Construction collaborative d'un dictionnaire japonais-français de qualité, à large couverture et libre de droits

Mathieu Mangeot[1,2]

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

[2] GETALP, 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

Ce projet de recherche se situe dans le domaine du traitement automatique des langues (TAL), à la croisée de l'informatique et de la linguistique, plus précisément sur la lexicographie et la lexicologie multilingues.

Lors d'un premier long séjour au Japon de novembre 2001 à mars 2004, nous avions fait le constat que les ressources lexicales français-japonais disponibles sur le Web étaient quasi inexistantes. Ce qui avait donné naissance au projet Papillon de construction d'une base lexicale multilingue à structure pivot (Sérasset et al., 2001). Depuis, des progrès ont été faits dans plusieurs domaines (technique, théorique, social) (Mangeot, 2006) mais la production concrète de données a très peu progressé. D'autre part, la réutilisation de ressources lexicales est à la mode (désambiguïsation lexicale, utilisation de ressources en source ouverte (Wiktionary, dbpedia), fusion avec des ontologies, etc.). Même si elles permettent de consolider et d'élargir la couverture des ressources existantes, ces expériences partent toujours de données créées à la main par des lexicographes.

Partant de ce constat, nous avons défini le projet suivant qui consiste à construire un système lexical multilingue riche d'informations avec priorité sur le couple de langues français-japonais. La construction se fera d'une part par la réutilisation de ressources existantes (dictionnaires japonais-autre langue, Wikipedia) et leur exploitation automatique (lecture optique et corrections, calcul de liens de traduction) et d'autre part par des contributeurs bénévoles travaillant en communauté sur le Web. Ceux-ci seront amenés à contribuer sur les articles de dictionnaire en fonction de leur niveau d'expertise et de leurs connaissances dans le domaine de la lexicographie ou de la traduction bilingue.

Les ressources ainsi produites seront libres de droits et destinées à être utilisées aussi bien par des humains via des dictionnaires bilingues classiques que par des machines pour des outils de traitement automatique de la langue (analyse, traduction automatique, etc.).

Nous effectuerons d'abord un état des lieux des dictionnaires bilingues français-japonais, puis nous décrirons la ressource que nous souhaitons construire. Les parties suivantes concernent la récupération et la conversion de trois ressources : le dictionnaire Cesselin en version imprimée, les liens entre langues de Wikipedia ainsi que le dictionnaire JMdict en version électronique. Enfin,

nous terminerons par la mise en ligne de la ressource ainsi construite sur un site Web construit autour de la plate-forme Jibiki, permettant de consulter et modifier les articles en ligne.

2 État de l'art des dictionnaires bilingues japonais

Bien que le français et le japonais soient considérées comme des langues bien dotées au niveau des outils et des ressources linguistiques, le couple français-japonais est considéré comme un couple de langues peu doté. Il existe en effet peu de ressources lexicales bilingues électroniques de qualité et libres de droits. Les corpus bilingues alignés et les systèmes de traduction automatique français-japonais sont logiquement tout aussi rares.

Pour des raisons historiques autant que pratiques, les Japonais ont mis rapidement l'accent sur l'anglais. Le couple anglais-japonais est donc l'un des mieux dotés à l'heure actuelle avec des ressources très conséquentes comme le dictionnaire EDR (1993) et des systèmes de traduction automatique parmi les plus performants.

2.1 Dictionnaires éditoriaux français–japonais

Dans cette partie, nous présenterons les dictionnaires les plus marquants, soit pour des raisons historiques soit par les innovations qu'ils apportent. Il est à noter que jusqu'à très récemment d'une part qu'il existe deux traditions lexicographiques distinctes selon que l'équipe de rédaction est de langue maternelle français ou japonaise et d'autre part que les dictionnaires sont tous monodirectionnels (une langue vers une autre mais pas l'inverse). Il faut attendre 2009 avec la sortie du dictionnaire bidirectionnel d'Assimil (Hisamatsu et al., 2009). D'autre part, jusque dans les années 1950, les auteurs de langue maternelle française son tous des missionnaires catholiques. L'objectif premier était alors de traduire la bible en japonais.

2.1.1 Dictionnaires imprimés japonais→français

1603 : Vocabvlario da Lingoa de Iapam (nippo jisho). Ce dictionnaire japonais→portugais de 32 293 articles rédigé par des missionnaires jésuites portugais est considéré comme le premier dictionnaire bilingue japonais.

1862 : traduction du Nippo jisho en français par Léon Pagès (1814-1886). Celui-ci prend le soin d'ajouter une transcription en katakana des mots japonais (Griolet, 2008).

1904 : dictionnaire Lemaréchal rédigé par Jean Lemaréchal (1842-1912) contenant environ 60 000 articles sur 1 008 pages. Celui-ci abandonne un transcription latine adaptée au français pour le romaji Hepburn (Griolet, 2008).

1939 : dictionnaire Cesselin (Cesselin, 1939) rédigé par Gustave Cesselin (1873-1944), contenant 82 500 articles sur 2 340 p. Il est considéré comme « le meilleur du point de vue de ceux qui étudient la langue japonaise de façon approfondie, car il fournit de nombreux exemples présentés sous forme alphabétique. » (Griolet, 2008).

2009 : dictionnaire japonais Assimil. Ce dictionnaire (Hisamatsu et al., 2009) est à notre connaissance le premier dictionnaire bilingue bidirectionnel (français→japonais et japonais→français) Il contient 24 000 articles sur 1 280 pages. Il contient également 135 000 mots, expressions et traductions, 35 000 exemples d'utilisation. Tous les mots et expressions sont

transcrits en romaji. Cela en fait un outil très utile pour les francophones apprenant le japonais.

2.1.2 Dictionnaires imprimés français→japonais

1864 : dictionnaire futsugo-meiyo (élucidation de la langue française) par Murakami Hidetoshi (1811-1890). Ce savant est considéré comme le premier japonais a avoir appris le français et ceci par l'intermédiaire d'un dictionnaire français-hollandais (Koichi, 2010).

1866 : dictionnaire français-anglais-japonais par l'Abbé Eugène Mermet de Cachon (1828-1871) contenant environ 5 300 articles sur 433 pages. Le japonais a été revue par Léon Pagès.

1887 : dictionnaire universel français-japonais de Nakae Atsusuke et Nomura Yasuaki. Ce dictionnaire est une traduction en japonais du dictionnaire monolingue français Petit Littré. Il marque aussi une première prise en compte de la polysémie.

1905 : dictionnaire français-japonais par Émile Raguet (1854-1929) et Tōta Ono, 1 048 p.

1953 : deuxième édition revue et augmentée (Raguet & Martin, 1953) appelée « Raguet-Martin », rédigée par Émile Raguet et Jean Marie Martin (1886-1975). Ce dictionnaire contient environ 50 000 articles sur 1 445 pages. Il « demeure le seul grand dictionnaire à présenter les traductions japonaises sous forme alphabétique, et donc intelligibles sans connaître les sinogrammes. » (Griolet, 2008).

1983 : dictionnaire franco-japonais de notre époque édité par Mikasa-shobo. Contient environ 42 000 articles sur 1 763 pages. Le romaji est indiqué pour le japonais ainsi que la prononciation des mots français.

1988 : dictionnaire Shôgakukan-Robert. Ce dictionnaire, fruit de la traduction en japonais du dictionnaire monolingue français Robert est un travail considérable. Il reste à ce jour le plus gros dictionnaire français-japonais imprimé puisqu'il contient plus de 100 000 articles. Il n'y a malheureusement pas de romaji (transcription latine) ni de furigana (prononciation des kanji). Il est donc destiné principalement aux japonophones.

