The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons

Christian BOITET[1], Mathieu MANGEOT[2] & Gilles SÉRASSET[1]

(1) GETA, CLIPS, IMAG
385, av. de la bibliothèque, BP 53
F-38041 Grenoble cedex 9, France
Christian.Boitet@imag.fr

(2) National Institute of Informatics (NII)
2-1-2-1314, Hitotsubashi
Chiyoda-ku Tokyo 101-8430, Japan
Mathieu.Mangeot@imag.fr

Abstract

The PAPILLON project aims at creating a cooperative, free, permanent, web-oriented and personalizable environment for the development and the consultation of a multilingual lexical database. The initial motivation is the lack of dictionaries, both for humans and machines, between French and many Asian languages. In particular, although there are large F-J paper usage dictionaries, they are usable only by Japanese literates, as they never contain both original (kanji/kana) and romaji writing. This applies as well to Thai, Vietnamese, Lao, etc.

Introduction

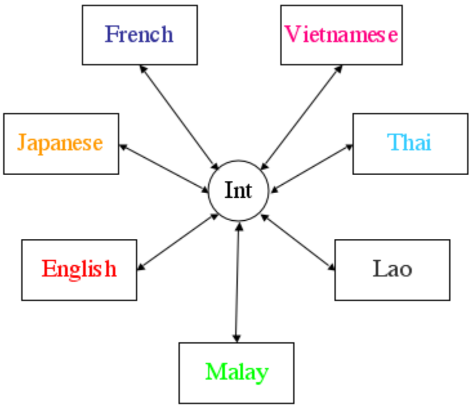
The project was initiated in 2000 and launched with the support of the French Embassy and NII (Tokyo) in July 2000, and took really shape in 2001, with a technical seminar in July 2001 at Grenoble, and concrete work (data gathering, tool building, etc.).

The macrostructure of Papillon is a set of monolingual dictionaries (one for each language) of word senses, called "lexies", linked through a central set of interlingual links, called "axies". This pivot macrostructure has been defined by Sérasset (1994) and experimented by Blanc (1994) in the PARAX mockup.

The microstructure of the monolingual dictionaries is the "DiCo" structure, which is a simplification of Mel'tchuk's (1981;1987;1995) DEC (Explanatory and Combinatory Dictionary) designed by Polguère (2000) & Mel'tchuk to make it possible to construct large, detailed and principled dictionaries in tractable time.

1. Languages included in the project

In 2000, the initial languages of the Papillon project were English, French, Japanese and Thai. Thai was included because there had been a successful project, SAIKAM (Ampornaramveth et al. 1998; 2000), supported by NII and NECTEC, of building a Japanese-Thai lexicon by volunteers on the web. Lao, Vietnamese, and Malay have been added in 2001 because of active interest of labs and individuals.



The star-like macrostructure of Papillon makes it easy to add a new language. Also, the DiCo microstructure of each monolingual dictionary is defined by an XML schema, containing a large common core and a small specialization part (morphosyntactic categories, language usage).

2. Interlingual links

Axies, also called "interlingual acceptions", are not concepts, but simply interlingual links between lexies, motivated by translations found in existing dictionaries or proposed by the contributors.

In case of discrepancies, 1 axie may be linked with lexies of some languages only, e.g. FR(mur#1), EN(wall#1), RU(stena#1), and to other axies by refinement links:

Example:

Axie#234	--lg--> FR(mur#1), EN(wall#1), RU(stena#1)
	--rf--> Axie#235, Axie#236
Axie#235	--lg--> DE(Wand#2), IT(muro#1), ES(muro#1)
Axie#236	--lg--> DE(Mauer#2), IT(parete#1), ES(pared#1)

It is also possible to have 2 axies for the same "concept" at a certain stage in the life of the database, because the monolingual information is not yet detailed enough.

Suppose the level of language (usual, specialized, vulgar, familiar…) is not yet given for FR(maladie#1), FR(affection#2), EN(disease#1), EN(affection#3).

Then we might have, for translational reasons:

Axie#456 --lg-->	FR(maladie#1), EN(disease#1)
Axie#457 --lg-->	FR(affection#2), EN(affection#3)

When this information will be put in each of the above 4 monolingual entries, we may merge the 2 axies and get:

Axie#500 --lg-->	FR(maladie#1, affection#2), EN(disease#1, affection#3)

Axies may also be linked to "external" systems of semantic description. Each axie contains a (possibly empty) list for each such system, and the list of systems is open. The following are included at this stage; UNL UWs (universal words), ONTOS concepts, WordNet synsets, NTT semantic categories.

3. Building the content

3.1. Recuperating existing resources

Building the content of the data base has several aspects. To initiate it, the project starts from open source computerized data, called "raw dictionaries", which may be monolingual (4,000 French DiCo entries from UdM, 10,000 Thai entries from Kasetsart Univ.), bilingual (70,000 Japanese-English entries and 10,000 Japanese-French entries in J.Breen's JDICT XML format, 8000 Japanese-Thai entries in SAIKAM XML format, 120,000 English-Japanese entries in KDD-KATE LISP format), or multilingual (50,000 French-English-Malay entries in FEM XML format).

