

Collaborative construction of a good quality, broad coverage and copyright free Japanese-French dictionary

Mathieu Mangeot-Nagata^{1,2}

¹Department of Digital Media
Hosei University
3-7-2 Kajino-cho, Koganei,
Tokyo 184-8584 Japan

²GETALP, LIG-campus,
Université de Savoie Mont Blanc
BP 53 - 41 rue des mathématiques
F-38041 Grenoble Cedex 9 - France

mathieu.mangeot@imag.fr

1 Introduction

This research project is located in the field of natural language processing (NLP), at the intersection of computer science and linguistics, specifically multilingual lexicography and lexicology.

Concerning the Web, although French and Japanese are two well resourced languages (Berment, 2004), is not the case of the French-Japanese couple:

- Electronic French-Japanese bilingual dictionaries (denshi jishô) can not be copied to a computer or reused;
- There is a French-Japanese dictionary on the Web¹, but it only contains 40 000 entries, no examples and is not available for download.

There are collaborative Web dictionaries such as the Japanese-English JMdict project led by Jim Breen (2004) that contains over 173,000 items. These resources are freely downloadable. It is therefore possible to carry out such projects.

During a first stay in Japan from November 2001 to March 2004, we had already noticed the lack of French-Japanese bilingual resources on the Web. Which gave rise to the Papillon project about the construction of a multilingual lexical database with a pivot structure (Sérasset et al., 2001). Since then, progress has been made in several areas (technical, theoretical, social) (Mangeot, 2006) but the actual production of data has made very little progress. On the other hand, there is a new trend in reusing existing lexical resources (word sense disambiguation, using open source resources (Wiktionary, dbpedia) merging with ontologies, etc.). Although they allow to consolidate and expand the coverage of existing resources, these experiences still use data created by hand by professional lexicographers. There are printed French-Japanese dictionaries of good quality and sufficiently old to be royalty free. It should be possible to reuse these resources as part of our project to build a good quality dictionary and broad coverage available on the Web.

Based on this observation, we defined the following project to build a rich multilingual lexical system with priority over French-Japanese languages. The construction will be done first by reusing existing resources (printed Japanese-French dictionaries, Japanese-other language dictionaries,

¹<http://www.dictionnaire-japonais.com>

Wikipedia) and automatic operations (scanning and corrections, calculating translation links) and then by volunteer contributors working as a community on the Web. They will have to contribute to dictionary articles according to their level of expertise and knowledge in the field of lexicography or bilingual translation.

The resulting resources will be royalty-free and intended for use by both humans via conventional bilingual dictionaries and by machines for automatic language processing tools (analysis, machine translation, etc.).

First, we will conduct an inventory of French-Japanese bilingual dictionaries, then describe the resource we want to build. The following sections concern the conversion of three resources: the Cesselin printed dictionary, the language links between Wikipedia pages and the JMdict electronic dictionary. Finally, we conclude with the release of the resource on a Web site built around the Jibiki platform allowing to view and edit articles online.

2 State of the art of Japanese bilingual dictionaries

Although French and Japanese are regarded as well-resourced languages concerning tools and linguistic resources, the French-Japanese couple is considered an under-resourced language pair (Berment, 2004). Indeed, there are few bilingual electronic lexical quality resources and royalty copyright free. French-Japanese bilingual aligned corpora and machine translation systems are logically equally rare.

For historical as well as practical reasons, Japanese people quickly put the emphasis on English. The English-Japanese couple is one of the best equipped at present with very substantial resources like the EDR dictionary (1993) and machine translation systems among the best performers.

2.1 French-Japanese Printed Dictionaries

In this section, we present the most significant dictionaries, either for historical reasons or by the innovations they bring. It should be noted that until recently that there are two distinct lexicographical traditions as the writing team is of French or Japanese mother tongue and also that dictionaries are unidirectional (a language to another but not vice versa). It was the case until 2009 with the release of Assimil bidirectional dictionary (Hisamatsu et al., 2009). Moreover, until the 1950s, French mother tongue authors are all Catholic missionaries. The primary objective was to translate the Bible into Japanese.

2.1.1 Japanese → French dictionaries

1603: *Vocabvlario da Lingoa de Iapam (nippo jisho)*. This Japanese → Portuguese dictionary contains 32,293 articles written by the Portuguese Jesuit missionaries is considered Japan's first bilingual dictionary.

1862: Translation of the *Nippo jisho* into French by Léon Pagès (1814-1886). He takes care of adding a katakana transcription of Japanese words (Griole, 2008).

1904: Lemaréchal dictionary written by Jean Lemaréchal (1842-1912) containing around 60,000 articles on 1,008 pages. He abandons the Latin transcription adapted to French for the more widespread Hepburn romaji (Griole, 2008).

1939: Cesselin dictionary (Cesselin, 1939) written by Gustave Cesselin (1873-1944), containing 82,500 articles on 2,340 p. It is considered "the best from the perspective of those who study the Japanese language in depth, as it provides many examples presented in alphabetical form." (Griole, 2008).

2009: Assimil Japanese dictionary. This dictionary (Hisamatsu et al., 2009) is to our knowledge the first two-way bilingual dictionary (French → Japanese and Japanese → French). It contains 24,000 articles on 1,280 pages. It also contains 135,000 words, phrases and translations, 35,000 usage examples. All words and phrases are transcribed into romaji. This makes it a very useful tool for French speakers learning Japanese.

2.1.2 French → Japanese Dictionaries

1864: futsugo meiyō dictionary (elucidation of the French language) by Hidetoshi Murakami (1811-1890). This scholar is considered the first Japanese to have learned French and this through a French-Dutch dictionary (Koichi, 2010).

1866: French-English-Japanese dictionary by Father Eugene Mermet de Cachon (1828-1871) containing about 5,300 articles on 433 pages. The Japanese was reviewed by Léon Pagès.

1887: "dictionnaire universel français-japonais" (French-Japanese universal dictionary) by Nakae Katsusuke and Nomura Yasuaki. This dictionary is a Japanese translation of the French monolingual dictionary Petit Littré. It also marks the first integration of multiple word senses.

1905: French-Japanese dictionary by Emile Raguet (1854-1929) and Tota Ono, 1,048 pages.

1953: second edition of the previous French-Japanese dictionary (Raguet & Martin, 1953) called "Raguet-Martin", revised and expanded by Jean Marie Martin (1886-1975). This dictionary contains about 50,000 items on 1,445 pages. It "remains the only major Japanese dictionary to submit translations in alphabetical form, and therefore intelligible without knowing ideograms." (Griole, 2008).

1983: "dictionnaire franco-japonais de notre époque" (Franco-Japanese dictionary of our time) published by Mikasa-shobo. Contains about 42,000 articles on 1,763 pages. The romaji is indicated for the Japanese as well as the pronunciation of French words.

1988: Shôgakukan-Robert dictionary. This dictionary, result of the Japanese translation of the monolingual French Robert dictionary is a considerable work. It remains to this day the largest French-Japanese dictionary printed since it contains more than 100,000 articles. Unfortunately, there is no romaji (Latin transcription) or furigana (kanji pronunciation). It is intended primarily for Japanese-speaking users.

2.1.3 Electronic Dictionaries (denshi-jishō)

The Crown French → Japanese dictionary (Sanseido, 1978) contains 47,000 articles. There is neither romaji nor furigana (see Figure 1).

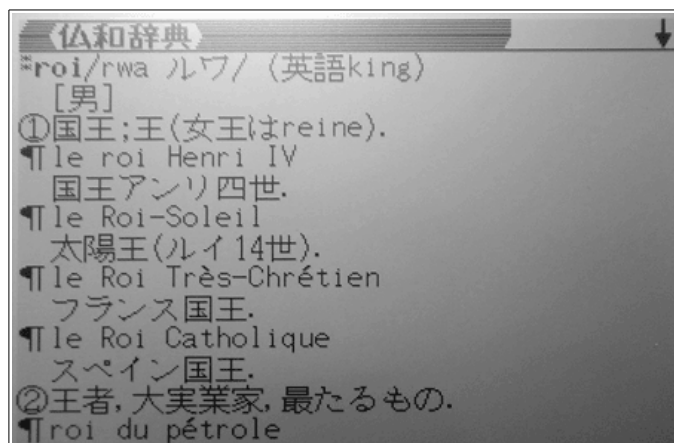


Figure 1 : Screenshot of the Crown dictionary in electronic version

The Japanese → French Concise dictionary (Sanseido) contains 38,000 articles. There is neither romaji nor furigana (see Figure 2).

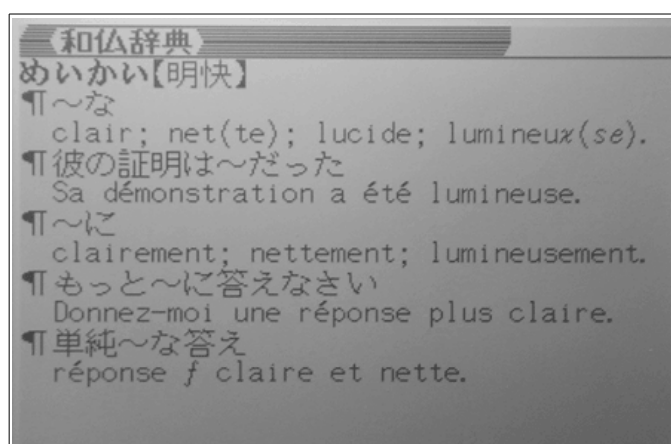


Figure 2 : Screenshot of the Concise dictionary in electronic version

Francophones with a level sufficient to read Japanese certainly need a wider coverage dictionary. These dictionaries are thus designed for Japanese-speaking users.

