Frameworks, Implementation and Open Problems for the Collaborative Building of a Multilingual Lexical Database

Mathieu MANGEOT-LEREBOURS
Software Research Division, NII
Hitotsubashi, 2-1-2-1913 Chiyoda-ku
101-8430 Tokyo, Japan
mangeot@nii.ac.jp

Gilles SÉRASSET
GETA-CLIPS-IMAG
185, rue de la bibliothèque, BP 53
F-38041 GRENOBLE CEDEX 9, France
Gilles.Serasset@imag.fr

Abstract

Many NLP systems are based on lexical data. The development costs of such data are a major drawback in such NLP systems. In order to cut these costs, we adopt a strategy inspired from "open-source" projects to allow volunteers to collaborate in the creation of a multilingual lexical database.

For this, we had to specify and develop tools to manage a lexical database containing information complete and detailed enough to be usable for a wide range of applications.

This paper presents our project and details the tools, frameworks and structures used to manage such a database. We will also show some research problems still to be addressed in this context.

Résumé

La connaissance linguistique reste une constituante importante de nombreux systèmes de traitement automatique des langues (TAL). Le coût de création d'un dictionnaire est l'un des freins majeurs dans le développement de ces systèmes.

Afin de réduire les coûts de création de cette connaissance lexicale, nous adoptons une méthode inspirée des projets "open-source" afin de créer une base lexicale multilingue.

Pour cela, nous avons spécifié et développé des outils de gestion d' une base lexicale contenant des informations suffisamment complètes et détaillées pour êtres utilisées dans de nombreuses applications différentes.

Cet article présente notre projet et détaille les outils, les cadres et les structures utilisées pour la gestion de cette base. Nous montrons aussi certains problèmes de recherche ouverts qu' il nous faut aborder dans ce contexte.

Introduction

Many NLP systems are based on lexical data. The development costs of such data are a major drawback in such NLP systems. Furthermore, the existing lexical data have generally been developed for a specific purpose and can't be reused easily in other applications.

The Papillon project applies some tools and methods to develop multipurpose, multilingual lexical data collaboratively on Internet. This data is complete and detailed enough to be eventually used either by NLP systems (MT engines for example) or by human users (language learners, translators…).

After presenting the motivations of the Papillon project, we will show the management of existing data. Then we will describe the structure of the Papillon dictionary, and the tools that are used to allow contributions from Internet volunteers.

1 The Papillon Project

1.1 Motivations

The Papillon project is the result of the gathering of different people sharing common problems and solutions.

1.1.1 A Lack of Resources

On the Internet, a lot of free dictionaries are available but very few of them imply more than 2 languages. Most of these dictionaries include English as one of their languages.
Furthermore, the existing dictionaries often lack information essential for beginners or NLP systems.
Another point contributing to this lack: the high costs of development of large lexical resources for NLP involves also a high price, dissuasive for the end-user.

1.1.2 Existing Structures and Tools for Multilingual Dictionaries

Some partners of the Papillon project have been involved in research on the definition of structures and tools to handle multilingual lexical databases.
They were looking for an opportunity to apply their research results on real scale lexical data.

1.1.3 Collaborative Development on the Internet

Most partners were participating, as computer scientists, in the development of open source products. With the democratisation of Internet access in a lot of countries, came the opportunity to apply the open source principles to the development of a multipurpose, multilingual lexical database.
Cooperation projects for bilingual dictionaries are already going on such as EDICT, a Japanese-English dictionary lead by Jim Breen (2001) for more than 10 years and more recently, SAIKAM, a Japanese-Thai dictionary (see Ampornaramveth (2000)).
With the Papillon project, the dictionary is extended to a multilingual lexical database. Volunteers will find lexicons developed by others and some tools to complete or correct the Papillon multilingual dictionary. Users will also be able to define their own personal views of the database.

