Papillon project: Retrospective and Perspectives

Mathieu Mangeot

Condillac-LISTIC
F-73376 Le Bourget du Lac Cedex
Mathieu.Mangeot@univ-savoie.fr

Abstract

This paper describes the first five years of life of the Papillon project with four main phases: the birth with the motivations of such a project; the extension with the decision to build a multilingual pivot dictionary; the implementation with the realization of "Jibiki", a generic dictionary management platform and the population with the use of semantic vectors for linking entries and an ongoing project: word games, for creating specific lexical information.

1. Introduction

This paper describes the first five years of life of the Papillon project which goal is to build a multilingual pivot dictionary with a rich microstructure. The idea is that everyone can contribute online to the dictionary. The resulting data is freely and publicly available.
The paper is divided in four sections, one for each phase of the project in historic order: the birth, the extension, the implementation and the population.

2. Phase I, birth: a French-Japanese bilingual dictionary

2.1. History

The Papillon project (first named FeJ for French-English-Japanese) (Boitet et al., 2002; Mangeot et al., 2004) was launched in early 2000 by Emmanuel Planas, François Brown de Colstoun and Mutsuko Tomokiyo. Emmanuel Planas was postdoc researcher at NTT Research Centre, located in Keihanna, Japan and François Brown de Colstoun was scientific attaché at the embassy of France in Tokyo, Japan. Mutsuko Tomokiyo was a linguist Ph.D. student in Grenoble, France.
They were confronted every day with the needs of a good French-Japanese dictionary. That was the starting point of the project.
The first institutional partners were the home institutions of the initiators: the GETA-CLIPS laboratory in Grenoble, France and the Embassy of France in Japan. The National Institute of Informatics (NII) also joined the project through contacts with NII researchers.

2.2. Motivations

The first motivations of the project were the following:

- **Few resources** The main problem is the lack of free and good French-Japanese dictionaries. The few complete French-Japanese resources are expensive, and tailored for Japanese speakers. The free lexicons available on the Web are very insufficient even for simple vocabulary (10,000 entries). Thus, the majority of French speakers have no choice but using English-Japanese dictionaries. This is also true for many other languages. Even for those with a good knowledge of English, it automatically adds confusion.
- **Lack of information** The most complete French-Japanese dictionaries were built for Japanese speakers, thus there is a lack of information necessary for French speakers: transliteration of kanji, numerical specifiers, etc.
- **High construction costs** The traditional way of building a dictionary needs lots of money and time. As an example, the construction of the EDR English-Japanese dictionary cost 1,200 human-year for about 300,000 entries in each language. The public price, 14,3 millions of yens (~100,000 €) is so expensive that only companies can afford it. Furthermore, it does not even reflect the construction costs. The initiators had no choice but finding another way to build their dictionary.
- **Collaborative projects** An interesting way seems to launch a collaborative project like the LINUX construction paradigm. People contribute at their level. The result is free of rights and free so that every can benefit from it. At that time, there were already dictionaries building projects that were using this method, like the Edict Japanese-English dictionary project launched and still managed by Jim Breen for more than ten years. Now, the success of the Wikipedia project confirms our idea.

2.3. Meetings

The initiators had a user point of view of the dictionary. They were not specialists of computational lexicography. They decided to ask other researchers (mainly from GETA-CLIPS) to join the project and the decision was taken to hold the first Papillon meeting (Tomokiyo et al., 2000) at the National Institute of Informatics, Tokyo, Japan in August 2000.
Since then, we decided to organize a meeting every year. The 2001 meeting took place in July in Grenoble. The dictionary structures (Sérasset and Mangeot, 2001) were adopted during this meeting.
The 2002 meeting took place in July in Tokyo. We took there important decisions concerning the data built in the framework of the project: it is free of rights and freely and publicly available. In order to ensure a long life to the Papillon project, we organized our way of working in a way that

it would not depend on any specific founds. The scientific leaders are university researchers with a full time position. The project advances also thanks to Ph.D. fellows or post-doctorate researchers whose subject integrates a scientific issue of the project. We decided also to organize every meeting as a workshop with scientific reviewing committee in parallel with an international conference so that it would be easier for researchers to obtain founds for coming.