2.1.3 Dictionnaires électroniques (denshi-jishō)

Le Crown (Sanseido, 1978), dictionnaire français → japonais contient 47 000 entrées. Il n'y a pas de romaji ni de furigana (voir figure 1).

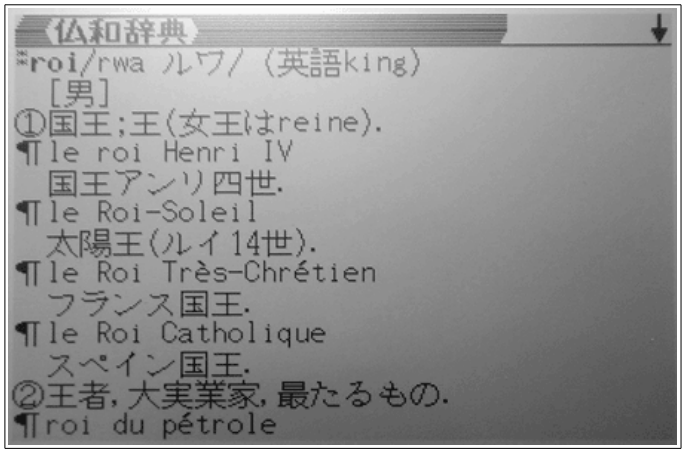


Figure 1 : Capture d'écran du dictionnaire Crown en version électronique

Le Concise (Sanseido), dictionnaire japonais → français contient 38 000 entrées. Il n'y a pas de

romaji ni de furigana (voir figure 2).

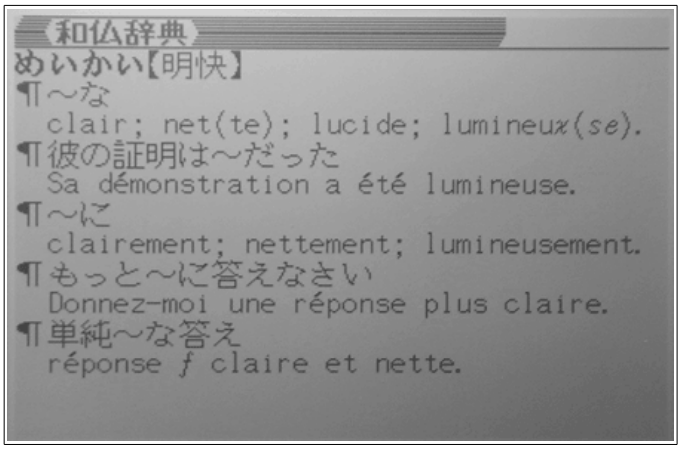


Figure 2 : Capture d'écran du dictionnaire Concise en version électronique

Les francophones ayant un niveau suffisant pour lire le japonais ont certainement besoin d'un dictionnaire de couverture plus large. Ces dictionnaires sont donc destinés aux japonophones.

2.1.4 Conclusion

Les dictionnaires électroniques ne sont pas réutilisables en dehors du support dans lequel ils sont vendus. De plus, ils sont conçus pour des japonophones et leur couverture n'est pas très large.

Les seuls dictionnaires japonais-français existant de bonne qualité et à large couverture sont des dictionnaires éditoriaux qui n'existent qu'au format papier pour lesquels il n'existe pas d'interface de consultation en ligne. Par contre, certains sont suffisamment anciens pour être libres de droits.

La figure 3 montre l'évolution du nombre d'articles dans les dictionnaires français-japonais en fonction des années. On remarquera un pic dans les années 1950. Il doit être possible de réutiliser certaines de ces ressources dans le cadre de notre projet pour construire un dictionnaire de bonne qualité et à large couverture disponible sur le Web, à condition de les actualiser avec du vocabulaire

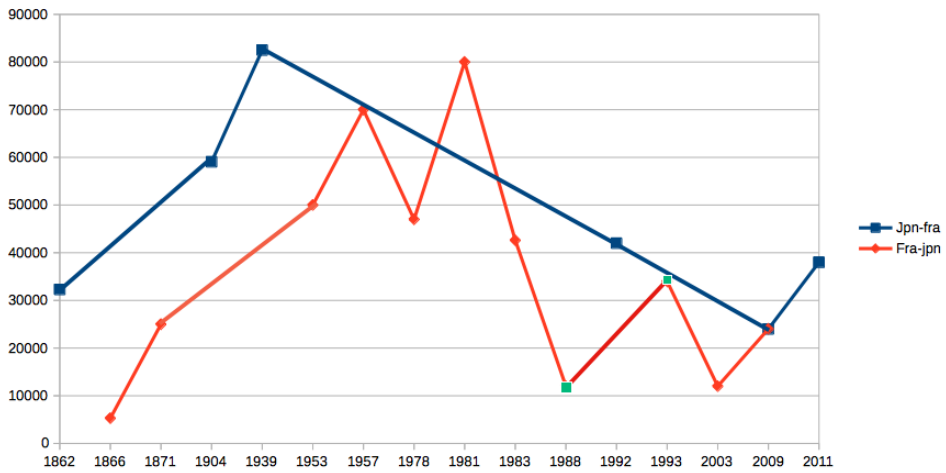


Figure 3 : Évolution du nombre d'articles des dictionnaires français-japonais

moderne.

2.2 Projets Wiktionary

Le Wiktionary français compte actuellement 2,2 Mo d'entrées dont 1,2 Mo d'entrées françaises et avec un peu moins de 7 000 traductions en japonais dont une sur deux environ est un nom propre (c'est souvent une simple transcription en syllabaire japonais du mot-vedette). On trouve également quelques traductions reprises du site dictionnaire-japonais.com (voir plus bas). Les traductions sont indiquées au niveau de l'entrée et non des sens de mot. Il n'y a aucune description de contexte de traduction (glose, exemples), ni d'informations sur la traduction japonaise (classe grammaticale, etc).

Le Wiktionary japonais compte 83 000 entrées dont 26 000 entrées japonaises et 2 800 entrées françaises traduites en japonais (on y trouve des formes fléchies ou formes verbales conjuguées, par exemple 32 entrées pour le verbe « aimer »). La couverture est très insuffisante.

Les projets Wiktionary sont intéressants et à la mode mais ils ont plusieurs limitations :

- La structure des articles est libre. Il n'est pas possible d'utiliser la même microstructure précise pour tous les articles.
- Même s'il est possible de décrire, dans le Wiktionary d'une langue A, un sens de mot d'une langue B dans la langue A, l'interface de départ n'est pas conçue pour rédiger des dictionnaires bilingues. Par exemple, la description du lien inverse langue A→langue B doit être fait à la main dans le projet Wiktionary de la langue B.
- Il n'est pas non plus possible d'ajouter automatiquement des données existantes provenant d'autres sources pour construire un brouillon à raffiner ultérieurement.
- Les contributions sont anonymes. Il n'est pas possible d'utiliser un niveau de qualité des données ou un système de relecture/validation.

Même si le succès du projet Wikipedia peut logiquement nous amener à penser qu'il en sera de même pour la construction collaborative de dictionnaires bilingues ou multilingues de qualité, ce n'est pas le cas. Nous citons à ce propos, Larry Sender, co-fondateur de 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.”.