1.2 Dictionary Markup Language Framework

Mathieu Mangeot-Lerebours (2001) defines a complete framework for the consultation and the construction of dictionaries. The framework is completely generic in order to manage heterogeneous dictionaries with their own proper structures. This framework is extensively used in Papillon project.

1.2.1 Dictionary Markup Language (DML)

The framework consists in the definition of an XML namespace[1] called DML (Dictionary Markup Language). All lexical data of a lexical database can be described with DML elements. The entire hierarchy of the XML files, elements and attributes is described using XML schemata and grouped into the DML namespace. Figure 1 describes the organisation of the main DML elements.

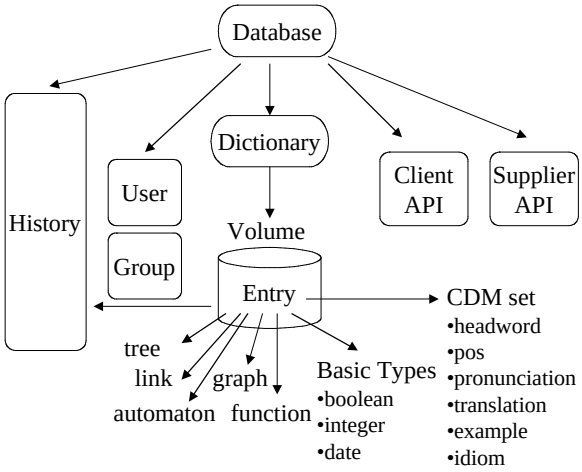

Figure 1. The DML Framework

The XML schemata are available online. This allows users to edit and validate their files online with an XML schema validator.

1.2.2 Common Dictionary Markup (CDM)

The DML framework may be used to encode many different dictionary structures. Indeed, two dictionary structures can be radically different. In order to handle such heterogeneous structures with the same tools, we have defined a subset of DML element and attributes that are used to identify which part of the different structures

[1] http://www-clips.imag.fr/geta/services/dml

represent the same lexical information. This subset is called Common Dictionary Markup (CDM). This set is in constant evolution. If the same kind of information is found in several dictionaries then a new element representing this piece of information is added to the CDM set. It allows tools to have access to common information in heterogeneous dictionaries by way of pointers into the structures of the dictionaries.

1.3 Three Layers for the Lexical Data

The lexical data repository of the Papillon project is divided into 4 subdirectories:

- *Administration* contains guidelines and administrative files
- *Hell* (data in original format)
- *Purgatory* (data in XML & UTF-8)
- *Paradise* (data in Papillon format)

The name of the files and directories is normalised in order to allow easy navigation into the repository.
All lexical data stored in the repository is free of rights or protected by a GPL-like licence.

1.3.1 Hell Directory

This directory contains lexical data in their original format. When a dictionary is received, it is first stored there while waiting to be "recycled". For each dictionary, we create a metadata file containing all available information concerning the dictionary (name, languages covered, creation date, size, authors, domain, etc.). It is then used to evaluate the quality of the dictionary and to guide the recycling process. These dicitonaries are freely downloadable as they are.

1.3.2 Purgatory Directory

The Purgatory directory receives the lexical data once the recuperation process is over. This process consists in converting the lexical data from its original format into XML encoded in UTF-8. To perform this task, we use the RECUPDIC methodology described in Doan-Nguyen (1998) regular expression tools like Perl scripts.

If a dictionary is already encoded in XML, the recuperation process consists in mapping the elements of information into CDM elements and storing the correspondence into the metadata file.

Internet users access these dictionaries as classical online dictionaries, retrieving individual entries by way of requests on the Papillon web site.

1.3.3 Paradise Directory

The *Paradise* directory contains only one dictionary often called the "Papillon dictionary". This dictionary has a particular DML structure. Internet users access entries of this dictionary by way of requests to the Papillon web site.
It is possible to retrieve only one entry, or any subset of entries in any available output format. The "native" format is the Papillon textual XML DML format in UTF-8. Users also have ways to add new entries or correct existing ones online.
Other *purgatory* dictionaries may be integrated into the Papillon dictionary with the help of the CDM elements.