The 2003 meeting took place in July in Sapporo. We discussed mainly about the platform used for building the dictionary. The 2004 meeting took place in August in Grenoble, France. The 2006 meeting took place in Chiang Rai, Thailand.

Every meeting gathers about roughly 50 people from all parts of the world. Nowadays, the main actors are Christian Boitet, Gilles Sérasset and Mutsuko Tomokiyo from GETA-CLIPS, Grenoble, France; Mathieu Lafourcade from LIRMM, Montpellier, France, Michael Zock from LIF, Marseille, France; Yves Lepage from ATR, Keihanna, Japan; Asanee KAwtrakul from Kasetsart U., Bangkok, Thailand; Jim Breen from Monash U., Melbourne Australia and myself ;-).

3. Phase II, extension: a multilingual pivot dictionary

3.1. History

The first idea was to build a multi-target French to English and Japanese dictionary following the model of the FeM French-English-Malay dictionary. But then, the research conducted at GETA-CLIPS on pivot structures and the opportunity to open the project to many languages led us to decide to build a multilingual pivot dictionary. In the same way, we decided to use an entry microstructure based on the word sense level and very detailed in order for the dictionary to be used both by humans and by machines.

3.2. Macrostructure

The multilingual pivot macrostructure with interlingual links is based on Gilles Sérasset's Ph.D. Thesis(Sérasset, 1994b; Sérasset, 1994a; Sérasset, 1994c) and has been experimented at a small scale by Etienne Blanc (Blanc, 1995) with the PARAX database.

This structure consists in one monolingual volume for every language of the dictionary and one pivot volume in the middle see Figure 1.

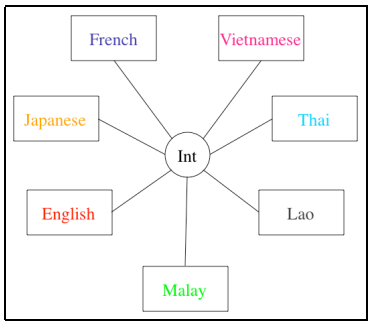

Figure 1: Multilingual Pivot Macrostructure

The monolingual volumes gathers monolingual entries at a word sense level, i.e. monolingual acceptions (called lexies). The entries of different languages are then linked between each others via interlingual acceptions (called axies) that can be seen as complex translation links. These acceptions may also be linked together by refinement links in order to cope with the semantic discrepancies between languages.

Each sense or meaning of each entry of a monolingual volume is linked to one or more acceptions of the pivot volume. For example, like in figure 2 in French “ affection ” has two meanings: “affection” and “disease”. The vocable “affection” will consequently be linked to two "lexies" (corresponding to two word senses) in the French monolingual dictionary, which in turn will be linked to two interlingual acception or "axies" in the pivot volume.

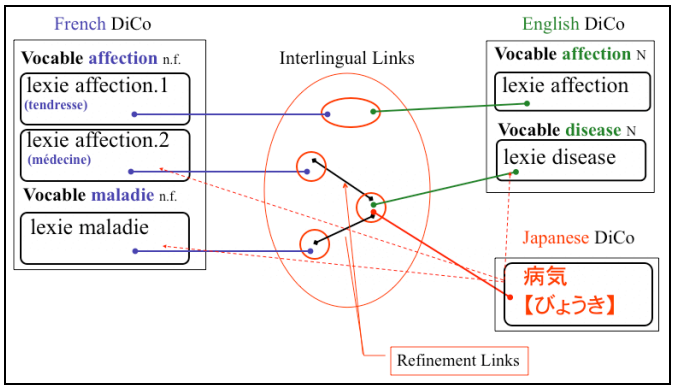

Figure 2: Macrostructure in Detail with Interlingual Links

3.3. Microstructure

The structure of the entries or microstructure of the monolingual volumes is based on the structure used for the formal lexical database DiCo (Polguère, 2000) of the OLST laboratory in Université de Montréal. The encoding methodology is directly borrowed from the Explanatory and Combinatorial Lexicology (ECL)(Mel'čuk et al., 1995), which is part of the Meaning-Text Theory elaborated by Igor Melčuk and his colleagues first in Moscow, Russia and then in Montreal, Canada.