En effet, chaque article de Wikipedia peut être rédigé par un spécialiste du domaine en question, mais pour un dictionnaire de langue générale, il n'est pas possible de trouver un spécialiste pour seulement quelques articles. Seuls les linguistes spécialistes de la langue et traducteurs professionnels (après avoir été formés en lexicographie) peuvent rédiger un article dans son intégralité.

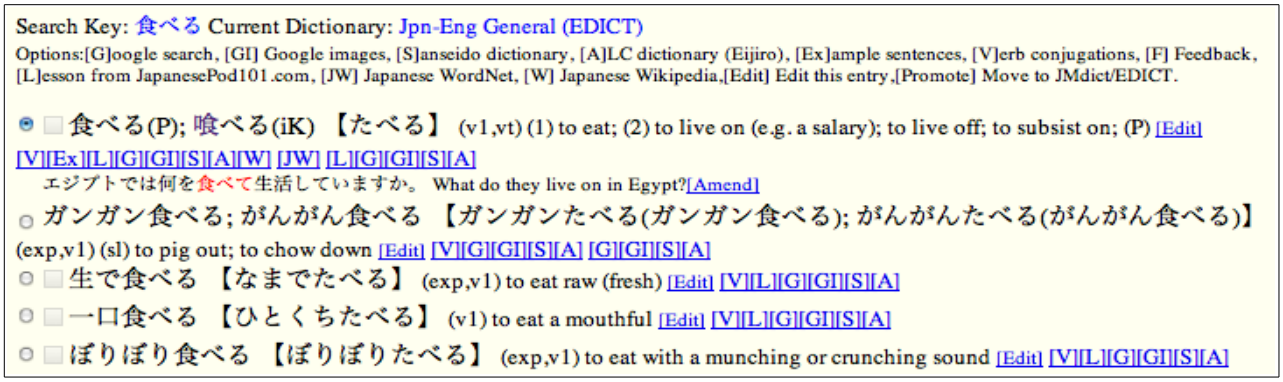
2.3 Ressources japonais–autre langue en ligne

2.3.1 Dictionnaire japonais-anglais : Jmdict

Le JMdict[1] (Japanese-Multilingual Dictionary) (Breen, 2004) est un projet mené par Jim Breen. Il contient 173 000 entrées japonais-anglais avec des ajouts de traductions dans d'autres langues :

1 http://www.csse.monash.edu.au/~jwb/jmdict.html

allemand (provenant de WaDokuJiten), 31 000 équivalents français (provenant du dico FJ), russe, 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] Feedback, [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; がんがん食べる 【ガンガンたべる(ガンガン食べる); がんがんたべる(がんがん食べる)】 (exp,v1) (sl) to pig out; to chow down [Edit] [V][G][GI][S][A] [G][GI][S][A]
生で食べる 【なまでたべる】 (exp,v1) to eat raw (fresh) [Edit] [V][L][G][GI][S][A]
一口食べる 【ひとくちたべる】 (v1) to eat a mouthful [Edit] [V][L][G][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][G][GI][S][A]

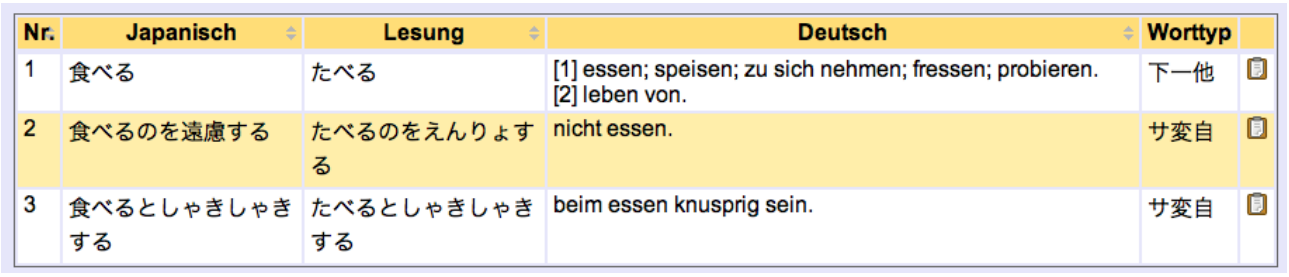
Figure 4 : Article “食べる” (taberu) du dictionnaire JMdict

Avantages : 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.

2.3.2 Dictionnaire japonais-allemand : WaDokuJiten

Le WaDokuJiTen[2] de Ulrich Apel (Apel, 2002) est constitué de plus de 280 000 entrées. Sa large couverture ainsi que sa microstructure sont plus développées que le 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	たべるのをえんりょする	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 : Article “食べる” (taberu) du dictionnaire WaDokuJiTen

Avantages : plus complet que le JMdict en terme de couverture et d'informations, libre de droits et disponible gratuitement au téléchargement.

Inconvénients : tout comme le JMdict, le dictionnaire est unidirectionnel. Il ne comporte pas non plus d'exemples d'usage pour illustrer les contextes de traduction.

Ce dictionnaire est à ce jour la ressource la plus complète japonais-autre langue disponible gratuitement en téléchargement. Il constitue un objectif à atteindre pour notre ressource en termes de couverture.

2.4 Ressources français-japonais disponibles en ligne

2.4.1 Le Projet Dico FJ

Le projet dico FJ, précurseur dans le domaine, a été lancé début 2000 par Jean-Marc Desperrier

2 http://www.wadoku.de

(Desperrier, 2002). Il contient un peu plus de 10 000 entrées provenant de traduction du dictionnaire japonais-anglais JMdict de Jim Breen. Il n'y a pas eu d'évolution depuis 2003.

Avantages : libre de droits et disponible gratuitement au téléchargement.

Inconvénients : en plus des inconvénients du JMdict, on trouve des erreurs de traduction dues au fait que certains contributeurs maîtrisant mal le japonais ont traduit directement les traductions anglaises au lieu des entrées japonaises, ce qui augmente le nombre de contresens.

2.4.2 Dictionnaire-japonais.com

Le projet dictionnaire-japonais.com[3] contient actuellement un peu plus de 40 000 mots. Il constitue un net progrès par rapport aux autres projets de dictionnaire japonais-français en ligne. Chaque utilisateur peut contribuer directement en rajoutant des entrées. La communauté de contributeurs semble assez active comme en témoigne l'activité sur le forum du projet. Les informations disponibles pour chaque entrée sont relativement limitées à un "type grammatical", une "catégorie" (domaine), un registre de langue, et parfois une "origine du mot" (étymologie).