2 The Papillon Multilingual Dictionary

2.1 Macrostructure

The architecture of the Papillon multilingual dictionary is based on Gilles Sérasset (1994) and has been prototyped by Blanc (1999). This architecture uses a pivot structure based on multiple monolingual volumes linked to an interlingual acception volume.

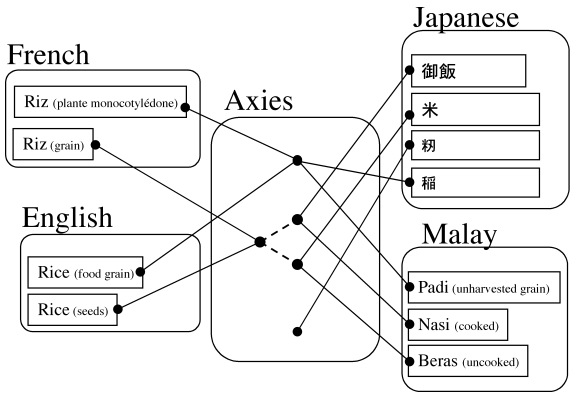

Figure 2. Illustration of Papillon's macrostructure.

Each entry of a monolingual volume represents a word sense. In this document, we use the term of "lexie" as in the Explanatory and Combinatory Dictionary to name a monolingual entry. The meaning of "lexie" is not the same as "lexeme". A lexie is a complete monolingual entry.

The interlingual volume gathers all the interlingual acceptions. An interlingual acception represents the union of word-senses or "lexies" considered as "equivalent" among different monolingual volumes. This equivalence is calculated from translation links. In this document, we use the term of "axie" to name an interlingual acception.

Real contrastive problems in lexical equivalence (not to be confused with monolingual polysemy, homonymy or synonymy as clearly explained in Mel'cuk and Wanner (2001) are handled by way of a special kind of link between axies. Figure 2 illustrates this architecture using a classical example involving "Rice" in 4 languages. In this example, we used the word senses as given by the "Petit Robert" dictionary for French and the "Longman Dictionary of Contemporary English" for English. As shown, the French and English dictionaries do not make any word sense distinction between cooked and uncooked rice seeds. However, this distinction is clearly made in Japanese and Malay. No axie may be used to denote the union of the word senses for Malay "nasi" and "beras" unless we want to consider them as true synonyms in Malay (which would be false). Hence, we have to create 3 different axies: one for the union of "nasi" and 御飯 (gohan), the other for the union of "beras" and 米 (kome) and one for the union of "rice" and "riz". A link (non-continuous line in Figure 1 has to be added between the third axies and the others in order to keep the translation equivalence between the word-senses.

Note that the links between axies do not bear any particular semantics and should not be confused with some kind of ontological links.

Bilingual dictionaries can be obtained from the multilingual dictionary.