This structure, rather complex (see Figure 3) has been chosen for mainly two reasons:

1. It has been proven language independent and thus, appropriate for any of our languages present in our dictionary. Of course, there are some parts that are language dependent such as the grammatical properties or the language levels, but the main part remains the same.

2. It has been elaborated to be theoretically used both by humans or machines.

Each lexie or lexical unit is made of a name, grammatical properties (mainly a part of speech), a semantic formula which can be seen as a formal definition. In the case of a predicative lexie, it describes the entire predicate and its

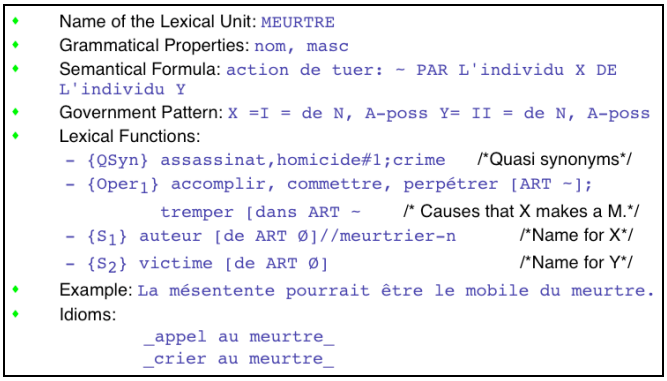

Figure 3: Microstructure of the French lexie "MEURTRE"

arguments, a government pattern which describes the syntactic realization of the arguments of the predicate, a list of lexico-semantic functions. There is a fixed number of 56 basic functions that can be applied in any language. These functions can be combined to create more elaborated ones; a list of examples; a list of full idioms.

4. Phase III, implementation: an online generic dictionary management platform

4.1. History

I began my Ph. D. (Mangeot, 2001) in 1998 taking the results of Gilles Sérasset's Ph.D. (Sérasset, 1994c) Ph.D. as a starting point and having the goal to implement a demonstrator. I implemented a first prototype in Perl called DicoWeb. It was able to query several dictionaries with different structures and display the results in the same window (this tool is still used daily at XRCE laboratory).
After the first Papillon meeting in July 2000, Gilles Sérasset and I began to implement a more robust prototype in Java based on the specifications described in my Ph.D. thesis with the goal to obtain a generic platform for managing (querying, editing, importing, exporting) dictionaries in any structure.
In order to follow the LINUX construction paradigm not only for the data but also for the software, we chose to use only free open-source software for building the platform. Furthermore, we plan to release it in the future as a sourceforge project.

4.2. The Jibiki Platform

The Jibiki platform[1] (Mangeot and Sérasset, 2002), (Sérasset, 2004) is a community web site primarily developed for the Papillon project. This platform is entirely written in Java using the "Enhydra[2]" web development Framework. All XML data is stored in a standard relational database (Postgres). This community web site proposes several services:

- a unified interface to simultaneously access the Papillon MLDB and several other monolingual and bilingual dictionaries;
- a specific edition interface to contribute to the dictionaries stored on the platform,
- an open document repository where registered users may share writings related to the project; among these documents, one may find all the papers presented in the different Papillon workshops organized each year by the project partners;
- a mailing list archive,

To encourage volunteers, we think that it is important to give a real service to attract as many Internet users as possible. As a result, we began our development with a service to allow users to access to many dictionaries with different structures but in a unified way (see Figure 4). This service currently gives access to thirteen (13) multilingual, bilingual and monolingual dictionaries, representing more than one million entries.