2.1.4 Conclusion

Electronic dictionaries are not reusable outside the medium in which they are sold. In addition, they are designed for Japanese-speaking users and their coverage is not very wide.

The only existing French-Japanese dictionaries with good quality and broad coverage are publishing dictionaries that exist only in paper format for which there is no online lookup interface. However, some are old enough to be copyright free.

Figure 3 shows the changing number of articles in the French-Japanese dictionaries based on years. Note a peak in the 1950s. It must be possible to reuse some of these resources in our project to build a dictionary of good quality and broad coverage available on the web, provided that it is updated with modern vocabulary.

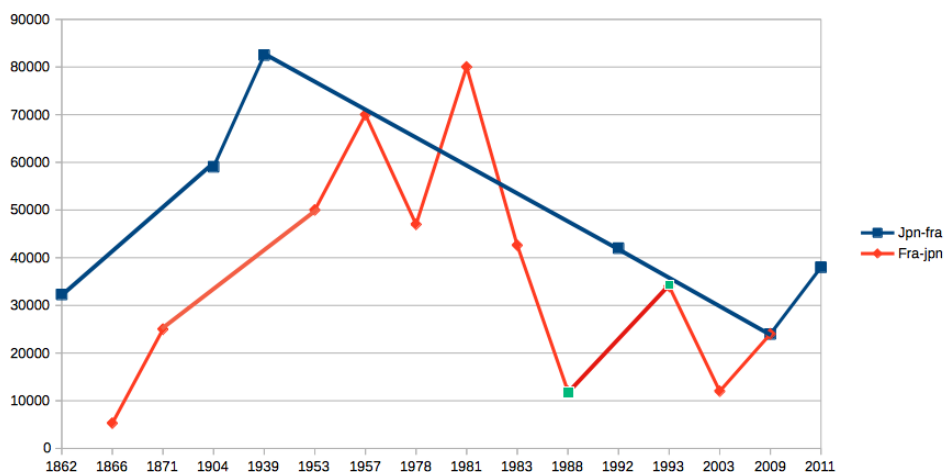


Figure 3 : History of the number of articles in French-Japanese dictionaries

2.2 Wiktionary Projects

The French Wiktionary currently has 2.2 Millions of articles including 1.2 Millions of French articles and with a little less than 7,000 Japanese translations of which about 1/2 is a proper noun (often a simple transcript in Japanese syllabary of the headword). There are also a few translations coming from the dictionnaire-japonais.com website (see 2.4.2). Translations are indicated at the entry level and not at the sense level. There is no translation of context description (gloss, examples, etc.), or information on the Japanese translation (part-of-speech, etc.).

The Japanese Wiktionary has 83,000 articles with 26,000 Japanese articles and 2,800 French articles translated into Japanese. There are also inflected forms or oral conjugated forms, e.g. 32 articles for the verb "aimer" (to like/love). The coverage is very insufficient.

Wiktionary projects are interesting and fashionable but they have several limitations:

- The structure of the articles is free. It is not possible to use the same precise microstructure for all articles.
- Although it is possible to describe, in a language A Wiktionary, a word sense of a language B into language A, the initial interface is not designed for writing bilingual dictionaries. For example, the description of the reverse language link $A \rightarrow B$ must be done by hand in language B Wiktionary.
- It is not possible to automatically add existing data from other sources in order to build a draft to be refined later.
- The contributions are anonymous. It is not possible to use a quality level for data or a review / validation system.

Although the success of the Wikipedia project can logically lead us to think that it will be the same for the collaborative construction of quality bilingual or multilingual dictionaries, this is not the case. We quote in this regard, Larry Sender, co-founder of Wikipedia:

“To try to develop a dictionary by collaboration among random Internet users, particularly in a completely uncontrolled wiki format, now strikes me as a nonstarter.”

Indeed, every Wikipedia article can be written by a specialist in the field in question, but for a general language dictionary, it is not possible to find a specialist for only a few articles. Only linguists specialists of the language and professional translators (after being trained in lexicography) can write an entire article.

2.3 Online Japanese–other language Resources

2.3.1 Japanese → English Dictionary: JMdict

The JMdict² (Japanese-Multilingual Dictionary) (Breen, 2004) is a project led by Jim Breen. It contains 173,000 Japanese entries translated into English with additional translations in other languages: German (from WaDokuJiten), 31,000 French equivalents (from the dico FJ), Russian, Dutch, etc.

Search Key: 食べる Current Dictionary: Jpn-Eng General (EDICT)

Options:[G]oogle search, [GI] Google images, [S]anseido dictionary, [A]LC dictionary (Eijiro), [Ex]ample sentences, [V]erb conjugations, [F]eedback, [L]esson from JapanesePod101.com, [JW] Japanese WordNet, [W] Japanese Wikipedia,[Edit] Edit this entry,[Promote] Move to JMdict/EDICT.

● 食べる (P); 喰べる (iK) 【たべる】 (v1,vt) (1) to eat; (2) to live on (e.g. a salary); to live off; to subsist on; (P) [\[Edit\]](#)
[\[V\]\[Ex\]\[L\]\[G\]\[GI\]\[S\]\[A\]\[W\]](#) [\[JW\]](#) [\[L\]\[G\]\[GI\]\[S\]\[A\]](#)
 エジプトでは何を食べて生活していますか。 What do they live on in Egypt?[\[Amend\]](#)

○ ガンガン食べる; がんがんと食べる 【ガンガンたべる(ガンガン食べる); がんがんとたべる(がんがんと食べる)】
 (exp,v1) (sl) to pig out; to chow down [\[Edit\]](#) [\[V\]\[GI\]\[S\]\[A\]](#) [\[GI\]\[S\]\[A\]](#)

○ 生で食べる 【なまでたべる】 (exp,v1) to eat raw (fresh) [\[Edit\]](#) [\[V\]\[L\]\[GI\]\[S\]\[A\]](#)

○ 一口食べる 【ひとくちたべる】 (v1) to eat a mouthful [\[Edit\]](#) [\[V\]\[L\]\[GI\]\[S\]\[A\]](#)

○ ぼりぼり食べる 【ぼりぼりたべる】 (exp,v1) to eat with a munching or crunching sound [\[Edit\]](#) [\[V\]\[L\]\[GI\]\[S\]\[A\]](#)

Figure 4 : “食べる” (taberu) article of the JMdict dictionary

Advantages: ressource à large couverture, libre de droits et disponible gratuitement au téléchargement. Elle est aussi régulièrement révisée et complétée.

Inconvénients: dictionnaire unidirectionnel japonais → autre langue. Il n'existe pas de dictionnaire inverse anglais → japonais. La microstructure est limitée : les contextes de traduction ne sont pas décrits. Il manque également une définition et des exemples.

Advantages: broad coverage resource, royalty-free and available for download. It is also regularly revised and updated.

Disadvantages: unidirectional dictionary Japanese → other language. There is no English → Japanese reverse dictionary. The microstructure is limited: the translation contexts are not described. It also lacks a definition and examples.

2.3.2 Japanese → German Dictionary: WaDokuJiten

The WaDokuJiten³ from Ulrich Apel (Apel, 2002) consists of more than 280,000 entries. Its wide

² <http://www.csse.monash.edu.au/~jwb/jmdict.html>

³ <http://www.wadoku.de>

coverage as well as its microstructure are more developed than the JMdict.




Nr.	Japanisch	Lesung	Deutsch	Worttyp
1	食べる	たべる	[1] essen; speisen; zu sich nehmen; fressen; probieren. [2] leben von.	下一他 
2	食べるのを遠慮する	たべるのをえんりよする	nicht essen.	サ変自 
3	食べるとしゃきしゃきする	たべるとしゃきしゃきする	beim essen knusprig sein.	サ変自 

Figure 5 : “食べる” (taberu) article of the WaDokuJiTEn dictionary

Advantages: more complete than the JMdict in terms of coverage and information, free of charge and available for download.

Disadvantages: like the JMdict, the dictionary is unidirectional. It also does not contain usage examples to illustrate the translation contexts.

This dictionary is to date the most comprehensive Japanese-other language resource available for free. It constitutes a goal to reach for our resource in terms of coverage.

2.4 French-Japanese Resources Available Online

2.4.1 Dico FJ Project

The dico FJ dictionary project, pioneer in the field of Japanese-French resources on the Web, was launched in early 2000 by Jean-Marc Desperrier (Desperrier, 2002). It contains just over 10,000 entries mainly coming from translation of the Japanese-English dictionary JMdict by Jim Breen. There has not been any change since 2003.

Advantages: royalty-free and available for download.

Disadvantages: more disadvantages than the JMdict. Translation errors arise because some contributors with poor level of Japanese translated the English translations directly to French instead of translating the Japanese headword, increasing the number of misinterpretations.

2.4.2 Dictionnaire-japonais.com

The dictionnaire-japonais.com¹ project currently contains just over 40,000 words. There is a clear progress compared to other Japanese-French online dictionary projects. Each user can contribute directly by adding entries. The community of contributors looks quite active as evidenced by the activity on the project forum. The information available for each entry are relatively limited to a "grammatical type," a "category" (field), a language level, and sometimes an "origin of the word" (etymology).