2.2 Microstructure

The structure of the lexies (units of the monolingual dictionaries) is based on Polguère (2000) and Mel'cuk's work on the combinatorial and explanatory lexicography, a part of the meaning-text theory. An XML schema using the DML framework has been defined to represent this structure as accurately as possible.

```
<lexie xmlns="http://www-clips.imag.fr/geta/services/dml"
       xmlns:d="http://www-clips.imag.fr/geta/services/dml"
       xmlns:xlink="http://www.w3.org/1999/xlink"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       basic="true" d:id="meurtre$1" frequency="0.3"
       name="Papillon-fra" source-language="fra"
       ...>
<headword hn="1">meurtre</headword>
  <pronunciation encoding="GETA">meu+rtr(e)</pronunciation>
  <pos>n.m.</pos>
  <semantic-formula>action de tuer: ~ PAR L'
    <sem-label>individu</sem-label><actor>X</actor> DE L'
    <sem-label>individu</sem-label><actor>Y</actor>
  </semantic-formula>
  <government-pattern>
   <mod nb="1">
    <actor>
      <sem-actant>X</sem-actant><synt-actant>I</synt-actant>
      <surface-group>
       <surface>de N</surface>
       <surface>A-poss</surface></surface-group></actor>
    <actor>
     <sem-actant>Y</sem-actant><synt-actant>II</synt-actant>
      <surface-group>
       <surface>de N</surface>
       <surface>A-poss</surface></surface-group></actor>
   </mod></government-pattern>
  <lexical-functions>
   <function name="Qsyn">
     <valgroup>
      <value>
       <reflexie xlink:href="#assassinat$1">assassinat
       </reflexie></value>
     <value>
      <reflexie xlink:href="#homicide$2">homicide</reflexie>
     </value><value>
      <reflexie xlink:href="#crime$1">crime</reflexie>
     </value></valgroup></function>
   <function name="V0">
    <valgroup>
     <value><reflexie xlink:href="#tuer$1">tuer</reflexie>
     </value></valgroup></function>
    ...
  </lexical-functions>
  <examples>
    <example d:id="#meurtre$1-e1">
      C'est ici que le double meurtre a été commis.</example>
    ...</examples>
  ...
</lexie>
```

Figure 3. XML encoding of the French entry "meurtre" (excerpt)

This structure is common to all the monolingual dictionaries. In order to cope with language

differences, small variations are authorised for each monolingual lexicon. Up to now, these variations have been used to define the parts of speech for each language and to add information specific to each language, such as level of politeness and counters for Japanese.
Figure 3 presents an excerpt of the XML encoding of the French entry "meurtre" (murder) and Figure 4 shows a DEC-like view.
The general schema has been presented in detail in Gilles Sérasset & Mathieu Mangeot-Lerebours (2001).

3 Implementation of the Collaborative Web Site

For the external user, the Papillon project is viewed as a dynamic web site providing access the existing dictionaries and giving ways to contribute to the Papillon dictionary.

3.1 General Architecture

The Papillon web site is built with a Java based open source framework called Enhydra[2]. It is designed around a standard 3-tier architecture

- a presentation layer in charge of the interface with the user. We currently use classical HTML/CSS rendering, but plan to integrate WML access to the dictionaries (for mobile phones),
- a business layer in charge of data manipulation and transformation. We currently use XML data (in UTF-8) and XSL transformations for data manipulation,
- a data layer in charge of the communication with the database via a JDBC driver. The data layer should be managed by an XML database allowing language dependent sorting. For the moment, XML databases are still in an early stage. In order to advance in the project, a mapping system for DML has been defined in order to store the XML data into conventional relational databases. PostgreSQL is used at this point.

3.2 Particular features

As different users may have different needs (translators, learners…) we define different views of the Papillon dictionary. Each view is encoded as a XSL stylesheet that is applied on the result of each user query. In the future, we will also allow users to define their own custom views and store them on the server. All these transformations are done on the server in order to allow users to use their preferred browser (even if it is not XML aware). Figure 4 shows an example of the French entry "MEURTRE" (murder) viewed as in Mel'cuk's DEC dictionary.

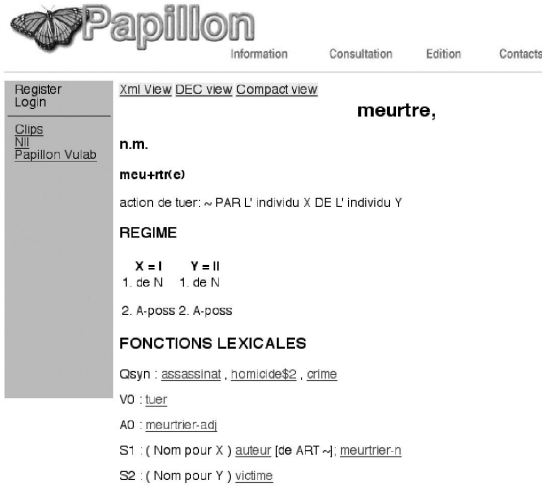

Figure 4. French entry "meurtre" dynamically displayed using Mel'cuk's classical view