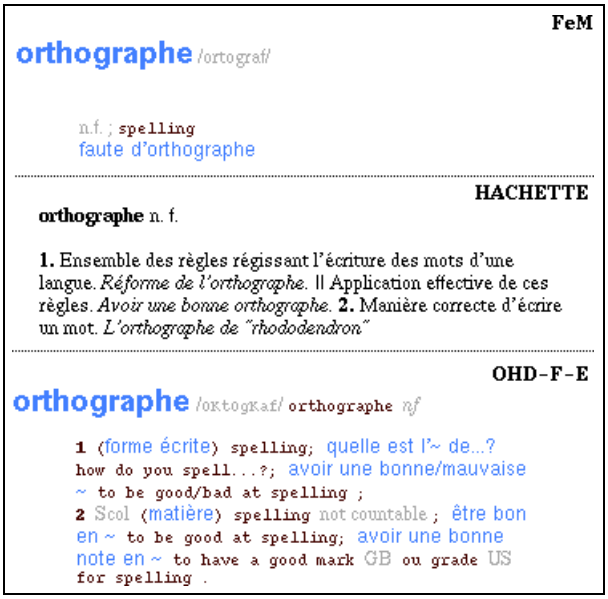

Figure 4: Query of "Orthographe" in three dictionaries

Every available dictionary will be queried according to its own structure from a multi-criteria search interface (see 4.2.). Moreover, all results will be displayed in a form that fits the structure. Any monolingual, bilingual or multilingual dictionary may be added in this collection, provided that it is available in XML format. With the Jibiki platform, giving access to a new, unknown, dictionary is a matter of writing two XML files: a dictionary description and an XSL stylesheet. For currently available dictionaries, this took an average of about one hour per dictionary.
The description file gathers dictionary meta-information and a minimum set of information in the dictionary's XML structure. The Jibiki platform defines a standard structure of an abstract dictionary containing the most frequent subset of information found in most dictionaries. This abstract structure is called the Common Dictionary Markup (Mangeot, 2002). To describe a new dictionary, one has to write an XML file that associate CDM elements to pointers in the original dictionary structure.

[1] see http://jibiki.univ-savoie.fr/jibiki
[2] see http://www.enhydra.org/

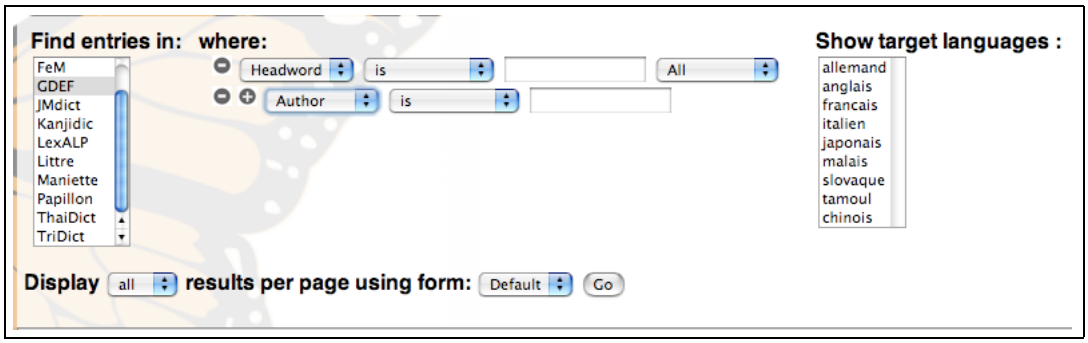

Figure 5: Multicriteria Advanced Search Interface in Several Dictionaries

Along with this description, one has to define an XSL style sheet that will be applied on requested dictionary elements to produce the HTML code that defines the final form of the result. If such a style sheet is not provided, the Jibiki platform will itself transform the dictionary structure into a CDM structure and apply a generic style sheet on this structure.