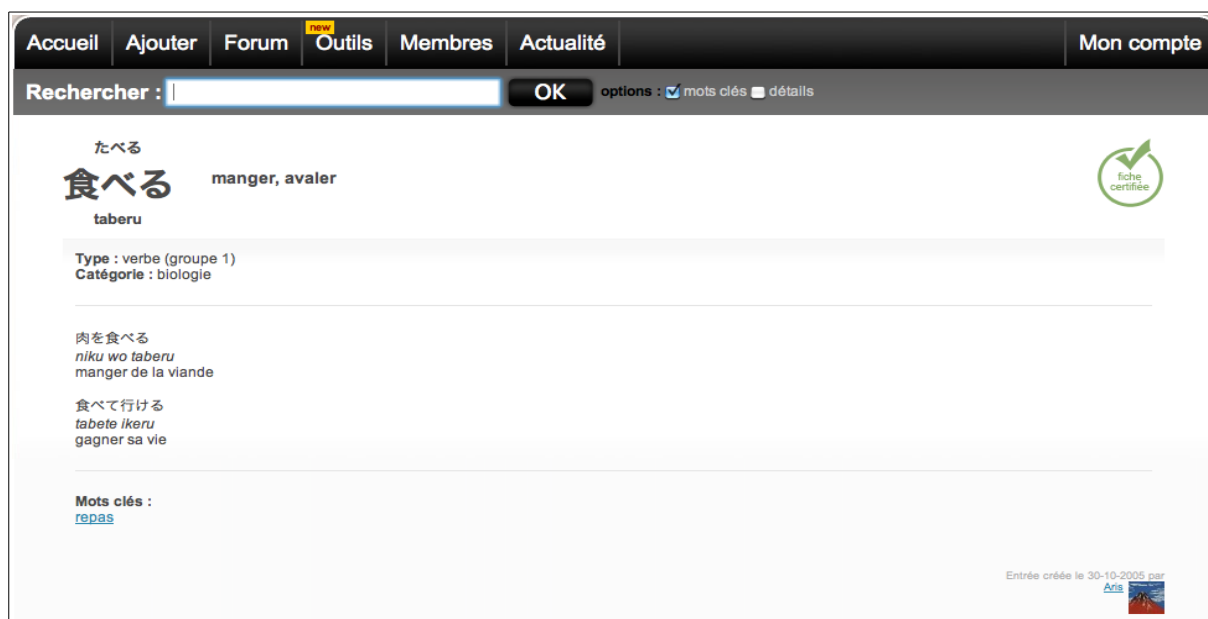


Figure 6: 食べる (taberu) Article of dictionnaire-japonais.com

Advantages: available online, the coverage is wider than the dico FJ, there is an active community of volunteer contributors.

Disadvantages: added to the disadvantages of the dico FJ dictionary, the data is not freely available for download.

2.5 Conclusion

The French-Japanese dictionaries available online do not have a broad coverage and they are all oriented from Japanese to French.

Most French-Japanese dictionaries also lack of information to be used by both French speaking and Japanese-speaking users. For example, to our knowledge, there is no dictionary with both kanji (ideograms), kana (syllabic) and romaji (transcription in the Roman alphabet).

Dictionaries for Japanese-speaking users do not indicate the romaji and most of the time, the examples are written with kanji without furigana. Beginners learners of Japanese can not read them. Dictionaries for French speaking users do not indicate the pronunciation of French words or the gender (masculine/feminine) which are essential to Japanese-speaking readers. They also lack some important information such as counters (we do not count the same way objects: one car = ichi dai, one dog = ippiki, etc.) or language levels.

In conclusion, for personal use, it is possible to find printed dictionaries (or their electronic version) of pretty good quality if you know how to read kanji; but when one look for a free dictionary or a resource reusable in other tools, there is no choice but to use an English-Japanese dictionary, which, as we know, can only increase the misunderstandings and translation errors.

However, JMdict and WadokuJiten projects show that it is possible to carry out collaborative dictionary construction projects online.

At this point, we can redefine our project this way: to retrieve and convert quality French-Japanese printed dictionaries copyright free thanks to an optical recognition process and to make them available online for users that will be able correct the remaining errors and update the data.

3 Description of the resource to build

3.1 History of the project

In 2001, already faced with the same problem of lack of French-Japanese bilingual lexical resources, we launched the Papillon Project (Mangeot et al., 2004) which allowed us to move forward on theoretical aspects with the definition of a pivot macrostructure.

In 2003, the launch of the GDEF project (Mangeot & Chalvin 2006) of an Estonian-French bilingual dictionary was an opportunity to move forward on the software part with Jibiki, a generic online platform for lexical resources management.

In 2010 the MotÀMot project (Mangeot, 2014) of a French-Khmer dictionary was the opportunity to work on defining quality levels.

In 2012, the DiLAF project (Enguehard & Mangeot, 2014) led us to define a precise methodology for data recovery from standard text processors (Word).

3.2 Microstructure of the Articles

3.2.1 General Microstructure

Generally speaking, our articles will be based on a combination of a lexeme and a part-of-speech. The articles will therefore not have any grammatical block. The articles structuring will follow that defined by the Lexical Markup Framework (LMF) standard (Francopoulo et al., 2009): each article contains a form block which includes information related to the form: headword, pronunciation, part-of-speech and a semantic block with a list of sense blocks. Each sense block describes a word meaning. It also contains the translation in another language as well as a list of examples. Each example is translated into the other language.

The data come primarily from the conversion of existing resources. At first, the structure of articles will therefore follow that of the original resource. Then, in the medium term, our goal is to strive for a richer microstructure based on the Explanatory and Combinatorial Lexicography (Mel'čuk et al., 1995), part of the Meaning-Text Theory (MTT). For each word sense (formally a lexie), add a semantic formula that can be seen as a formal definition. In the case of a predicative lexie, the formula describes the predicate and its arguments and also the schema that describes the syntactic realization of arguments. Then, add a list of lexico-semantic functions. There are 56 basic functions applicable to any language that can be combined together. Finally, add a list of examples and possibly idioms.

3.2.2 Japanese Articles

The articles of most Japanese dictionaries are based on the lexeme. There are no homographs articles. We will follow this division.

Regarding the headword, each kanji has several possible pronunciations. Therefore, it is necessary to indicate the pronunciation. This is usually done by using the hiragana syllabary. If we simply add kanji and hiragana, beginners in Japanese can not easily read the articles. We must also use a Latin transcription of the Japanese called romaji. There are several methods for official romaji: the oldest and most used is the Hepburn method, introduced by the American missionary James Hepburn in

1887. The Kunrei is a method introduced by the Japanese Ministry of Education and described as the ISO standard number 3602:1989. It is based on Japanese phonology. Its main advantage is that it illustrates better the grammar, while Hepburn biggest problem is that it changes radical verbs, which does not reflect the underlying morphology of Japanese. However, non-native speakers prefer Hepburn method because it gives a better indication of English pronunciation. In our dictionary, the goal is to help beginner learners of Japanese who are non-native speakers, not to indicate native speakers how to write romaji. We therefore choose the Hepburn method.

It is unfortunately not possible to automatically generate the romaji only from the hiragana pronunciation. Indeed, the "う" (u) letter is transcribed in two different ways depending on whether it extends a vowel "o" or not. The combination of the two vowels "o" and "u" can be written in two different ways in romaji depending if the "u" is the beginning of a morpheme (kanji) or not. For example, the word “東京” (Tokyo) is written in hiragana "とうきょう" and romaji "Tokyo" (not "toukyou"), while the word "子牛" (calf) is written in hiragana "こうし" and romaji "koushi" (not "kōshi"). We therefore choose to represent each headword with three parts:

- a first part "Japanese-headword" indicates the Japanese word as it appears in the texts. This can be a kanji word only (for nouns), a combination of kanji and hiragana (for verbs and adjectives in particular), a word in hiragana only (adverbs), a word in katakana (word of foreign origin) or any combination of the three writing systems (kanji, hiragana and katakana);
- a second part "hiragana-headword" always indicates the pronunciation of the word in hiragana (even if the Japanese-headword is already a word in hiragana);
- a third part "romaji-headword" indicates the transcription of the Japanese-headword in modern Hepburn romaji. This part can itself contain two versions: one for display that can contain spaces, dashes or dots and one for lookup that contains only letters.

In the remaining of the article, for each Japanese text segment, we add the pronunciation in hiragana above the text (called furigana) and a transcription in Hepburn romaji.

3.2.3 French Articles

Traditional lexicography distinguishes two vocables as homographs if they have no clear semantic link between them (Polguère, 2008). But in practice, it frequently happens that dictionaries with the same source language follow a different division between homograph vocables. We believe that this distinction is arbitrary. Furthermore, to our knowledge, there is not of objective criteria for assessing automatically a semantic link between two words with automatic processing tools. We therefore choose not to distinguish homograph vocables if they have the same grammatical category. The combination of a lexeme and a part-of-speech therefore constitutes a single article.

For the non-French-speaking users of the dictionary, pronunciation of the headword will be added. The gender (masculine / feminine) of each noun representing a French translation in a Japanese article will be added.

3.3 Quality Levels

Each article is assigned a level of quality. The levels range from 1 star for a draft (converted data

whose quality is unknown) to 5 stars, for an article certified by an expert (e.g., translation link validated by a professional translator).

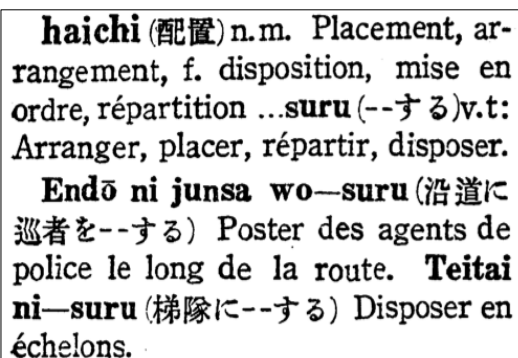
Likewise, contributors will be assigned a skill level (1-5 stars as well). 1 star being the level of an unknown novice in the community and 5 stars being the level of a recognized expert.

Then, for example, when a level 3 contributor revises a level 2 article, the article automatically moves up to level 3. Similarly, if the work of a contributor is systematically validated without corrections by other senior contributors, s/he can automatically switch to the next level after a certain threshold (e.g. 10 contributions).