To avoid the unintentional pollution of the database by erroneous data, the contributions of a user are to be validated by a central group of trusted users. In the mean time, the contributions are stored as XSL stylesheets in the cntributor's private space.
Each time a user requests a corresponding entry, the request is performed in the main database and in the user space. The results from the user space are used to modify results from the main database. This way, the contribution is immediately visible to the user exactly as if it had been integrated into the main database.
While contributions are waiting to be validated and integrated into the common space, The contributors may choose to share them with other users or groups of users.
Every user can contribute at his/her level. For example, a linguist specialist of lexical functions will enter values of lexico-semantic functions, a phonologist pronunciations and a professional bilingual translator will enter new interlingual links or check the semi-automatically generated

[2] available at www.enhydra.org

ones. For this, different interfaces will be developed to accommodate the various user profiles.

3.3 Annex Tools

As the web site hosts a rather complex collaborative work, we have added some tools that are not related to lexicography, but that have to work in a multilingual context.
First, there is a tool to archive our Papillon mailing list. Such a tool is very common on Internet sites. However, as we found out, these tools may not be used in our multilingual context, where mails may contain discussion in different languages, written with different tools, and encoded using different standards. Hence we patched an existing tool so that it archives all mail in UTF-8, regardless of its original encoding.
To avoid the considerable work of the webmaster and to facilitate the communication and the exchange of informations between the users of the database, we are developing tools to facilitate the use of a document repository.
After registration and login, users can easily upload online a file in whatever format. It will immediately be stored into the document repository and made accessible online on the web.

4 Actual Research and Development Directions

The Papillon project is a extremely interesting experimentation platform. We are currently working on validation of monolingual data, management of axies and acquisition of new data.

4.1 Validation of the Monolingual Data

A team of trusted lexicographers validates user contributions before they are integrated into the main database.
This validation is a time consuming process and implies a good level in linguistics and lexicography. Moreover, we may not find enough specialists volunteering for such a work and we may have to pay a core team for this.
This task is essential and should be conducted as quickly as possible lest the users will be discouraged by the delays implied by the central team.
Hence, even in this validation process, we wish to enroll users as much as possible. For this task, we plan to implement tools for indirect validation of information using vote mechanisms and generating questions answerable without any special knowledge in linguistics.
As a first experiment, we will use a French generator in order to produce a lot of examples using the word to be validated and a set of known words (already validated). These examples will be presented to native speakers and they will simply have to accept or reject them. This strategy is very interesting in our context, as it will help validating the lexical functions.

4.2 Management of the Interlingual Links

The use of a pivot dictionary to represent translation equivalence is challenging. This macrostructure is very satisfying on a theoretical level, but introduces a high complexity of management.
In Sérasset (1994), we envisaged that these interlingual acceptions would be created and managed by hand by a team of specialists, helped by tools that would detect inconsistencies and propagate decisions among the different languages. This appeared to be unrealistic.
However, we now have means to manage these acceptions automatically. For this, we use the fact that the interlingual acceptions volume does not, in any way, represent a semantic pivot. It is not related to an ontology.
In fact, the only relevant purpose of this interlingual volume is to factorise the bilingual links we find in classical bilingual dictionaries (or the ones that will be specified by the users).
Hence, given a set of translation equivalences between monolingual acceptions of different languages, it is possible to compute a minimal set of acceptions (and their links) that conforms to a set of well-formedness criteria.
One of the difficult tasks is to obtain bilingual translation equivalences between monolingual acceptions when bilingual dictionaries often provide bilingual links between mere lemmas. For this, we will use aligned corpora and translations memories to add contextual information to the translation pairs.