4.3. The key feature: an online generic editor

The main purpose of the Jibiki platform is to gather a community around the development of one or several dictionaries. Thus, the crucial challenge that we faced was to provide a way to edit the dictionary entries directly on the platform. It was specifically difficult because we wanted to be able to edit any kind of dictionary entry (the editor had to adapt itself to the structure of the entries) and to edit them online with a simple browser (it had to be bult only with a combination of HTML forms and simple javascripts). We did not even want to use java applets because of compatibility problems.

A preliminary version of the editor (Mangeot and Thevenin, 2004) was developed in collaboration with David Thevenin with his tool called ARTStudio for the development of adaptative plastic user interfaces. It was fragile and very difficult to handle. Furthermore, some parts of the code were not open source. Thus, a new simplified version has been recoded from scratch afterwards.

The new editor works with a template XHTML interface that is instanciated with the entry that the user wants to edit. This template can be generated automatically from a description of the entry structure in XML schema. It can be modified afterwards for improving the rendering on the screen. Thus, the only data needed to edit a dictionary entry on the jibiki platform (apart from the dictionary metadata described previously) is the XML schema of the structure of the entry and furthermore, any type of dictionary entry as long as it is encoded in XML.

We chose to use XML schema because it allows for a finer description compared to DTDs (for instance, we may define the set of valid values of the textual content of an XML element). Moreover XML schemata provides a simple inheritance mechanism that is useful for the definition of a dictionary.

HTML forms are very limited. The available interactors are text fields, radio buttons, check boxes and pop up menus. It was not enough to be able to edit complex entries. Thus, we had to build more complex interactors from the combination of the previous ones in order to handle lists (adding,deleting, moving an item on a list) and links (links to entries in the same volume or other ones). These elements can be themselves complex objects containing lists of other objects, etc.

Any user, who is registered and logged in to the Papillon web site, may contribute to the Papillon dictionary by creating or editing an entry. Moreover, when a user asks for an unknown word, he is encouraged to contribute it to the dictionary. Contribution is made through a standard HTML interface (see Figure 6).

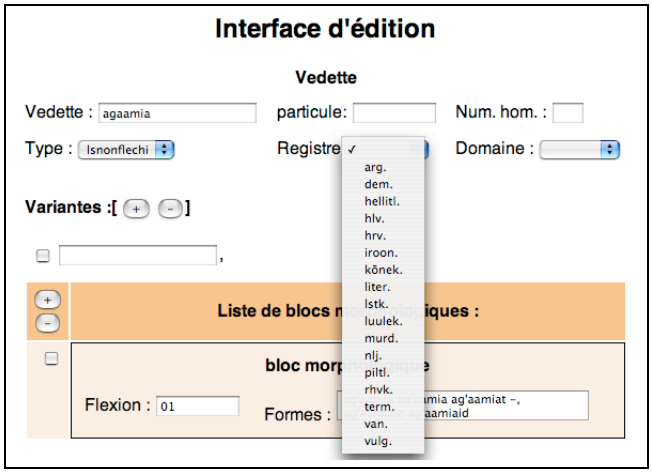

Figure 6: Interface for Writing an Entry

Every change made in the entry is stored in a history. It is then possible to come back to any previous version of the entry just like the usual "undo" commands. The writing process is divided in several steps depending on the project. The GDEF is the most complete with three steps:

- A contributor writes an entry;
- It is next revised by a reviewer;
- It is then validated by a validator;

4.4. Conclusion

The platform is now used by four different projects:

- the Papillon[3] project,
- the GDEF[4] project (Chalvin and Mangeot, 2006), about a bilingual Estonian-French dictionary,
- the LexALP[5] project (Sérasset, 2005), about a multilingual (English, French, German, Italian and Slovene) terminological database on the legal terms of the alpine convention,
- the TriDict trilingual (Sinhala, Tamil, English) dictionary.

There are still lots of ongoing developments on the platform with still a perspective of genericity in the different resources handled.
For those who want to use the platform for their projects, we are open to any collaboration. The only condition is that all the data produced with the platform must be publicly available and free of rights.