To go further, we plan to analyze the work of the contributors. If a person contributes massively e.g. on a specific domain, the system will automatically send regular contribution proposals in this domain.

4 Conversion of the Cesselin Dictionary

4.1 Presentation of the Dictionary



haichi (配置) n.m. Placement, arrangement, f. disposition, mise en ordre, répartition ...**suru** (--する) v.t: Arranger, placer, répartir, disposer.
Endō ni junsā wo—suru (沿道に巡者を--する) Poster des agents de police le long de la route. **Teitai ni—suru** (梯隊に--する) Disposer en échelons.

Figure 7: Scan of the « 配置 » (haichi) article of Cesselin dictionary

The "Cesselin" (Cesselin, 1944) is a French → Japanese dictionary developed by Gustave Cesselin, apostolic missionary who died in 1944 having spent his entire career in Japan. The dictionary contains 2,365 pages and over 82,600 articles. The headword is noted in romaji and Japanese (kanji or kana). It is followed by a part-of-speech in French and a list of French translations. Next comes a list of phrases containing the headword in romaji and Japanese (kana and kanji). Each phrase is translated into French. The article concludes with a list of examples, each noted in romaji, Japanese and translated into French (see Figure 7).

4.2 Copyright Negotiation

To ensure that a book is copyright free, one must verify that all authors are dead for a fixed term. In the case of a dictionary written by a significant number of authors, such verification may be difficult. Furthermore, the duration varies from one country to another. In Japan, it is currently 50 years. In France, it is 70 years, duration shared with many countries.

Gustave Cesselin is the only official author of his dictionary. He died in 1944. Thus, in Japan like in France, the "Cesselin" dictionary is copyright free.

If the targeted dictionary is not copyright free, another solution is to negotiate directly with the copyrights holders. In the case of the missionaries who are the most numerous authors until the 1950s, holders are the congregations.

In the case of "Raguet Martin" French-Japanese dictionary (Raguet & Martin, 1953), Jean-Marie Martin died in 1975. We must therefore wait 10 years in Japan and 30 years in France for the dictionary to be finally copyright-free. So we contacted the "Missions Étrangères de Paris", congregation holding the copyrights of this dictionary. An agreement was concluded for the use of the "Raguet Martin" data on our website.

4.3 Dictionary Scanning

Several scanning techniques were tested:

- manual scanning with a flat scanner. The resulting image gets a black band in the middle of the two pages that hides some characters at the beginning of each line, so potentially headwords words. The result is not usable. Moreover, the operation is tedious because it requires lifting the dictionary and turning each page by hand.
- scanner with camera on top. The procedure is quick (3 hours) because it is not necessary to move the book. Unfortunately, a small bend remains in the middle of the two pages.
- scanner with one camera above and another aside. The side camera is used to calculate the thickness of the book and then automatically straighten the curvature in the middle of the two pages. We could not test this type of machine.
- book cutting and automatic scan. This technique gives the best scan quality since there is no curvature or black area. Moreover, it is very fast because automatic. However, it requires to sacrifice a copy because the binding is cut in order to scan the book page by page.

4.4 Optical Character Recognition

Once the dictionary scanning is achieved in the best possible conditions, we must find an optical reader software that is able to:

- recognize multiple languages simultaneously;
- allow the training of the optical recognition;
- recognize the kanji.

We tested a dozen of software. Abbyy was the only one to comply with most of our requirements. Its major drawback is that it does not allow the training of the optical recognition for ideograms. We then performed two passes. The first pass was performed by choosing only French as recognition language and training the optical recognition on the first page. The French text parts have been correctly recognized with very few errors. Text portions in Japanese (kanji or kana) were not recognized at all. The second treatment was performed by choosing both French and Japanese as recognition languages, which forbid us to train the recognition (because Japanese uses ideograms). The French text parts have been recognized with a higher error rate than during the first pass and the Japanese text parts have been recognized with a standard error rate. It takes about 12 hours for a

pass throughout the whole dictionary.

The results of the process are then exported into OpenXML (.docx) or OpenDocumentFormat (odt). These documents are actually zip archives. They are then decompressed and the XML files containing the text of each page of the dictionary are extracted. The result files of the two passes are then merged using an algorithm for calculating string distance such as Levenshtein.

4.5 Headword Detection

The most important part of the conversion of a dictionary is located in the headwords detection. They delimit the entries and are also access keys. This part consists of three phases:

1. Headwords extraction from the header of each page. This operation is very important because they will be used to delimit the alphabetical order of the headwords in the page. In the .docx or .odt archives, the pages headers are separated from the main text. A tool was written to automatically extract these headers. Then, the headers were verified (is it romaji and are they sorted in alphabetical order) and corrected if necessary. In the example of Figure 8 for the page 323 of the Cesselin, the headers headwords are 'haichai' and 'haigeki'.
2. Comparison of each beginning of line in strict alphabetical order done first with the headers: header 1 < Word < header 2 and then between the detected headwords: headword 1 < headword 2 < headword 3. In Figure 8, the following headwords are extracted: haichai, haichi, haichi, haichüritsu, haidan
3. Approximate comparison of each beginning of line. To take into account OCR errors, a second comparison is performed. This time, an accepted percentage of errors is introduced by using Levenshtein string distance algorithm. In Figure 8, headwords « iichi » and « haichutsu » are extracted and automatically corrected to become « haichi » and « haichutsu ».

A problem of over-detection remains. Indeed, some phrases or examples included in an article start at the beginning of a line and reuse the article headword. Example, page 1,038 of the Cesselin, "kuwadate iru" was detected as headword but it is an example of the article "kuwadateru" and "Kuwashiku monoshiberu" was detected as a headword but it is a phrase of the article "kuwashii".

```

<p>de poitrine ...no(—cd) ...sbitsu no<j>--</j>a. Phitisque ...wo wazurat-</p>
<p>te iru <j> を患つてゐる </j>Souffrir d'une</p>
<p>maladie de poitrine.</p>
<p>haichai <j> はいちやい </j>V' b ^ ^T enf : Au re-</p>
<p>voir! Adieu!</p>[...]
<p>haichi <> 配置 </j>n.in. Placement, ar-</p>
<p>rangement, f. disposition, mise en</p>
<p>ordre répartition.</p>
<p>iichi <j> 廢她 </j> n.m. Déclin, relâche-</p>
<p>t, f. decadence ...suru <j> する </j></p>
<p>Décliner, se détériorer.</p>
<p>haichi <> 廢置 </j> n. l.f. Abolition et</p>[...]
<p>haichūritsu <j> 排中律 </j> n.f. Loi ex-</p>
<p>cluant l'intervention d'un tiers.</p>
<p>haicbutsu <> 廢黜 </j> n. f. Déposition</p>
<p>et expulsion</p>[...]
<p>haidan <j> 俳談 </j> n.m. Conte comique</p>
<p>et à double sens.</p>

```

Premier passage

Deuxième passage

Figure 8: Headword detection for the page 323 of the Cesselin dictionary

4.6 Post-OCR Corrections

After the optical recognition, it is possible to automatically correct text characters depending on the language in the text.

4.6.1 French

Corrections to the French mainly focus on the use of diacritics. Examples: Â + ' ⇒ À; Etre ⇒ Être; ç[[^]àaou] ⇒ c; etc.

4.6.2 Romaji

The romaji uses only macron as diacritical mark on the vowels: ā,ī,ū,ē,ō. Other diacritic letters are automatically converted: à ⇒ a; äâ ⇒ ā, etc.

Some character strings do not exist in romaji. They are also converted: lt + [aiueo] ⇒ h + [aiueo]; rn + [aiueo] ⇒ m + [aiueo], etc.

4.6.3 Japanese

Errors on the Japanese text segments appear mainly at the beginning or end of segment. When there is a change of language between French and Japanese, the software detects this change too late. Therefore, the first and last characters of a Japanese segment are sometimes replaced by ASCII characters. Some patterns are often repeated. There are below three examples of replacement. The <j> opening tag indicates the start and the </j> end tag the end of the Japanese segment:

- 9 ⇒ り; | c ⇒ に; = ⇒ ニ
- v') ⇒ い Ex : <j>と忙はし</j>v') ⇒ <j>と忙はしい</j>
- 't) ⇒ す Ex : <j>心を越</j>'t) ⇒ <j>心を越す</j>

4.6.4 Priority Lists for Corrections

In order to prioritize corrective work for the contributors when the resource will be online, priority lists were calculated from word frequency lists in a Japanese monolingual corpus. The frequency

list of words used (JapFreqList_5109_Novels) comes from a corpus of 5,109 Japanese novels of modern Japanese literature. It was built for the cbJisho⁴ project and can be downloaded by following this link⁵. It contains 188,218 entries. The highest frequency is the Japanese comma ",", with 26,244,137 occurrences.

To prioritize the remaining work on the headwords with undetected kanji, a new frequency list based on the hiragana was generated from the JapFreqList list. The JMdict dictionary is first queried via the REST⁶ Application Programming Interface (API) of the Jibiki platform¹⁷. Then, the frequencies obtained for each hiragana are summed and the resulting list is sorted. The last step is to compare this list with the one of undetected kanji headwords in the Cesselin.

4.7 Modernization of the Japanese

Since when the dictionary was published, the Japanese language has undergone many changes including in its writing. The goal of the project is not to reproduce as faithfully as possible the Cesselin but to build a modern dictionary reflecting the current use of the language. This is why we decided to modernize the Japanese automatically where possible.