3.2. Integrating the data into Papillon

In the second phase, the "raw dictionaries" are transformed into a "lexical soup" in M.Mangeot's (2001) intermediary DML format (an XML schema and namespace). The transformation into almost empty DiCo entries and the creation of axies for the translational equivalences is semi-automatic. A tool has been programmed at NII for that task.

3.3. Enriching the data with contributions

After that, it is hoped that many contributors will "fill in" the missing information. The basis for that third and continuous phase is a server for cooperative building of the data base, where each contributor has his/her own space, so that contributions can be validated and integrated into the DB by a group experts. Users can establish user groups with specific read and write access rights on their spaces.

4. Consultation of the resulting data

4.1. Online consultation

Consultation is meant to be free for anybody, and open source. Requests produce personalizable views of the data base, the most classical of which are fragments of bilingual dictionaries. However, it is also possible to produce multitarget entries, on the fly and offline. Users (consumers) are encouraged to become contributors. To contribute, one may propose a new word sense, a definition, an example of use, a translation, the translation of an example, a correction, etc., or an annotation on any accessible information: Every user can contribute with his own knowledge level.

4.2. Download of entire files

Users can also retrieve files, and can contribute to define new output formats. The files retrieved can contain structural, content-oriented tags. This open source orientation contrasts with the current usage of allowing users to retrieve files containing only presentation-oriented tags.

4.3. Coverage of the dictionary

An interesting point is that the project wants to cover both general terms and terminological terms.

Another one is that it contains a translation subproject, because definitions, examples, citations, etc. have to be translated into all languages of the collection. For this, the notion of complex lexie, already present to account for lexical collocations such as compound predicates (e.g. "to kick the bucket"), is extended to cover full sentences. Axies relating them are special because they can't in general relate them to external semantic systems such as WordNet. An exception is UNL: the UNL list for an axie may contain one UNL graph, produced automatically, manually, or semi-automatically. This graph may be automatically sent to available UNL "deconverters" to get draft translations.

5. Project organisation

In the current stage, the project has no legal implementation as a fundation, association, company, etc., although many participants have already established official MOUs and other types of agreements on which to base their cooperative work.

There is a steering committee of about 10-12 members, who represent Papillon where they are, and not the converse. There is a set of tasks, and for each task a working group and an advisory committee. One of the tasks is the management of the project. In between, there is a coordinating group containing the heads of the tasks and chaired by the head of the management task.

Sponsors may not donate money to the project, which has no bank account. Rather, they are encouraged to donate data, to assign personal part time to the project, and to fund participating organizations and persons as they see fit.

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 these 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.

Rerefences

Ampornaramveth V., Aizawa A. & Oyama K. (2000) *An Internet-based Collaborative Dictionary Development Project: SAIKAM*. Proc. of 7th Intl. Workshop on Academic Information Networks and Systems (WAINS'7), Bangkok, 7-8 December 2000, Kasetsart University.

Blanc É., Sérasset G. & Tchéou F. (1994) *Designing an Acception-Based Multilingual Lexical Data Base under HyperCard: PARAX*. Research Report, GETA, IMAG (UJF & CNRS), Aug. 1994, 10 p.

Connolly, Dan (1997) *XML Principles, Tools and Techniques* World Wide Web Journal, Volume 2, Issue 4, Fall 1997, O'REILLY & Associates, 250 p.

Ide, N. & Veronis, J. (1995) *Text Encoding Initiative, background and context*. Kluwer Academic Publishers, 242 p.

Mangeot-Lerebours M. (2000) *Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links*. Proc. of 7th Workshop on Advanced Information Network and System Pacific Association for Computational Linguistics 1997 Conference (WAINS'7), Bangkok, Thailande, 7-8 décembre 2000, Kasetsart University, 6 p.

Mangeot-Lerebours M. (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue.* Nouvelle thèse, Université Joseph Fourier (Grenoble I), 27 September 2001, 280 p.

Mel'tchuk I., Clas A. & Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*. AUPELF-UREF/Duculot, Louvain-la-Neuve, 256 p.

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

Sérasset G. (1994a) *Interlingual Lexical Organisation for Multilingual Lexical Databases*. Proc. of 15th International Conference on Computational Linguistics, COLING-94, 5-9 Aug. 1994, 6 p.

Sérasset G. (1994b) *SUBLIM, un système universel de bases lexicales multilingues; et NADIA, sa spécialisation aux bases lexicales interlingues par acceptions*. Nouvelle thèse, UJF (Grenoble 1), déc. 1994.

Sérasset G. (1997) *Le projet NADIA-DEC : vers un dictionnaire explicatif et combinatoire informatisé ?* Proc. of La mémoire des mots, 5ème journées scientifiques du réseau LTT, Tunis, 25-27 septembre 1997, AUPELF•UREF, 7 p.

Sérasset G. & Mangeot-Lerebours M. (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.

Tomokiyo M., Mangeot-Lerebours M. & Planas E. (2000) *Papillon : a Project of Lexical Database for English, French and Japanese, using Interlingual Links*. Proc. of Journées des Sciences et Techniques de l'Ambassade de France au Japon, Tokyo, Japon, 13-14 novembre 2000, Ambassade de France au Japon, 3 p.

-o-o-o-o-o-o-o-o-o-