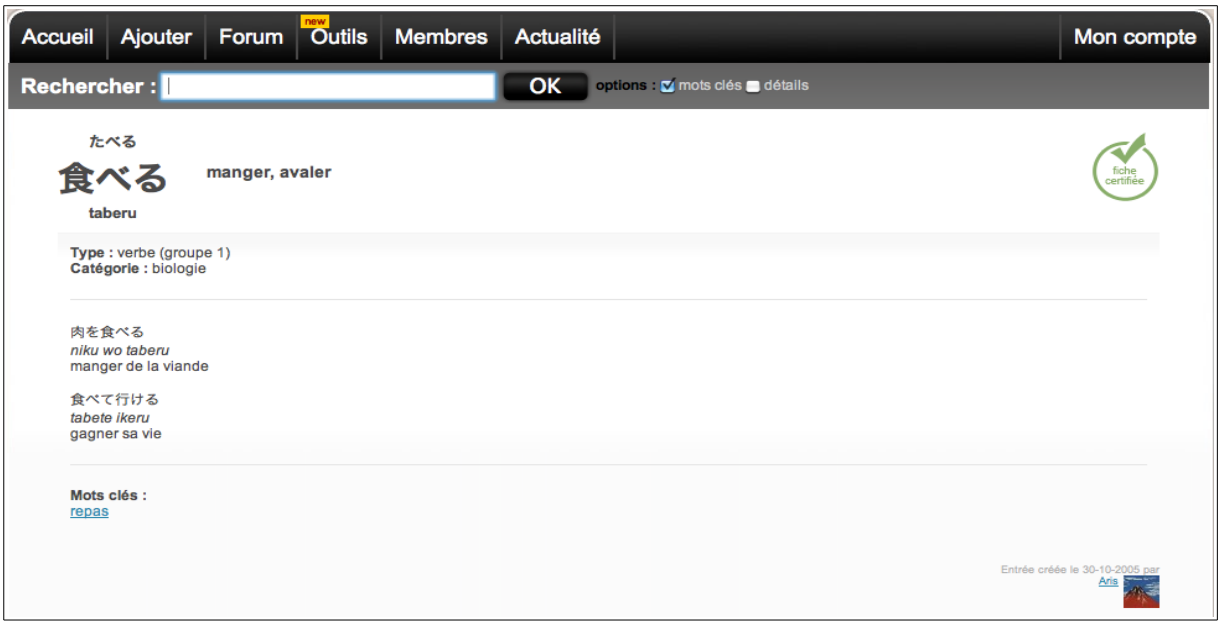


Figure 6: Entrée "食べる" (taberu) du dictionnaire-japonais.com

Avantages : disponible en ligne, couverture un peu plus large que le dico FJ, communauté de active de contributeurs bénévoles.

Inconvénients : en plus des inconvénients du dico FJ, les données ne sont pas disponibles au téléchargement.

2.5 Bilan

Les lexiques japonais-français disponibles en ligne sont d'une part de petite taille et d'autre part tous orientés japonais vers français.

La plupart des dictionnaires français-japonais manque également d'informations pour être utilisés à la fois par des francophones et par des japonophones. Par exemple, il n'existe par exemple pas à

3 http://www.dictionnaire-japonais.com

notre connaissance de dictionnaire présentant à la fois les kanji (idéogrammes), les kana (syllabaires) et le romaji (transcription en alphabet romain).

Les dictionnaires pour japonophones n'indiquent pas le romaji et la plupart du temps, les exemples ne sont écrits qu'avec des kanji sans furigana. Les francophones débutants en japonais ne peuvent pas les lire. Les dictionnaires pour francophones n'indiquent pas la prononciation des mots français ne leur genre (masculin/féminin) pourtant indispensables aux lecteurs japonophones. Il manque aussi certaines informations importantes telles que les compteurs (on ne compte pas de la même manière les objets : une voiture = ichi dai, un chien = ippiki, etc) ou les niveaux de langue.

En conclusion, pour un usage personnel, il est possible de trouver des dictionnaires imprimés (ou leur version électronique) d'assez bonne qualité à condition de savoir lire les kanji; mais lorsque l'on cherche un dictionnaire gratuit ou une ressource réutilisable dans d'autres outils, il n'y a bien souvent pas d'autre choix que d'utiliser un dictionnaire anglais-japonais, ce qui, on le sait, ne peut que multiplier les erreurs de compréhension et de traduction.

Toutefois, les projets JMdict et WadokuJiten montrent qu'il est possible de mener à bien des projets de construction collaborative de dictionnaires en ligne.

À ce stade, nous pouvons redéfinir notre projet de cette manière : récupérer des dictionnaires imprimés français-japonais de qualité et libres de droits grâce à un processus de lecture optique puis les rendre disponibles en ligne pour que les utilisateurs puissent corriger les erreurs restantes et actualiser les données.

3 Description de la ressource à construire

3.1 Historique du projet

En 2001, déjà confrontés au même problème de manque de ressources lexicales bilingue français-japonais, nous avons lancé le projet Papillon (Mangeot et al., 2004) de construction d'une base lexicale multilingue à structure pivot. Ce projet nous a permis d'avancer principalement dans deux directions : la collecte de données lexicales existantes (plus d'un million d'entrées en 11 langues) et les aspects théoriques avec la définition d'une macrostructure pivot et d'une microstructure utilisant les fonctions lexico-sémantiques de la lexicographie explicative et combinatoire.

En 2003, le lancement du projet GDEF (Mangeot & Chalvin, 2006) de création d'un dictionnaire bilingue estonien-français sera l'occasion d'avancer sur la partie logicielle avec la mise en œuvre de Jibiki, plate-forme générique de gestion de ressources lexicales en ligne.

En 2010, le projet MotÀMot (Mangeot, 2014) de dictionnaire français-khmer sera l'occasion de travailler sur la la définition de niveaux de qualité pour les données et les contributeurs ainsi que la gestion des liens bilingues et pivot.

En 2012, le projet DiLAF Dictionnaire langue africaine-français (Enguehard & Mangeot, 2014) nécessitera de définir une méthodologie précise de récupération de données issues d'éditeurs de texte standard (Word) ;

3.2 Microstructure des articles

3.2.1 Microstructure générale

De manière générale, nos articles seront basés sur une association d'un lexème et d'une catégorie grammaticale. Les articles n'auront donc pas de bloc grammatical. La structuration des articles suivra celle définie par la norme Lexical Markup Framework (LMF) (Francopoulo et al., 2009) : chaque article contient un bloc forme qui regroupe les informations liées à la forme : vedette, prononciation, catégorie grammaticale puis une suite de blocs sens. Chaque bloc sens décrit un sens de mot. Il contient également les traductions dans l'autre langue ainsi qu'une liste d'exemples. Chaque exemple est traduit dans l'autre langue.

Les données proviendront principalement de la récupération de ressources existantes. Dans un premier temps, la structure des articles suivra donc celle de la ressource d'origine. Ensuite, à moyen terme, notre but est de tendre vers une microstructure plus riche fondée sur la lexicographie explicative et combinatoire (Mel'čuk et al., 1995), partie de la théorie sens-texte (TST). Pour chaque sens de mot (formellement lexie), ajouter une formule sémantique qui peut être vue comme une définition formelle - dans le cas d'une lexie, prédicative, la formule décrit le prédicat et ses arguments et on trouve aussi le régime qui décrit la réalisation syntaxique des arguments - , puis une liste de fonctions lexico-sémantiques - il y a 56 fonctions de base applicable à toute langue et pouvant se combiner entre elles -, suivie d'une liste d'exemples et enfin d'expressions idiomatiques.

3.2.2 Articles japonais

Les articles de la plupart des dictionnaires japonais sont basés sur le lexème. Il n'y a pas d'articles homographes. Nous reprendrons ce découpage.

Concernant le mot-vedette, chaque kanji ayant plusieurs prononciations possibles, il est nécessaire d'indiquer leur prononciation. Celle-ci est généralement indiquée en utilisant le syllabaire hiragana. Si l'on se contente des kanji et du hiragana, les lecteurs débutant en japonais ne pourront pas lire facilement les articles. Il faut donc utiliser également une transcription latine du japonais, appelée romaji. Il existe plusieurs méthodes de romaji officielles : la plus ancienne et la plus utilisée est la méthode Hepburn, introduite par le missionnaire américain James Hepburn en 1887. La méthode Kunrei quant a elle a été introduite par le ministère japonais de l'Éducation et a fait l'objet de la norme ISO 3602:1989. Elle se fonde sur la phonologie japonaise. Son principal avantage est qu'elle illustre mieux la grammaire, alors que le plus gros problème du Hepburn est qu'il change le radical des verbes, ce qui ne reflète pas la morphologie sous-jacente du japonais. Toutefois, les locuteurs non natifs préfèrent la méthode Hepburn qui donne une meilleure indication de la prononciation anglaise. Dans notre dictionnaire, nous souhaitons ajouter le romaji pour aider les locuteurs débutants non natifs. Nous choisissons donc la méthode Hepburn.