4.7.1 Romaji

The romaji used at the time of writing the Cesselin, based on Hepburn transcription is called romajikwa. It uses the letters "kwa" and "gwa" to transcribe certain syllables that are now simplified into "ka" and "ga". Example: kwaikwan ⇒ kaikan. The "m" was also used to transcribe the letter hiragana letter “ん” before "m, b, p" consonants because it is actually pronounced "m". Now, the "n" is used everywhere. Example: jimbōchō ⇒ jinbōchō; gumma ⇒ gunma.

The transcription of the "n" consonant is ambiguous in Japanese as it can be the final syllable "ん" or the beginning of a syllable (na, ni, nu, ne, no). It is customary to note in romaji when the n is a final syllable. In the Cesselin, a dash is added after "n". But the dash is also sometimes used to separate syllables. On the other hand, current romaji uses other characters. The modern Hepburn uses an apostrophe "'". Francophones use the dot ".". So we replaced the dash marking a "ん" with a dot. Example: ran-i ⇒ ran.i

4.7.2 Kanji Simplification

At the time of the writing of the dictionary, there was no encoding table for the kanjis. The first Japanese encoding, JIS (Japanese Industrial Standard) 208 appeared in 1978. It contains about 7,000 characters. This encoding has been widely used during the computerization of the Japanese. So, the kanji not included in this encoding gradually disappeared. To remedy this situation, other coding appeared such as JIS 212 in 1990 which specifies 6,067 characters, and JIS 213, which specifies 11,233 characters. But the damage was already done and most of the current Japanese words use only the JIS 208 kanjis. We have therefore replaced all the kanji that were not included in the JIS 208 with their JIS 208 variant. In the same way, we replaced the JIS 208 variants by those of the lowest grade when available.

- JIS 213 → JIS 208 replacements: 状⇒状,歩⇒歩,黄⇒黄,縁⇒縁,晩⇒晩,黒⇒黒

4 <http://subs2srs.sourceforge.net/cbJisho/help.html>

5 <http://forum.koohii.com/viewtopic.php?pid=132030#p132030>

6 https://en.wikipedia.org/wiki/Representational_State_Transfer

- JIS 212 → JIS 208 replacements: 啞⇒啞,搔⇒搔,頰⇒頰,上⇒上,伙⇒火
- JIS 208 variants replacements: 阪 8⇒坂 3,弍⇒一 1,埜 10⇒野 2,京⇒京 2,區⇒区 3

Some ideograms were even not included in any JIS. They are part of the Han Chinese characters. Some are actually used in the original Cesselin dictionary but the others come from character recognition errors. For these two series, we had to find a JIS 208 equivalent by hands.

- Examples of Han characters included in the Cesselin: 絶⇒絶, 説⇒説, 青⇒青.
- Examples of Han characters coming from errors: 内⇒内, 戸⇒戸, 出⇒出.

4.7.3 Japanese

Japanese language also evolved since the 1950s, particularly the verb endings. The evolution was already indicated in the romaji (romajikwai → romaji) but the hiragana used for the verb endings in Japanese kept track of old pronunciations. Examples: the "ふ" (fu) ending is pronounced "u"; the “へる” (heru) ending is pronounced "eru". Therefore this situation created a mismatch between the romaji and hiragana. So we changed the hiragana to match romaji and thus the modern pronunciation. E.g.: kokitsukau 扱使ふ ⇒ 扱使う; kikikaeru 切替へる ⇒ 切替える.

Some hiragana letters were also replaced systematically: ゐ⇒い (i); ゑ⇒え (e). Example: 光ってゐる ⇒ 光っている.

The sokuon, letter indicating a geminate consonant noted with a macron in Hepburn romaji, is usually noted with a small tsu "っ". In the Cesselin dictionary, the sokuon is noted with a standard size tsu "つ". In the following verifications, all variants with and without small tsu are generated to avoid over-detection problems.

4.8 Information Tagging

Once all corrections made, it is time to move to the markup of each part of information. Initially, only segments containing Japanese (kana or kanji) are tagged. The result of the previous steps also tagged out the headword in romaji and Japanese. All other segments must be tagged.

We first listed all abbreviations used in the dictionary which allowed us to tag the etymology, part-of-speech, the domain and language levels. We also included in the list frequent errors from the optical reading. Example: "n.in." instead of "n.m." on Figure 9.

Then, the Japanese segments are preceded with segments in romaji. The latter begin either with "... " or a capital letter. Japanese segments are also followed by French segments ending with a dot. These rules allowed us to tag the examples.


```

<article><vr>haichi</vr><vj> 配置 </vj> (n.in. Placement, arrangement f. disposition,
mise en ordre répartition ...suru </r><j> する </j><v.t.> Arranger, placer, répartir, disposer.
Endô ni junsâ wo—suru </r><j> 沿道に巡者を -- する </j><fra> Poster des agents de police le
long de la route. Teitai ni—suru </r><j> 梯隊に -- する </j><fra> Disposer en échelons. </article>

catégories 品詞      français フランス語      romaji ローマ字

<article>
  <vr>haichi</vr><vj> 配置 </vj>
  <cat>n.in.</cat><fra>Placement, arrangement, <cat> f.</cat> disposition, mise en
ordre, répartition</fra></r><j> ...suru </r><j> する </j> <cat>v.t.</cat><fra>Arranger, placer,
répartir, disposer.</fra></r><j> Endô ni junsâ wo—suru </r><j> 沿道に巡者を -- する
</j><fra>Poster des agents de police le long de la route. </fra></r><j> Teitai ni—suru </r><j> 梯
隊に -- する </j><fra>Disposer en échelons. </fra>
</article>

```

Figure 9: Information tagging for the « 配置 » (haichi) entry of the Cesselin dictionary
 Finally, the only remaining segment to tag was the French translation of the headword.

Note: because of optical recognition errors, it is complicated to use a language detection tool to differentiate between romaji and French.

4.9 Entry Structuring

During this step, the aim is to structure the articles according to the normative part of the LMF standard (Francopoulo et al., 2009). This will allow automatic export into the informative part of the standard (LMF syntax). The normative part specifies the structure of different blocks but gives no constraint on how to represent them (XML elements or attributes, name of the elements, etc.). It is therefore possible that the resource complies with the LMF standard while keeping its own tags.

The informative part of the standard provides an example of LMF syntax but we consider that it is not convenient to use (& Enguehard Mangeot, 2013). So we choose to use our own tags. The structuring consists primarily in gathering informations about the word form into one "<forme>" block and each word sense into a "<sens>" block. Examples in the Cesselin are not attached to a particular word sense. Therefore, we did not separate the examples into different sense blocks.

```

<article>
  <forme><vedette> <vr>haichi</vr><vj> 配置 </vj></vedette> <cat>n.in.</cat></forme>
  <sens><fra>Placement, arrangement, <cat> f.</cat> disposition, mise en ordre,
répartition</fra></sens>
  <exemples>
    <exemple>
      </r>...suru </r><j> する </j><cat>v.t.</cat><fra>Arranger, placer, répartir,
disposer.</fra>
    </exemple>
    <exemple>
      </r>Endô ni junsâ wo—suru </r><j> 沿道に巡者を -- する </j><fra>Poster des
agents de police le long de la route.</fra>
    </exemple>
    <exemple>
      </r>Teitai ni—suru </r><j> 梯隊に -- する </j><fra>Disposer en échelons.</fra>
    </exemple>
  </exemples>
</article>

```

Figure 10 : Structuring of the « 配置 » (haichi) article of the Cesselin dictionary

4.10 Headword Verification

Once the dictionary is structured, at this stage, we need to detect a maximum of potential errors to in order mark them so they can be easily corrected later online by users of the dictionary. We also add additional information such as furigana for Japanese segments.

A first detection consists in certifying the presence of the headwords in other dictionaries. To this end, we used the 'super daijirin' English-Japanese dictionary (Matsumura, 2006) included in MacOs to program a tool to automate the dictionary lookup. We have also downloaded and installed the JMdict for a better verification. The algorithm is as follows:

- Generation of the hiragana from the headword in romaji and adding it in the block form of article;
- Verification of the presence of the headword in kanji in the 'super daijirin'.
 - if not present, verification in the JMdict.
 - if the headword in kanji is attested in one of these two dictionaries, then check if the hiragana corresponds.
 - if not, convert the hiragana found in one of the verification dictionaries into romaji and then calculate the approximate distance between the two romaji using the Levenshtein algorithm.
 - if the distance is small, replace the romaji and hiragana of the headword by those found in the verification dictionary
 - if the distance is large, identify the problem for later correction.
 - if the headword in kanji is not attested, use the hiragana to seek an alternative proposal in the verification dictionaries
 - if an alternative proposal is found, then add it in the article as a possible alternative.
 - if no alternative is found, use the Google transcription API to offer an alternative.
- if the headword in kanji is empty due to an OCR error, check the hiragana in the two verification dictionaries
 - if there is only a single headword in kanji in the two verification dictionaries, then replace the empty headword in kanji by that found in the verification dictionary.

Optical reading errors are frequent on letters with macron in romaji. It is used to generate the hiragana. Therefore, for comparison with the hiragana with the verification dictionaries, all variants with and without long vowel are generated (\bar{a}/a , \bar{i}/i , \bar{u}/u , \bar{e}/e , o/\bar{o}). This avoids over-detection problems.

The page number in the original printed version of Cesselin is added to each article. This later will allows to display a link to the PDF file of the scanned page allwoing contributors to correct OCR errors by viewing directly the source file.

Finally, about one out of two headwords has been certified in another dictionary, and about 10% of headwords in kanji remain empty.

4.11 Error Detection for French

The detection of potential errors in French is carried out using the tree tagger⁷ morphological analyzer. Each French sentence is sent to the parser. If an unknown word is detected, it is tagged for later correction. If an unknown word is detected, it is marked for subsequent correction.

Example: "L'un des huit enfers glacés du boaddhisme.". "boaddhisme" was not recognized by the analyzer. In this case, it is an OCR error. The correct word is "bouddhisme".