5. Phase IV, population: semantic vectors and word games

In order to facilitate the construction of Papillon dictionary, we decided to reuse existing data. The hypothesis are that it is easier to correct existing data than to build new data from scratch and that the public users prefer to have slightly incorrect data that no data at all when they lookup words in a dictionary.
The population faces two serious issues: the building of interlingual links between the lexies and the specific lexical information that is not available in any dictionary. We decided to use semantic vectors for the first issues and word games for the second one.

5.1. Semantic Vectors

The first problem is augmented by the fact that we chose to work at the word sense level, not at the vocable level. There is no unique way to divide a word into senses. In two dictionaries of the same language, for many entries, the division into word senses will be different. Thus, when one wants to merge the entries of two different dictionaries at the word sense level, s/he has to find a way to cope with this problem.
The solution we found uses semantic vectors in order to calculate the semantic distance between two lexies of two different dictionaries we want to merge and to determine if they can be merged or not.
The conceptual vectors model has been presented in (Lafourcade, January 2001; Lafourcade et al., 2002). Each textual segment (word, phrase, text) is linked with a thematic association that is represented by a vector of concepts. The set of concepts is predefined and constitutes a multidimensional vector space on which the word senses can be projected.

In this vector framework, it is possible to use the notion of *similarity* (usually used in information retrieval) and *angular distance* between two vectors. It will be used as an evaluation of the *thematic distance* beween word senses.

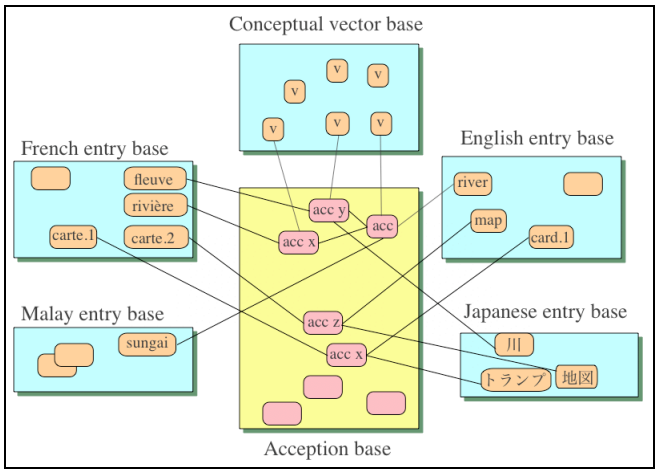

Figure 7: Linking Acceptions with Vectors

In order to merge lexies coming from different dictionaries, the first step is to calculate the conceptual vector that is linked to each of the lexies. For example, in French, the set of concepts is predefined with the 873 concepts of the Larousse thesaurus.
The second step is to bootstrap the computation by manually indexing 5,000 terms in each language.
Then the definition of each lexie is analyzed with a morphological analyzer. Then, using the manual indexed vectors of known words and the resulting analysis tree, we compute the vectors associated to each lexie and word-form. The process is reiterated until a stability is reached.
Once the process is finished, the dictionary is "vectorized". It is then possible to merge two dictionaries of the same language by looking at the thematic distance of the conceptual vectors of each lexie.
We consider that two conceptual vectors are close enough if their thematic distance is less than a threshold t. The more the threshold is low, the more the lexies can be considered as being merged. Nevertheless, it may be difficult to merge completely automatically the lexies. An acceptable value for the threshold is $\pi/4$.