Il n'est malheureusement pas possible de générer automatiquement le romaji uniquement à partir de la prononciation hiragana. En effet, la lettre « う » peut être transcrite de deux manières différentes selon si elle prolonge une voyelle « o » ou non. L'association des deux voyelles « o » et « u » peut donc s'écrire de deux manières différentes en romaji selon si le « u » est au début d'un morphème (un kanji) ou non. Par exemple, le mot 東京(Tokyo) s'écrit en higanana « とうきょう » et en romaji « tōkyō » (et non « toukyou »), alors que le mot « 子牛 » (veau) s'écrit en hiragana « こうし » et en romaji « koushi » (et non « kōshi »). Nous choisissons donc de représenter chaque vedette avec trois

parties.

- une première partie « vedette-japonais » indique le mot japonais tel qu'il se apparait dans les textes. Cela peut être un mot en kanji uniquement (pour les noms), une combinaison de kanji et hiragana (pour les verbes et adjectifs notamment), un mot en hiragana seulement (adverbes), un mot en katakana (mot d'origine étrangère) ou toute combinaison des trois systèmes d'écriture (kanji, hiragana et katakana) ;
- une deuxième partie « vedette-hiragana » indique systématiquement la prononciation du mot en hiragana (même si la vedette-japonais est déjà un mot en hiragana) ;
- une troisième partie « vedette-romaji » indique la transcription de la vedette-japonais en romaji Hepburn moderne. Cette partie peut elle-même contenir deux versions : une pour l'affichage qui peut contenir des espaces, des tirets ou des points et une pour la recherche qui ne contient que des lettres.

Dans le reste de l'article, pour chaque segment de texte japonais, nous ajoutons la prononciation en hiragana au dessus du texte (appelé furigana) ainsi qu'une transcription en romaji Hepburn.

3.2.3 Articles français

La lexicographie traditionnelle distingue deux vocables comme homographes se elles n'ont aucun lien sémantique clair entre eux (Polguère, 2008). Mais en pratique, il arrive fréquemment que les dictionnaires avec la même langue source suivent une division entre vocables homographes différente. Nous pensons que cette distinction et en fait arbitraire et qu'il n'y a pas non plus à notre connaissance de critère objectif permettant d'évaluer un lien sémantique entre deux mots pouvant être mis en œuvre de manière automatique avec des outils de traitement automatique des langues. Nous choisissons donc de ne pas distinguer de vocables homographes s'ils ont la même catégorie grammaticale. La combinaison d'un lexème et d'une catégorie grammaticale constitue donc un seul article.

Pour les utilisateurs non francophones du dictionnaires, nous ajouterons la prononciation du mot-vedette. Nous indiquerons également le genre (masculin/féminin) de chaque nom représentant une traduction française dans un article japonais.

3.3 Niveaux de qualité

Chaque partie d’information de chaque article se verra attribuer un niveau de qualité . Les niveaux s’échelonnent de 1 étoile pour un brouillon (données récupérées dont la qualité n’est pas connue) à 5 étoiles, qualité certifiée par un expert (par exemple, un lien de traduction validé par un traducteur assermenté).

De la même manière, les contributeurs se verront assigner un niveau de compétence (1 à 5 étoiles également). 1 étoile étant le niveau d’un débutant inconnu dans la communauté et 5 étoiles étant le niveau d’un expert reconnu.

Ensuite, lorsqu’un contributeur de niveau 3 révise un article de niveau 2, l’article monte automatiquement au niveau 3. De même, si le travail d’un contributeur est systématiquement validé sans corrections par d’autres contributeurs de niveau supérieur, celui-ci peut passer automatiquement au niveau supérieur au bout d’un certain seuil (par exemple 10 contributions).

Pour aller plus loin, nous envisageons d'analyser le travail des contributeurs. Si une personne contribue massivement par exemple sur un domaine particulier, le système pourra de manière automatique lui envoyer régulièrement des propositions de contribution dans son domaine.

4 Récupération du dictionnaire Cesselin

4.1 Présentation du dictionnaire

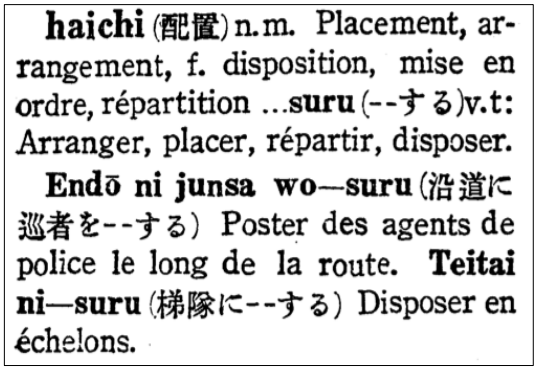

haichi (配置) n.m. Placement, arrangement, f. disposition, mise en ordre, répartition ...**suru** (--する) v.t: Arranger, placer, répartir, disposer.
Endō ni junsa wo—suru (沿道に巡者を--する) Poster des agents de police le long de la route. **Teitai ni—suru** (梯隊に--する) Disposer en échelons.

Figure 7: Scan de l'article « haichi » du dictionnaire Cesselin

Le « Cesselin » (Cesselin, 1944) est un dictionnaire japonais→français élaboré par Gustave Cesselin, missionnaire apostolique décédé en 1944 et ayant effectué toute sa carrière au Japon. Le dictionnaire contient 2 365 pages et plus de 82 600 articles. Le mot-vedettes est noté en romaji puis en japonais (kanji ou kana). Il est suivi par une catégorie grammaticale en français puis une liste de traductions en français. Ensuite, vient une liste d'expressions contenant le mot-vedette en romaji puis en japonais (kana et kanji). Chaque expression est traduite en français. L'article se termine par une liste d'exemples, chacun étant noté en romaji, en japonais puis traduit en français (voir figure 7).

4.2 Négociation des droits d'auteur

Pour s'assurer qu'un ouvrage est libre de droits, il faut vérifier que tous les auteurs soient décédés depuis une durée déterminée. Dans le cas d'un dictionnaire qui peut être rédigé par un nombre conséquent d'auteurs, cette vérification peut être ardue. D'autre part, la durée varie d'un pays à l'autre. Au Japon, elle est actuellement de 50 ans. En France, elle est de 70 ans, durée partagée par un grand nombre de pays.

Gustave Cesselin est le seul auteur déclaré de son dictionnaire. Il est décédé en 1944. Au Japon comme en France, le dictionnaire « Cesselin » est donc libre de droits.

Si le dictionnaire visé n'est pas libre de droits, une autre solution consiste à négocier directement avec les détenteurs des droits. Dans le cas des missionnaires qui sont les auteurs les plus nombreux jusque dans les années 1950, les détenteurs sont les congrégations.

Dans le cas du dictionnaire français-japonais « Raguet-Martin » (Raguet & Martin, 1953), Jean-Marie Martin est décédé en 1975. Il faut donc attendre 10 ans au Japon et encore 30 ans en France pour que le dictionnaire soit enfin libre de droits. Nous avons donc contacté les Missions Étrangères

de Paris, congrégation détenant les droits de cet ouvrage. Un accord a été conclu pour l'utilisation des données sur notre site Web.