4.3 Acquisition of new data

To depend entirely on volunteer work is of course unrealistic, especially while beginning to build the lexical database. That is why we first reuse existing dictionaries in order to build the kernelof the database.
Contributors will come in later, filling in missing informationin existing entries and creating partial or complete new entries as well as links. However, as we are using a rather complex structure which require some skills that are not shared by all Internet users, we will have to help them help us.
In particular, we are beginning to use corpus-based techniques to extract lemmas that will be candidates as a value of a lexical function. Determining the appropriate lexical function is one of the jobs of our contributors, but they will be helped in this task by tools that will provide them with questions and candidate paraphrases.
For a complement of information or to help the contributors in their task, the database should also propose the consultation of other dictionaries stored locally or available online on the web.
Moreover, to be really useful for the reader, and especially to the learners, the examples found in the dictionaries will be translated in other languages literally and semantically. Some of these translations will be extracted from aligned corpora.

Conclusion

The theoretical frameworks for the whole database, the macrostructure and the microstructure are very well defined. It constitutes a solid basis for the implementation.
A lot of open problems still have to be addressed for the Papillon project to be a success. In this respect, the Papillon project appears to be a very interesting experimentation platform for a lot of NLP research as data acquisition or human access to lexical data, among others.
All this research will improve the attraction of such a project to the Internet users. This attraction is necessary for the project to go on, as it is highly dependent on its users motivations.
This way, we will be able to provide a very interesting multilingual lexical database that we hope useful for a lot of persons.

References

Vutichai Ampornaramveth, Akiko Aizawa, Keizo Oyama & Tanasee Methapisit (2000) *An Internet-Based Collaborative Dictionary Development Project: SAIKAM*., Proc. AdInfo 2000, 9-10 March 2000, NACSIS, Tokyo, Japan, 4 p.

Etienne Blanc (1999) *PARAX-UNL: a Large Scale Hypertextual Multilingual Lexical Database.* Proc. NLPRS 1999, Tsinghua University Press, Beijing, 1999, pp. 507-510.

Jim W. Breen (2001) *A WWW Dictionary and Word Translator: Threat or Aid to Language Acquisition?*, in R Gitsaki-Taylor and P Lewis (eds), Proc. JALT-CALL 2001, Gunma, Japan, 26-27 May 2001, 10 pp.

Haï Doan-Nguyen (1998) *Accumulation of Lexical Sets: Acquisition of Dictionary Resources and Production of New Lexical Sets*. COLING-ACL'98, Montréal, 10-14 August 1998, pp 330-335.

Mathieu Mangeot-Lerebours (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue.* PhD Thesis in Computer Sciences Université Joseph Fourier Grenoble I, 27 September 2001, 280 p.

Igor Melc'uk & Leo Wanner (2001) Towards a Lexicographic Approach to Lexical Transfer in Machine Translation (Illustrated by the German–Russian Language Pair). Machine Translation 16: 21–87, 2001. © 2001 Kluwer Academic Publishers. Printed in the Netherlands.

Alain Polguère (1998) *La théorie Sens-Texte* .Dialangue, Vol. 8-9, Université du Québec à Chicoutimi, pp 9-30.

Alain Polguère (2000) *Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French.* Proc. EURALEX'2000, Stuttgart, pp 517-527.

Gilles Sérasset (1994) *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA.* In Proc. COLING-94, Kyoto, 5-9 August 1994, M. Nagao ed. vol. 1/2 : pp. 278-282.

Gilles Sérasset & Mathieu Mangeot-Lerebours (2001) *Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links.* Proc. NLPRS'2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan, 27-30 November 2001, vol 1/1, pp. 119-125.

Mutsuko Tomokiyo et al. (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 November 2000, Tokyo, Japan, 3 p.