This step could be refined further. Indeed, all the Latin words (eg: name of plants) were not recognized by the analyzer. On the other hand, it should be possible to automatically correct errors such as the one mentioned in the example.

4.12 Error detection for Japanese

The Japanese analyzers can not detect spelling errors such as the ones found in French as the Japanese language does not use separators and all kanji have a meaning. For the detection of potential errors, we compared the romaji transcription with the Japanese version. First we converted the romaji segment into hiragana and then we used the Mecab⁸ Japanese morphological analyzer to generate the furigana of the kanji included in the Japanese corresponding text segment. We then compared the hiragana resulting from the conversion of romaji with the furigana coming from the analyzer. Comparing the hiragana is done by generating all variants as shown in 4.10. When a difference is found, it is tagged for subsequent correction.

- Example: 早いが重宝 [はやいがちようほう] and « hayai ga jūhō » [はやいがじゅうほう] ;
- Example: の徴 [のしるし] and « no kizashi » [のきざし].

Variants with and without sokuon “っ” are generated during the comparison (see 4.7.3).

At this stage, the hiragana was added to the Japanese examples thanks to the output of the analyzer.

This step could also be improved. Indeed, the two examples given above are not real errors but come from the fact that there may be several possible furigana for the same kanji. It should therefore be interesting to use the output of an analyzer giving all possible solutions.

5 Conversion of Wikipedia Links

The coverage of the Cesselin dictionary is already substantial: more than 82,000 articles. However, its release date is 1939, before World War II that resulted in the occupation of Japan by the US military from 1945 to 1952. Since that time, many English words have been incorporated into Japanese after being transcribed into katakana. A modern Japanese dictionary can not ignore them. We therefore reused two free resources available to complete the first set of data from the Cesselin: Wikipedia and JMdict.

We had originally planned to use Wikipedia, but we found that most Japanese translations in the French Wiktionary actually came from translation links between Wikipedia pages. Thus, we preferred to directly use the original resource. So we picked the links of the Japanese Wikipedia

⁷ <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

⁸ <https://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

Japanese to the French and English Wikipedia pages.

5.1 Conversion Process

For each language, Wikipedia offers to download all data as database export in SQL format⁹. We download the information on each page such as the ID and title of the page (jawiki-latest-page.sql.gz) and links to pages in other languages (jawiki-latest-langlinks.sql.gz). We then import that data into a MySQL database and then we create a new table called "translation" initially containing the identifier of each page, its title, the title of the linked pages for French and English ones and the language of the linked page (French or English).

5.2 Extraction of the hiragana

The title of Japanese wikipedia pages are in Japanese (kanji + kana). There is no kanji reading (furigana) or pronunciation (romaji). We must find a way to get them. The use of a morphological analyzer is not appropriate here because there are many proper names in the titles of pages and they are not all in the analyzer dictionary. However, in the first sentence of every Japanese wikipedia page, the hiragana reading of the title is indicated in parentheses.

With API Wikipedia¹, we automatically recover a sample of each page that contains the first sentence and we analyze it to extract the hiragana we include in a new field to the table "translation" created in the previous step.

Le titre des pages du wikipedia japonais sont en japonais (kanji+kana). Il n'y a pas de lecture des kanji (furigana) ni de prononciation (romaji). Il faut donc trouver un moyen de les obtenir. L'usage d'un analyseur morphologique n'est pas judicieux ici car il y a beaucoup de noms propres dans les titres des pages et ceux-ci ne sont pas tous dans le dictionnaire de l'analyseur. Par contre, chaque page du wikipedia japonais indique dans la première phrase la lecture du titre en hiragana entre parenthèses.

Avec l'API de Wikipedia¹⁰, nous récupérons automatiquement un extrait de chaque page contenant la première phrase et nous l'analysons pour en extraire le hiragana que nous incluons dans un nouveau champ de la table « traduction » créée à l'étape précédente.

5.3 Generation of the romaji

As explained in 3.2.2, the romaji cannot be automatically generated from the hiragana if it contains the couple of letters “おう” or “うう”. Furthermore, while Japanese is a language without separators, it is customary to add spaces between words and capitalization at the beginning of proper names in the romaji transcriptions, which greatly facilitates reading. If the romaji is generated from the hiragana (or kanji), there will be no spaces.

We observed that pages that are translations in other languages of Japanese proper nouns often indicate in the first sentence the Japanese spelling and the romaji of these nouns.

Example: Le parc quasi national d'Abashiri (網走国定公園, Abashiri Kokutei Kōen) est un parc quasi national situé sur la côte Nord-Est de l'île de Hokkaidō au Japon.

⁹ <https://dumps.wikimedia.org/jawiki/latest/>

¹⁰ <https://ja.wikipedia.org/w/api.php>

Like for the extraction of hiragana, we use the Wikipedia API to extract the romaji translations of Japanese pages.

When no romaji can be found in the translation page, we will seek the readings of each kanji that composes the Japanese word in the Kanjidic dictionary¹¹. Several types of readings are associated to one kanji: onyomi (Chinese origin) kunyomi (Japanese origin), okurigana (with additional hiragana for termination), nanomi (for proper names).

Compositional phenomena in Japanese must then be taken into account: rendaku (consonant harmonization, e.g. か → が); sokuon (geminate consonant: つ → っ); renjo (doubling of the 'n' character, e.g. ん あ → な); etc.

Then, the romaji generation is performed based on the hiragana conversion.

- E.g.: 東京 + [とうきょう] 東 = とう; 京=きょう ; とう+きょう => tōkyō
- E.g.: 子牛 + [こうし] 子= こ; 牛=うし ; こ+うし => koushi

5.4 Selection of the headwords to import

At this point, articles that we will be imported into the Cesselin dictionary have to be selected. The aim is not to import all the articles in order to "make up the numbers," but only those whose headwords are attested in other resources. First, we will select the available articles with French translations. Articles with English translations only will be imported after the conversion of the JMdict dictionary. The selection algorithm is the following:

- Check that the page title (headword in Japanese) is in the Daijirin dictionary;
 - If so, check if the page title is already in the Cesselin;
 - If so, add the link to Wikipedia articles in the Cesselin article,
 - If not, check if the romaji is in the Cesselin;
 - If so, check whether all the Japanese headwords of the Cesselin articles were verified during step 4.10 ;
 - If all the headwords are verified then import the article;
 - If the romaji is not in the Cesselin then import the article.

This algorithm does not guarantee a selection without problems. Indeed, because of potential errors in optical character recognition, it is possible that a headword may be imported when it is already in the Cesselin if the kanji is empty and if the romaji has an error. In this case, once the romaji and kanji of the original Cesselin article are verified and corrected, a search for Japanese-headword + hiragana-headword duplicates will allow us to remove duplicate articles. Another possible problem is to not import an article because the romaji is already in the Cesselin but not the kanji. In this case, once all the Japanese-headwords associated to that romaji will be verified, it will be possible to re-import the missing article.

A total of 23,456 articles were generated from Wikipedia links and imported into the Cesselin. Of

11 <http://www.csse.monash.edu.au/~jwb/kanjidic.html>

these, 20,825 articles are translated into French and 2,631 articles translated into English. The latter articles were imported after the JMdict conversion (see next section).

6 Conversion of the JMdict dictionary

JMdict is a Japanese → English dictionary (see Figure 4). Since the beginning, in order to increase the coverage of our dictionary, we decided to import JMdict entries though most translations are in English and not in French. On the one hand, English being close to French and much studied, most Francophones can understand written English and secondly, due to the lack of French-Japanese lexical resources, many French speaking learners of Japanese use English-Japanese dictionaries anyway. However, although we had the technical possibility of crossing JMdict with a French-English dictionary, we decided to leave the English translation as is to avoid misinterpretations. Contributors may themselves offer online translation into French when they lookup these articles.

Each JMdict article contains for the headword, a list of words in kanji (`k_ele`) and a list of words in hiragana or katakana (`r_ele`). Each word in kana is sometimes followed by a restriction list to indicate to which word in kanji it corresponds. On the other hand, if, in Japanese text, the word is written in hiragana or katakana, the word list in kanji may be empty. Therefore, reading the headwords (kanji + hiragana reading) is not immediate and demand to compute the correct correspondences. There is no transcription in romaji.

For every headword, we then clarified information and generated a list of headwords consistent with the structure chosen for Cesselin (see 3.2.2): if the `k_ele` list is empty, we copy the items from the `r_ele` list into it. If a `r_ele` element is written in katakana, it is copied in the `k_ele` list and then replaced by the hiragana in the `r_ele` list. Then each element of the `k_ele` list is a Japanese headword of which is linked an `r_ele` item in hiragana. If there are several readings for the same word in kanji, the Japanese headword is duplicated for each corresponding hiragana. Then, the romaji generation algorithm of section 5.3 is reused to generate the romaji from hiragana and kanji.

Then the Japanese kanji headwords have been simplified as in Section 4.7.2 and then the resulting duplicated headwords removed.

Finally, the algorithm of Section 5.4 is reused to select articles to import. When importing articles in the Cesselin dictionary, the unique identifier of the article in the JMdict is stored for future reference.

A list of articles to be translated primarily in French is also generated from the `JapFreqList_5109_Novels` frequency list (see 4.6.4).

A total of 47,810 articles from the JMdict in which 2,521 translated into French and 45,289 translated into English were imported in the Cesselin.

7 Online release on the Jibiki platform

7.1 Site Description

The Project¹² web site is built around the Jibiki platform (Mangeot, 2006) for the management of heterogeneous lexical resources online. The platform is programmed using Enhydra, a java object

¹² <http://jibiki.fr>

server based on a 3/tiers architecture. The database used for the data layer is Postgres. This platform, developed and continuously improved since 2001 is freely available with a LGPL licence on the LIG laboratory forge¹³.