4.3 Scan du dictionnaire

Nous avons essayé plusieurs techniques de scan :

- scanner manuel avec scanner à plat. Il en résulte une bande noire au milieu des deux pages qui cache certains caractères se situant de plus en début de ligne, donc potentiellement des mots-vedette. Le résultat n'est pas utilisable. D'autre part, la manœuvre est fastidieuse car elle nécessite de soulever le dictionnaire et tourner à la main chaque page.
- scanner avec caméra sur le dessus. La procédure est rapide (compter 3 h) car il n'est pas nécessaire de bouger le livre. Par contre, il reste une petite courbure au milieu des deux pages.
- scanner avec caméra au dessus et sur le côté. La caméra située sur le côté est prévue pour calculer l'épaisseur du livre et ainsi redresser la courbure au milieu des deux pages. Nous n'avons pas pu expérimenter ce type de machine.
- découpe du livre et scan automatique. Cette technique donne la meilleure qualité de scan puisqu'il n'y a pas de courbure ou de zone noire. De plus, elle est très rapide car automatique. Par contre, elle impose le sacrifice d'un exemplaire car la reliure est découpée pour pouvoir scanner les pages individuellement.

4.4 Lecture optique

Une fois le scan du dictionnaire effectué dans les meilleures conditions possibles, il faut trouver un logiciel de lecture optique qui soit capable de :

- reconnaitre plusieurs langues en même temps ;
- permettre d'entraîner la lecture optique ;
- reconnaitre les kanji.

Nous avons passé au banc d'essai une dizaine de logiciels. Seul Abbyy s'est distingué. Son inconvénient majeur est qu'il ne permet pas d'entraîner la lecture optique sur les idéogrammes. Nous avons alors effectué 2 passes. Le premier traitement a été effectué en choisissant uniquement le français comme langue et en entraînant la lecture optique sur une page. Les parties de texte en français ont été correctement reconnues avec très peu d'erreurs. Les parties de texte en japonais (kanji ou kana) n'ont pas été reconnues du tout. Le deuxième traitement a été effectué en choisissant le français et le japonais comme langues, ce qui n'a pas permis d'entraîner la lecture. Les parties de texte en français ont été reconnues avec un taux d'erreur plus important que lors du premier traitement et les parties de texte en japonais ont été reconnues avec un taux d'erreur standard. Il faut compter environ 12 h pour un traitement sur tout le dictionnaire.

Les résultats des traitement sont ensuite exportés au format OpenXML (.docx) ou OpenDocumentFormat (.odt). Ces documents sont en fait des archives zip. Elles sont décompressées puis les fichiers XML contenant le texte de chaque page du dictionnaire sont

extraits. Les fichiers résultat des deux traitements sont ensuite fusionnés en utilisant un algorithme de calcul de distance de chaînes type Levenshtein.

4.5 Détection des mots-vedette

La partie la plus importante de la récupération d'un dictionnaire se situe dans la détection des mots-vedette. Ils délimitent les articles et leur servent également de clé d'accès. Cette partie se compose de trois phases :

1. Récupération des vedettes en en-tête de chaque page. Cette opération est très importante car elle servira à borner l'ordre alphabétique des vedettes se trouvant dans la page. Dans les archives .docx ou .odt, les en-tête des pages sont séparés du texte principal. Nous avons programmé un script permettant d'extraire automatiquement ces en-têtes. Il a fallu ensuite vérifier les vedettes (est-ce du romaji et sont-elles classées dans l'ordre alphabétique) et les corriger le cas échéant. Dans l'exemple de la figure 8 pour la page 323 du Cesselin, les en-têtes sont « haichai » et « haigeki ».

2. Comparaison de chaque début de ligne par ordre alphabétique strict. La comparaison s'effectue d'abord avec les en-têtes : en-tête 1 < mot < en-tête 2 puis entre les vedettes détectées vedette 1 < vedette 2 < vedette 3. Dans la figure 8, les vedettes suivantes sont extraites : haichai, haichi, haichi, haichüritsu, haidan

3. Comparaison approximative de chaque début de ligne. Pour tenir compte des erreurs de lecture optique, un deuxième passage est réalisé. Cette fois-ci la comparaison s'affectue en introduisant une marge d'erreur grâce à un calcul de distances de chaînes type Levenshtein. Dans la figure 8, les vedettes « iichi » et « haicbutsu » sont extraites et corrigées automatiquement pour devenir « haichi » et « haichutsu ».

Il reste un problème de sur-détection. En effet, certaines expressions composées faisant partie d'un article commencent en début de ligne et reprennent le mot-vedette de l'article. Exemple, page 1038 du Cesselin, « kuwadate iru » a été détecté comme vedette alors que c'est une expression de l'article « kuwadateru » et « Kuwashiku monoshiberu » a été détecté comme vedette alors que c'est une expression de l'article « kuwashii ».

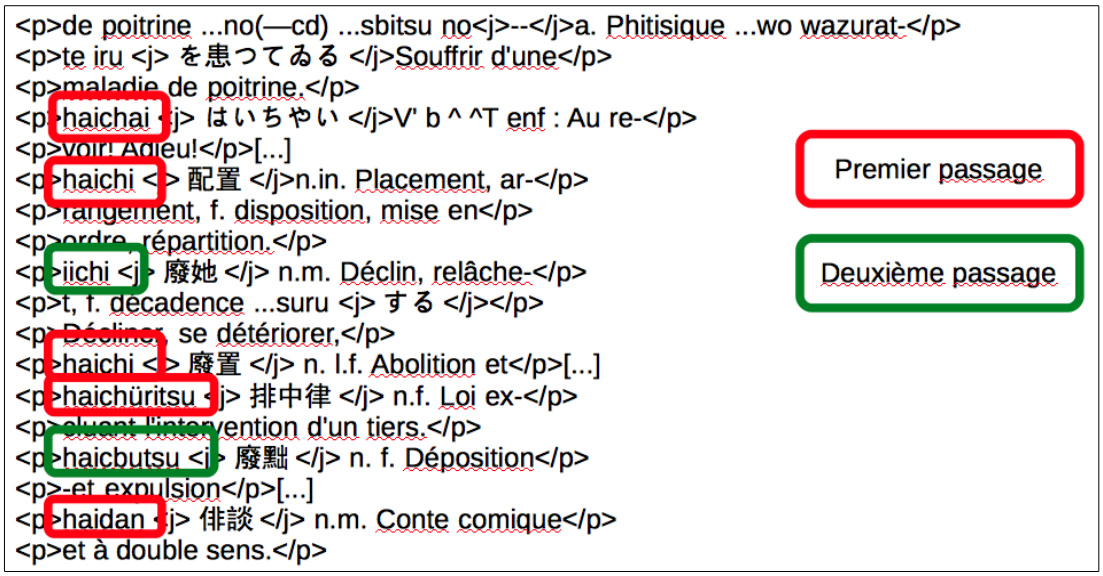


Figure 8 : Détection des vedettes pour la page 323 du Cesselin

4.6 Correction après lecture optique

Après la lecture optique, il est possible de corriger automatiquement certains caractères du texte selon la langue utilisée.

4.6.1 Français

Les corrections sur le français se concentrent sur l'utilisation des diacritiques. Exemples : Â + ' ' ⇒ À ; Etre ⇒ Être ; ç[^àaou] ⇒ c ; etc.