5.2. Word Games

The issue is to find methods for building some particular crucial lexical data which is furthermore not available in existing dictionaries. It is the case for collocations coded in our entries through lexical functions.
For example, in English, the notion of "fever" is intensified by the adjective "strong, the notion of "smoker" by the adjective "heavy", etc. or, more particularly for asian languages, special counters must be used for specific types of objects. In Japanese, "wa" is the counter for the rabbits (usagi san wa, 3 rabbits) and "hiki" is the counter for cats (neko ni hiki, 2 cats).
The goal of the project "jeu de mots" (word game) is to experiment and study the use of "word games" for building or

[3] see http://www.papillon-dictionary.org
[4] see http://www.estfra.ee
[5] LexALP: Legal Language Harmonisation System for Environment and Spatial Planning within the Multilingual Alps

collecting precise lexical information. The idea is to generate automatically or semi-automatically word games tat can take the shape of a multiple-choice test (e.g. Is it possible to say ... in English ?) or fill-In-the-blank exercises (complete "strong fever, heavy smoker, _____ rain")

Each generated exercise will be used to complete or validate the information available in the dictionaries. Te targeted languages in this project framework are Chinese, French, Japanese, Malay and Thai. The exercises will be submitted to students and web surfers, (via the Papillon website) who will work on their mother tongue. The answers collected will be analyzed and the method will be tested and evaluated on each language. The gathered information will be publicly available on the Papillon website.

This project has been accepted and funded by the French government under the STIC-Asia program driven by INRIA research organization.

6. Conclusion

We presented a very challenging project that is already six years old and has already produced interesting results theoretically with research on multilingual pivot structures and practically through "jibiki", an online generic dictionary management platform.

We are welcoming anybody who is motivated by the project and wants to join the project. It is mainly based on voluntary work and aims to build a reference lexical resource.

7. References

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.

Christian Boitet, Mathieu Mangeot, and Gilles Sérasset. 2002. The papillon project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. In Graham Wilcock, Nancy Ide, and Laurent Romary, editors, *Proc. of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop*, pages 93--96, Taipei, Taiwan, 1 September.

Antoine Chalvin and Mathieu Mangeot. 2006. Méthodes et outils pour la lexicographie bilingue en ligne : le cas du grand dictionnaire estonien-français. In *EURALEX 2006, à paraître*, Turin, Italie, 6-9 septembre.

Mathieu Lafourcade, Violaine Prince, and Didier Schwab. 2002. Vecteurs conceptuels et structuration émergente de terminologies. *TAL*, 43(1):43--72.

Mathieu Lafourcade. January 2001. Lexical sorting and lexical transfer by conceptual vectors. In *First International Workshop on MultiMedia Annotation (MMA'2001)*, page 6, Tokyo.

Mathieu Mangeot and Gilles Sérasset. 2002. Frameworks, implementation and open problems for the collaborative building of a multilingual lexical database. In Grace Ngai, Pascale Fung, and Kenneth W. Church, editors, *Proc. of SEMANET Workshop, Post COLING 2002 Workshop*, pages 9--15, Taipei, Taiwan, 31 August.

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.

Mathieu Mangeot. 2002. An xml markup language framework for lexical databases environments: the dictionary markup language. In *LREC Workshop on International Standards of Terminology and Language Resources Management*, pages 37--44, Las Palmas, Spain, 28 May.

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.

Gilles Sérasset and Mathieu Mangeot. 2001. Papillon lexical database project: Monolingual dictionaries and interlingual links. In *NLPRS-2001*, pages 119--125, Tokyo, 27-30 November.

Gilles Sérasset. 1994a. An interlingual lexical organisation based on acceptions, from the parax mock-up to the nadia system. In *ICLA-94*, pages 21--33, July.

Gilles Sérasset. 1994b. Interlingual lexical organisation for multilingual lexical databases in nadia. In Makoto Nagao, editor, *COLING-94*, volume 1, pages 278--282, August.

Gilles Sérasset. 1994c. *Sublim : un système universel de bases lexicales multilingues et Nadia : sa spécialisation aux bases lexicales interlingues par acceptions*. Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, Décembre.

Gilles Sérasset. 2004. A generic collaborative platform for multilingual lexical database development. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources Workshop*, pages 73--79, Geneva, Switzerland, 28 August.

Gilles Sérasset. 2005. Multilingual legal terminology on the jibiki platform: The lexalp project. In Mathieu Lafourcade, editor, *Proc. of Papillon 2005 Workshop*, pages 64--73, Chiang Rai, Thailand, 11-13 December.

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.