the meaning of 2,066 kanji by Yves Maniette for his book "Les kanjis dans la tête" (translation of the English book "Remembering the kanji" by James W. Heisig).
- **Armement** Small terminological lexicon about armament given by the French embassy in Tokyo, Japan with 1,116 terms. The terms are very specifics, it may not be usable for a dictionary of general domain.

5 Conclusion

This paper is mainly the result of my four years of work on Papillon project but also a summary of all the comments I listened every time I presented the project to newbies. I must admit that I was intentionally a little bit provocative. Things may not be that dramatic. Anyway, my aim here is first to give my opinion freely and second to shake the people consciousness about Papillon project. I am waiting for your comments.

I think that if we do not react quickly, Papillon project will soon be dying. Already many people lost interest in following the project. This sub-project is a proposal to add a new dynamics in Papillon. I need your comments on the feasibility of such a project.

Then, if we decide to follow my proposal, I will need to find groups of validators and specialists; and more specifically Japanese contributors.

References

Etienne Blanc, Gilles Sérasset, and Francois Tchéou. 1994. Designing an acception-based multilingual lexical data base under hypercard: Parax. Research report, GETA, IMAG (UJF & CNRS), August.

Etienne Blanc. 1995. Une maquette de base lexicale multilingue à pivot lexical : Parax. In *Lexicomatique et Dictionnairique, Actes du colloque LTT*, pages 43–58. Universités Francophones, Actualités scientifiques, AUPELF-UREF.

Jean-Marc Desperrier. 2002. Analyzis of the results of a collaborative project for the creation of a japanese-french dictionary. In *Papillon'2002 Seminar*, NII, Tokyo, Japan, July. http://www.papillon-dictionary.org/ConsultInformations.po.

Guy Lapalme and Gilles Sérasset. 2003. Batch creation of papillon entries from dico. In *Papillon'2003 Seminar*, Hokkaido University, Sapporo, Japan, July. http://www.papillon-dictionary.org/ConsultInformations.po.

Mathieu Mangeot and David Thevenin. 2004. Online generic editing of heterogeneous dictionary entries in papillon project. In *Proc. of the COLING 2004 conference*, volume 2, pages 1029–1035, Geneva, Switzerland, 26 August.

Mathieu Mangeot, Gilles Sérasset, and Mathieu Lafourcade. 2004. Construction collaborative d'une base lexicale multilingue. *Traitement Automatique des Langues*, 44(2):151–176, February.

Mathieu Mangeot. 2001. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Septembre.

Igor Mel'čuk, Andre Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Universites francophones et champs linguistiques. AUPELF-UREF et Duculot, Louvain-la Neuve.

Alain Polguère. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for french. In *Proceeding of EURALEX'2000, Stuttgart*, pages 517–527.

Mutsuko Tomokiyo, Mathieu Mangeot, and Emmanuel Planas. 2000. Papillon: a project of lexical database for english, french and japanese, using interlingual links. In *JST'00 Journées Science et Technologie*, page 3, National Olympic Memorial Youth Center, Tokyo, Japon, 13-14 novembre.

[4] http://www.oceandictionary.net/