4.6.2 Romaji

Le romaji n'utilise que le macron comme diacritique : ā,ī,ū,ē,ō. Les autres diacritiques sont donc automatiquement convertis : à ⇒ a; äâ ⇒ ā, etc.

Certaines suites de caractères n'existent pas en romaji. Elles sont donc converties également : lt + [aiueo] ⇒ h + [aiueo]; rn + [aiueo] ⇒ m +[aiueo], etc.

4.6.3 Japonais

Les erreurs sur le japonais apparaissent principalement en début ou en fin de chaîne. Lorsqu'il y a un changement de langue entre le français et le japonais, le logiciel détecte ce changement en retard. De ce fait, les premiers et les derniers caractères d'un segment japonais sont parfois remplacés par des caractères ASCII car le logiciel n'a pas détecté le changement de langue. Certains schémas se répètent souvent. Nous indiquons ci-dessous trois exemples de remplacement. La balise ouvrante <j> indique le début et la balise fermante </j> la fin du segment japonais :

- 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 Listes prioritaires de corrections

Afin de prioriser le travail de correction restant à la charge des contributeurs lorsque la ressource sera mise en ligne, nous avons calculé des listes prioritaires à partir de listes de fréquences de mots dans un corpus japonais. La liste de fréquences de mots utilisée (JapFreqList_5109_Novels) est tirée d'un corpus de 5 109 romans japonais de littérature moderne. Elle a été construire pour le projet cbJisho[4] et peut être téléchargée en suivant ce lien[5]. Elle contient 188 218 entrées. La fréquence la plus haute étant la virgule japonaise « 、 » avec 26 244 137 occurrences.

Pour prioriser le travail restant sur les vedettes avec kanji non détectés, nous avons généré une nouvelle liste de fréquence basée sur le hiragana à partir de la liste JapFreqList. Le dictionnaire JMdict est d'abord consulté pour obtenir la liste des hiragana possibles pour chaque mot japonais de la liste JapFreqList. Ensuite, on effectue la somme des fréquences obtenues pour chaque hiragana puis on trie la liste obtenue. La dernière étape consiste à comparer cette liste à celle des kanji non détectés dans le Cesselin.

4.7 Modernisation du japonais

Depuis l'époque de la rédaction du dictionnaire, la langue japonaise a subi de nombreux changements notamment dans son écriture. Le but du projet n'est pas de rendre accessible le plus fidèlement possible le Cesselin mais de construire un dictionnaire reflétant l'usage actuel des langues. C'est pourquoi nous avons décidé de moderniser le japonais automatiquement lorsque cela était possible.

4.7.1 Romaji

Le romaji utilisé à l'époque pour la rédaction du Cesselin, basé sur la transcription Hepburn, était appelé romajikwa. Il utilisait les lettres « kwa » et « gwa » pour retranscrire certaines syllabes qui sont maintenant simplifiées en « ka » et « ga ». Exemple : kwaikwan ⇒ kaikan. Le « m » était également utilisé pour retranscrire la lettre hiragana « ん » devant les consonnes « m,b,p » car elle se prononce effectivement « m ». Actuellement, le « n » est utilisé partout. Exemple : jimbōchō ⇒ jinbōchō ; gumma ⇒ gunma.

La transcription de la consonne « n » est ambigüe en japonais car cela peut être la lettre « ん » finale de syllabe ou un début de syllabe (na, ni, nu, ne, no). Il est donc d'usage de noter dans le romaji lorsque c'est une finale de syllabe. Dans le Cesselin, un tiret est ajouté après le « n ». Mais le tiret est aussi parfois utilisé pour séparer des syllabes. D'autre part, le romaji actuel utilise d'autres caractères. Le Hepburn moderne utilise une apostrophe '. Les francophones utilisent quant à eux un point. Nous avons donc remplacé le tiret marquant un « ん » par un point. Exemple : ran-i ⇒ ran.i

4.7.2 Simplification des kanji

À l'époque de la rédaction du dictionnaire, il n'existait pas de table de codage pour les kanji. Le premier codage japonais, le JIS (Japanese Industrial Standard) 208 est apparu en 1978. Il contient environ 7000 caractères. Ce codage a été largement utilisé lors de l'informatisation du japonais. Si bien que, les kanji qui n'étaient pas inclus dans ce codage ont petit à petit disparu. Pour corriger la situation d'autres codage sont apparus comme le JIS 212 en 1990 qui spécifie 6 067 caractères de

4 http://subs2srs.sourceforge.net/cbJisho/help.html
5 http://forum.koohii.com/viewtopic.php?pid=132030#p132030

plus puis ensuite le JIS 213 qui en spécifie 11 233 en tout. Mais le mal était fait et la plupart des mots de la langue japonaise actuels n'utilise que les kanji du standard JIS 208. Nous avons donc remplacé tous les kanji qui n'étaient pas inclus dans le JIS 208 par leur variante JIS 208. De la même manière, nous avons remplacé les variantes JIS 208 par celles de grade le plus bas.

- Remplacements JIS 213 → JIS 208 : 狀⇒状,歩⇒歩,黃⇒黄,緣⇒縁,晩⇒晩,黑⇒黒
- Remplacements JIS 212 → JIS 208 : 啞⇒唖,掻⇒掻,頬⇒頬,𠮟⇒叱,伙⇒火
- Remplacement de variantes JIS 208 : 阪 8⇒坂 3,弍⇒一 1,埜 10⇒野 2,亰⇒京 2,區⇒区 3

Certains idéogrammes n'étaient inclus dans aucun JIS. Il font partie de l'ensemble des caractères chinois Han. Une partie de ces caractères est effectivement utilisée dans le dictionnaire original Cesselin mais l'autre partie provient en fait d'erreurs de reconnaissance de caractères. Pour ces deux séries, il a fallu trouver un équivalent JIS 208 à la main.

- Exemples de caractères Han inclus dans le Cesselin : 絕⇒絶, 說⇒説,靑⇒青.
- Exemples de caractères Han provenant d'erreurs : 內⇒内,戶⇒戸,出⇒出.

4.7.3 Japonais

Le japonais a également évolué depuis 1950, notamment les terminaisons verbales. L'évolution était déjà indiquée dans le romaji mais le hiragana utilisé pour les terminaisons verbales en japonais gardait trace des anciennes prononciations. Exemples : la terminaison « ふ » (fu) se prononce « u » ; la terminaison « へ る » (heru) se prononce « eru ». Cela engendrait de ce fait un problème de correspondance entre le romaji et le kana. Nous avons donc modifié le hiragana pour qu'il corresponde au romaji et ainsi à la prononciation moderne : kokitsukau 扱使ふ ⇒ 扱使う ; kikikaeru 切替へる ⇒ 切替える.

Certains hiragana ont également été remplacés : ゐ⇒い (i);ゑ⇒え(e). Exemple : 光ってゐる ⇒ 光っている.

Le sokuon, lettre marquant les consonnes géminées marquées avec un macron dans le romaji Hepburn, est noté habituellement avec un petit tsu « っ ». Dans le Cesselin, le sokuon est marqué avec un tsu de taille standard « つ ». Lors des vérifications ultérieures, toutes les variantes avec et sans petit tsu sont donc générées pour éviter la sur-détection de problèmes.

4.8 Marquage des informations

Une fois toutes les corrections effectuées, il est temps de passer au balisage de chaque catégorie d'information. Au départ, seuls les segments contenant du japonais (kanji ou kana) sont balisés. Le résultat des étapes précédentes a également balisé la vedette en romaji et en japonais. Il va falloir baliser tous les autres segments.