In addition to the functions for the resources management (dictionary lookup and editing), several pages were added:

- a blog as homepage;
- a bilingual aligned corpus lookup interface (KWIC);
- a module for active reading.
- a download page for retrieving all the data (dictionaries and corpora).

The site is available in three languages : French, English and Japanese.

7.1.1 Home page: blog

Le blogue permet de présenter les dernières nouvelles du projet et présenter le meilleur contributeur de chaque mois. Chaque article est traduit dans les trois langues du site : français, anglais et japonais.

The home page consists of a standard consultation interface (see 7.1.4), a short text presenting the project, lists of articles to be corrected in priority and a blog programmed using WordPress.

Two lists of articles are displayed: the articles of the Cesselin dictionary which headword kanjis were not recognized during the optical reading (see 4.6.4) and the articles from the JMdict whose translation is in English and must be translated into French (see 6). For each list, the ten most frequent words are displayed. A button for calculating lists is provided in order to update the lists if corrections have been made.

The blog allows to present the latest news about the project and present the best contributor of the month. Each article is translated into the three languages of the site: French, English and Japanese.

7.1.2 Bilingual Aligned Corpora

The consultation module of the French-Japanese aligned bilingual corpora is programmed in perl using the IMS Open Corpus Workbench¹⁴ platform and its lookup tool Corpus Query Processor (CQP). This allows to use a regular expressions language to build complex queries. For example, the query "inter(ê|e)(t|ss)(é|e)(r|s)?" will get the words "intérêt", "intérêts", "intéressé", "intéresser", "intéressée", "intéressées". It is also possible to combine the levels of annotations. For example, the query [(lemma="sous.+") & (Cat="V. *")] will get all the conjugated forms of verbs beginning with the prefix "sous". This module is derived from a first version used in the GDEF project for a n Estonian-French corpus¹⁵.

The data come from one hand from the OPUS project¹⁶ for software (KDE, OpenOffice), the Koran and movie subtitles (OpenSubtitles) and from the other hand from corpora we built ourselves with

13 <https://jibiki.ligforge.imag.fr>

14 <http://cwb.sourceforge.net>

15 <http://corpus.estfra.ee>

16 <http://opus.lingfil.uu.se>

texts found on the Web (Le Monde Diplomatique, the Bible, a Franco-Japanese tax Convention and the Universal Declaration of Human Rights).

These data, aligned at the paragraph level, were then tagged using Tree tagger for French (see 4.11) and Mecab for Japanese (see 4.12). The furigana was added to the Japanese text, allowing to write queries at the kanji reading level. For example, a search for the word "はな" (hana) will get the words 花 (flower), 鼻 (nose), etc.

Figure 11 shows the search result for the Japanese word "配置" (haichi) in the “Le Monde Diplomatique” corpus. The corpus is parallel. Thus, it is also possible to search for French words.

At present, the whole corpus size is approximately 6 million words including:

- Newspapers: Le Monde Diplomatique (288,745 words);
- Legal texts: Franco-Japanese Tax Agreement (17,443 words) and Universal Declaration of Human Rights (2,208 words);
- Software: KDE 4 (1,179,000 words) and OpenOffice 3 (569,903 words);
- Religious texts: the Bible (904,914 words) and the Koran (Tanzil) (192,905 words);
- Movie subtitles (OpenSubtitles) (4,714,000 words).

9 occurrences trouvées dans 8 extraits	
<p>その配置を読み取って霊の加護を祈った後、蓮の花と樹皮の粉末でできた薬を たっぷりと患者に与える。</p>	<p>Pour soigner ses malades, M. Domingo revêt un tissu bariolé et des colliers de coquillages, jette dix-sept os d'agneau au sol, interprète leur disposition, puis invoque l'aide des esprits avant de prodiguer des traitements à base de fleur de lotus et de poudres d'écorce.</p>
<p>彼女は「産業の再建、および産業の再配置」を訴え、これが「唯一、真 のエコロジーにかなう」政策であるとする。</p>	<p>La encore, la dirigeante d'extrême droite puise allègrement dans des propos tenus par le bord opposé.</p>
<p>というのも、メキシコとの国境には1マイル [約1.6Km—訳注] ごとに 10人の警備隊員がすでに配置されているからだ。</p>	<p>« Les élus qui ont concocté la réforme de l'immigration semblent avoir voulu créer un chemin vers la citoyenneté plus décourageant qu'accessible », tranche la revue de gauche radicale Counterpunch.</p>
<p>中心市街に大きな建物を、その周りの近郊市街地にやや小さな建物を 配置し、住宅は最も外側の郊外地区に配置します」。</p>	<p>De grands immeubles au centre, des plus petits dans la première couronne autour de celui -ci, des maisons dans les quartiers les plus périphériques ».</p>
<p>すなわち、当該難民は当局によって国内の様々な場所に再配置されるとい うもので、それは「ブルンジの飛び地」がタンザニアにできないようにする ためだった。</p>	<p>Les cas de la Syrie et des anciens réfugiés irakiens qui y vivent, ainsi que celui de l'Afghanistan et de ses 5,7 millions de réfugiés (pour la plupart de longue durée), figuraient au programme.</p>
<p>2012年7月22日付『フィガロ』紙は、シリアの「化学兵器が監視下に置かれ ている」「化学兵器の配置を見極めるためにアメリカ特殊部隊が配備され た」と伝えている。</p>	<p>Et un diplomate en poste en Jordanie avertit : « C'est la menace des armes chimiques qui peut déclencher une intervention américaine ciblée. »</p>
<p>彼らの多くは、政治配置図の極左に位置し、トロツキスト系（レバノンの《 社会主義フォーラム》、エジプトの《革命社会党》）あるいは毛沢東主義 （モロッコの《民主主義への道》）の場合もある。</p>	<p>Souvent situés à l'extrême gauche du spectre politique, ils sont parfois de filiation trotskiste - le Forum socialiste au Liban, les socialistes révolutionnaires en Egypte - ou maoïste - la Voie démocratique au Maroc.</p>
<p>一方、アラブ世界における政治配置図の左派に位置している大半の勢カ グループは、シリア騒乱に対して慎重な距離感を保つことを特徴としてい る。</p>	<p>A l'inverse, une distance prudente à l'égard de la révolte syrienne caractérise la majorité des forces se situant à la gauche du spectre politique dans le monde arabe.</p>

Figure 11: Lookup result of the word « 配置 » (haichi) in the *Monde Diplomatique* corpus

7.1.3 Active Reading Module

The active reading module offers a reading aid for a user who knows a language but do not master it. The user enters a text in that language and the module will displays a word-for-word translation. In our case, the French speaking user can enter Japanese text and Japanese speaking user a French

text. Then, the module adds the pronunciation of words or furigana in the case of the Japanese and translations of each word. In order to avoid to interfere with the reading, the translations are displayed only when the user points a word with the mouse. See Figure 12 for an example of display on a Japanese text with the mouse pointed at the word "時代" (jidai).

The module first sends the text to a morphological analyzer: Tree tagger for French (see 4.11) and Mecab for Japanese (see 4.12) then retrieves the lemma for each word. It then uses the REST⁶ API of the Jibiki platform¹⁷ to lookup different resources. For Japanese text, the furigana is obtained with the morphological analyzer and translations with the Cesselin dictionary. For French text, translations are available for the moment in English until we obtain a French → Japanese dictionary. The pronunciation and translation are obtained using the FeM dictionary (Gut et al., 1996).

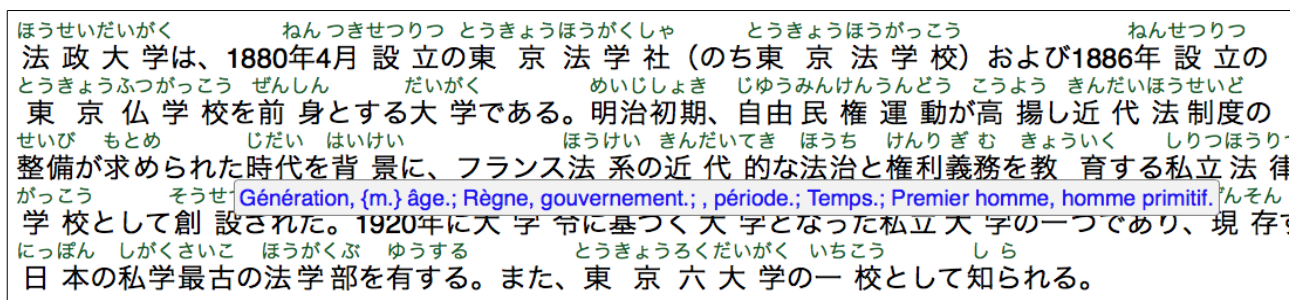


Figure 12: Display of a Japanese text with the mouse on the « 時代 » (jidai) word for the module

7.1.4 Standard Lookup

The standard lookup interface allows a user to lookup Japanese words typed in romaji, kana or kanji. It displays an advanced view of a printed dictionary: the left side displays the headwords of the immediate vicinity of the targeted word listed alphabetically. An infinite scroll can be used to browse all the dictionary headwords in alphabetical order. When a headword is clicked on the left side, the full article is displayed on the right side.

A menu is displayed at the top right of each article. It includes links to the editing form, the modification history, the XML source of the article, the original scanned page of the dictionary in PDF format and the result of the headword lookup in the bilingual corpus.

¹⁷ <http://jibiki.fr/jibiki/Api.po>

These can be combined in a single search. It is possible for example to look for articles which romaji starts with a "b" and pertaining to the botany domain (see Figure 15). The search results are displayed in an alphabetical list on the left side of the window. If the number of result is greater than 100, an AJAX request is made to get from the server the following results in alphabetical order, like for the standard consultation interface.

Figure 15: Advanced Lookup Interface with combination of two search criteria

7.2 Online Editing

To edit an article online, the user must be previously registered on the Jibiki platform.

7.2.1 Quick Editing

When viewing an article, it is possible to perform small modifications directly on the Web page. To do this, a double-click on the segment to modify turns it into a text field with an 'ok' button on the right to validate the edition (see Figure 16).

This editor is programmed using AJAX¹⁸ technology and it uses the REST⁶ API of the Jibiki platform¹⁷ to interact with the server. When clicking on the 'ok' button, the new string is sent to the server with the Xpath¹⁹ of the edited segment and the article ID.

Figure 16: Quick Editing of the French translation of the first example of the « 配置 » (haichi) article

7.2.2 Full Editing Form

The full editing interface is automatically generated from the articles noted as XML schema²⁰ (Mangeot & Thevenin, 2004). It consists of a standard HTML form with interactors for different data types (text field for free text, drop down menu for closed lists of values, check boxes for

18 [https://fr.wikipedia.org/wiki/Ajax_\(informatique\)](https://fr.wikipedia.org/wiki/Ajax_(informatique))

19 <http://www.w3.org/TR/xpath/>

20 <http://www.w3.org/XML/Schema>

booleans, radio buttons for choices of values, etc.) as well as more complex interactors to manage lists of objects (for example, adding or deleting one example in a list of examples by pressing the buttons "+" and "-" in Figure 17).

The image shows a complex web form for editing dictionary entries. It features several expandable sections, each with a header bar containing '+' and '-' icons. The sections are:

- Liste de sens**: Contains a 'sens' field with dropdowns for 'DOMAINE', 'Gram', and 'Non reconnu', followed by a text field for the sense definition.
- Liste d'exemples**: Contains an 'Exemple' section with fields for 'romaji', 'Japonais', and 'Français' (with its own dropdowns for 'DOMAINE', 'Gram', and 'Non reconnu').
- Liste de sous-sens**: Contains a 'sous-sens' field with dropdowns for 'DOMAINE', 'Gram', and 'Non reconnu', followed by a text field for the sub-sense definition.
- Liste de sous-sens** (second): Contains an 'Exemple' section with fields for 'romaji', 'Japonais', and 'Français'.

 The form uses a consistent color scheme with orange headers and blue text for labels and titles.

Figure 17: Full editing form for the examples of the « 配置 » (haichi) article

To access the full editing form, simply click on the link "Edit" menu at the top right of each article. The user can then edit the targeted article. At the end of the work, s/he previews the changes. They can be temporarily saved by affecting the "draft" status to the article. Modifications can also be canceled or saved permanently in the database. All previous versions are kept in the database. This allows the administrator to go back if systematic errors are detected for a specific contributor or search criteria.

7.3 Statistics

7.3.1 Number of articles

The dictionary contains 153,897 articles in total. Among these,

- 82,663 articles come from the Cesselin dictionary which:
 - 10 243 articles (12.39%) with headwords in kanji not recognized by the OCR;
 - 0 259 articles (48.70%) with headwords not yet verified;
- 47,721 articles come from the JMdict dictionary;

- 23,512 articles come from links between Wikipedia pages.

Of the 153,897 articles, 47,813 articles (31.07%) are translated into English (and therefore must be translated into French).

7.3.2 Number of contributions

After 3 months of online availability, the 25 October 2015, the site recorded 2,639 modifications of articles, among them:

- 86 headwords in kanji added,
- 225 headwords in kanji verified,
- 132 translations in French.

7.3.3 Number of visits

Three months after its opening, on July 22, 2015, the site registered 664 visits the 22 October 2015.

8 Conclusion

We showed in this paper that it is possible to launch a project for the collaborative construction of dictionaries on the Web using copyright free resources which allows to get a usable dictionary immediately. The site is open for only 3 months, but the high number of contributions already shows that the experiment is a success.

The methodology described in this article to convert a printed dictionary into XML can be reapplied to any printed resource (and there is a lot!).

The constitution of this resource is a starting point for future research.

Regarding the production of data, we plan to launch a similar process to retrieve a French → Japanese dictionary. We also plan to expand the current resource by adding new information (counters, quantifiers, frequencies of occurrences in corpora, etc.).

Regarding the use of data, obtaining a French → Japanese resource will allow us to experiment the convergence towards a pivot macrostructure (Mangeot et al., 2004). The examples and their translation can be reused to build an aligned bilingual corpus which can be then used for example to build a statistical machine translation system like Moses.

9 Acknowledgements

This project was made possible through the Hosei International Fund (HIF) program, which has allowed us to be welcomed at Hosei University, Tokyo from October 2014 to August 2015.

10References

- Apel U. (2002) WaDokuJT - A Japanese-German Dictionary Database. Papillon 2002 Seminar, 16-18 July 2002, NII, Tokyo, Japan, 13 p.
- Berment V. (2004) *Méthodes pour informatiser des langues et des groupes de langues "peu dotées"*. Thèse de nouveau doctorat, Université Joseph Fourier Grenoble I, Grenoble, France, 277 p.
- Breen JW. (2004) *JMDict: a Japanese-multilingual dictionary*. In: Coling 2004 workshop on multilingual linguistic resources, Geneva, Switzerland, pp. 71-78.

- Cesselin G. (1940) *Dictionnaire japonais-français*, Maruzen, Tokyo, juillet 1940, 2340 p.
- Desperrier J-M. (2002) *Analyse [sic] of the results of a collaborative project for the creation of a Japanese- French dictionary*. In: Proceedings of Papillon 2002 Seminar, Tokyo, Japan.
- EDR (1993) EDR Electronic Dictionary Technical Guide. Project Report, n°-042, Japan Electronic Dictionary Research Institute Ltd., 16 August 1993, 144 p.
- Enguehard Ch. & Mangeot M. (2013) *LMF for a selection of African Languages*. Chapter 7, book "LMF: Lexical Markup Framework, theory and practice", Ed. Gil Francopoulo, Hermès science, Paris, France, 17 p.
- Enguehard Ch., Mangeot M. (2014) *Computerization of African languages-French dictionaries*. Proc. of Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL), LREC 2014 workshop, Reykjavik, Island, 27 May 2014, 8 p.
- Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M., Soria C. (2009). *Multilingual resources for NLP in the Lexical Markup Framework (LMF)*. Language Resources and Evaluation, Vol. 43, pp. 57–70. ISBN: 10.1007/s10579-008-9077-5.
- Griollet Pascal (2008) *Plus de « cent cinquante ans » d’histoire de l’enseignement du japonais*. Le japonais au XXI^e siècle - Actes des États généraux pour l’enseignement du japonais en France, pp 47-63.
- Gut Y., Puteri R., Megat R., Zaharin Y., Chuah Choy K., Salina A. S., Boitet Ch., Nédobejkine, N. , Lafourcade M. et al. (1996) *Kamus Perancis-Melayu Dewan, dictionnaire français-malais*. Dewan Bahasa Dan Pustaka, Kuala Lumpur, 667 p.
- Hisamatsu K., Obataya Y., Hayakawa, F. et al. (2009) *Dictionnaire Japonais-Français / Français-Japonais*, Assimil, Paris, 1280 p. ISBN 978-2-7005-0445-3
- Koichi Hirao (2010) *La rénovation du dictionnaire français-japonais dans les années 1980 : le Dictionnaire général français-japonais de Hakuishia et le Shogakukan Robert Grand Dictionnaire français-japonais* dans Heinz, Michaela (éd.), *Cultures et lexicographies*, Berlin, Frank & Timme, p. 103-111.
- Mangeot M. (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, 280 p.
- Mangeot M. (2006) *Papillon project : Retrospective and perspectives*. In P. Zweigenbaum, Ed., *Acquiring and Representing Multilingual, Specialized Lexicons : the Case of Biomedicine*, LREC workshop, Genoa, Italy, 6 p.
- Mathieu Mangeot (2014) *MotàMot project: conversion of a French-Khmer published dictionary for building a multilingual lexical system*. Proc. of LREC 2014, Reykjavik, Island, 28-30 May 2014, 8 p.
- Mangeot M., (2015) *Construction of an open-source multilingual lexical system targeted on French and Japanese through contributive and automatic methods*. Internal Report, Hosei University, 12 p.
- Mangeot M. & Chalvin A. (2006) *Dictionary building with the Jibiki platform : the GDEF case*. In LREC 2006, Genoa, Italy, pp. 1666–1669.
- Mangeot M., Sérasset G. & Lafourcade M. (2004) *Construction collaborative d’une base lexicale multilingue*. *Traitement Automatique des Langues*, vol. 44(2), pp. 151–176.
- Mangeot M. & Thevenin D. (2004) *Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project*. Proc. of COLING 2004, ISSCO, Université de Genève, Switzerland, 23-27 August 2004, vol 2/2, pp 1029-1035.
- Matsumura A. (2006) *Daijirin Japanese-English dictionary (大辞林)*, 3rd edition, Sanseido, Tokyo, 2974 p. ISBN 4-385-13905-9.
- Mel’čuk I., Clas A. & Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*. Universités francophones et champs linguistiques. Louvain-la Neuve : AUPELF-UREF et Duculot, 256 p.
- Polguère A. (2008) *Lexicologie et sémantique lexicale. Notions fondamentales*. Paramètres, 304 pages. Les Presses de l’Université de Montréal, Montréal, 2e édition. Nouvelle édition revue et augmentée.
- Raguet É. & Martin J-M. (1953) *Dictionnaire français-japonais*, Hakuishia, Tokyo, 1467 p.