Nous avons d'abord listé toutes les abréviations utilisées ce qui nous a permis de baliser l'étymologie, les catégories grammaticales, le domaine et le registre de langue. Nous avons également inclus dans la liste les erreurs fréquentes provenant de la lecture optique. Exemple : « n.in. » au lieu de « n.m. » sur la figure 9.

Ensuite, les segments en japonais sont tous précédés de segments en romaji. Ceux-ci commencent

soit par « ... » soit par une majuscule. Ils sont également suivis par des segments en français qui se terminent par un point. Ces règles nous ont permis de baliser les exemples.

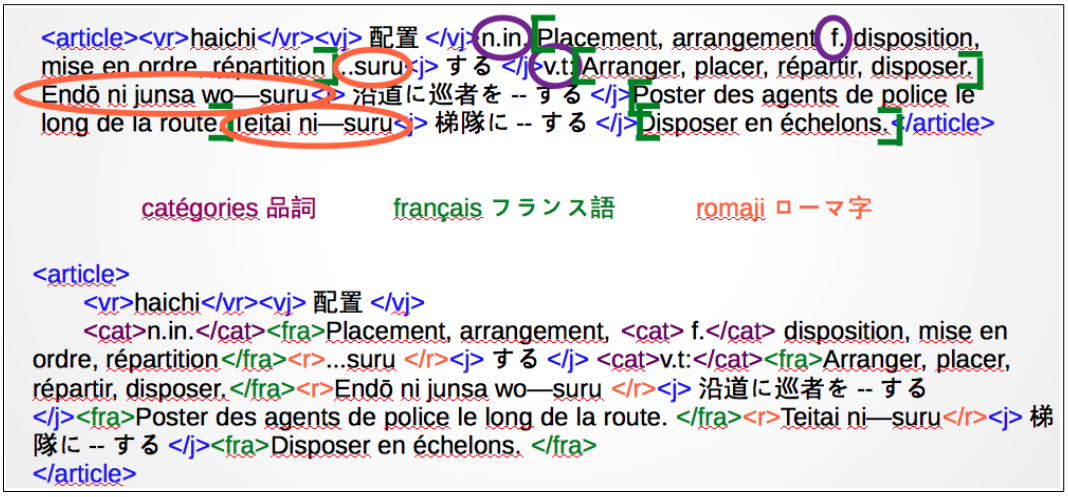


Figure 9 : marquage des informations pour l'article « 配置 » (haichi)

Il restait finalement à baliser le segment de la traduction de la vedette en français.

Note : à cause des erreurs de lecture optique, il est compliqué d'utiliser un outil de détection de langue pour faire la différence entre le romaji et le français.

4.9 Structuration des articles

Lors de cette étape, il s'agit principalement de structurer les articles en respectant la partie normative du standard LMF (Francopoulo et al., 2009) pour permettre un export automatique vers la partie informative (syntaxe LMF). La partie normative spécifie la structuration des différents blocs mais ne donne aucune contrainte sur la manière de les représenter (éléments et attributs XML). Il est donc possible qu'une ressource respecte la norme LMF tout en gardant ses propres balises. La partie informative donne un exemple de syntaxe LMF mais nous considérons qu'elle est peu pratique à utiliser (Enguehard & Mangeot, 2013). Nous choisissons donc d'utiliser nos propres balises. Cette structuration consiste principalement à regrouper les informations de forme dans un bloc « <forme> » et les informations de sens dans un bloc « <sens> ». Les exemples du Cesselin n'étant pas rattachés à un sens en particulier, nous n'avons pas séparé les exemples en blocs sens différents.

Figure 10 : structuration de l'article « 配置 » (haichi)

4.10 Complétion et vérification des vedettes

Une fois le dictionnaire structuré, il s'agit à cette étape de détecter un maximum d'erreurs potentielles afin de les marquer pour qu'elles puissent être facilement corrigées en ligne par la suite par les utilisateurs du dictionnaire et d'ajouter des informations complémentaires comme le furigana pour les segments japonais.

Une première détection consiste à attester la présence des vedettes dans d'autres dictionnaires. À cette fin, nous avons profité de la présence du dictionnaire japonais-anglais « super daijirin » (Matsumura, 2006) dans MacOs pour programmer un outil permettant d'automatiser la consultation de ce dictionnaire. Nous avons également téléchargé et installé le JMdict pour compléter la vérification. L'algorithme est le suivant :

- Génération du hiragana à partir de la vedette en romaji et ajout dans le bloc forme de l'article ;

- Vérification de la présence de la vedette en kanji dans le « super daijirin ».

- Sinon, vérification dans le JMdict.

- Si la vedette en kanji est attestée dans un de ces deux dictionnaires, alors vérifier si les hiragana correspondent.

- S'ils ne correspondent pas, convertir en romaji le hiragana trouvé dans un des dictionnaires de vérification puis calculer la distance approximative entre les deux romaji à l'aide de l'algorithme de Levenshtein.

- Si celle-ci correspond, modifier le romaji et le hiragana de la vedette par ceux trouvés dans le dictionnaire de vérification

- Si celle-ci ne correspond pas, marquer le problème pour correction plus tard.

- Si la vedette en kanji n'est pas attestée, utiliser le hiragana pour chercher une proposition alternative dans les dictionnaires de vérification

- Si une proposition alternative est trouvée, la rajouter dans l'article comme alternative possible.

- Si aucune alternative n'est trouvée, utiliser l'API de transcription de Google pour proposer une alternative.

- Si la vedette en kanji est vide suite à une erreur de lecture optique, vérifier la présence du hiragana dans les deux dictionnaires de vérification.

- S'il n'y a qu'une seule vedette en kanji dans les deux dictionnaires de vérification, alors remplacer la vedette en kanji vide par celle trouvée dans les dictionnaires de vérification.

Les erreurs de lecture optique son fréquentes sur les lettres avec macron en romaji. Celui-ci est utilisé pour générer le hiragana. Par conséquent, pour les comparaisons du hiragana avec les dictionnaires de vérification, toutes les variantes avec et sans voyelle longue sont générées (ā/a, ī/i, ū/u, ē/e, o/ō). Cela évite une sur-détection de problèmes.

Le numéro de page dans la version imprimée originale du Cesselin est ajouté dans chaque article. Cela permettra ultérieurement d'afficher un lien vers fichier PDF de la page scannée pour corriger les erreurs de la lecture optique.

Au final, environ une vedette sur deux a été attestée dans un autre dictionnaire et environ 10 % des vedettes en kanji restent vides.

4.11 Détection des erreurs en français

La détection d'erreurs potentielles en français s'est effectuée en utilisant l'analyseur morphologique tree tagger[6]. Chaque phrase française est envoyée à l'analyseur. Si un mot inconnu est détecté, celui-ci est marqué pour correction ultérieure. Exemple : “L'un des huit enfers glacés du boaddhisme.“. « boaddhisme » n'a pas été reconnu par l'analyseur. Dans ce cas, il s'agit d'une erreur de lecture optique. Le mot correct est « bouddhisme ».

Cette étape pourrait être affinée d'avantage. En effet, tous les mots latins n'ont pas été reconnus par l'analyseur. D'autre part, il doit être possible de corriger automatiquement certaines erreurs telles que celle mentionnée dans l'exemple.

4.12 Détection des erreurs en japonais

Les analyseurs du japonais ne peuvent pas détecter d'erreurs d'orthographe telles qu'on en trouve en français puisque cette langue est sans séparateurs et tous les kanji ont une signification. Pour la détection des erreurs potentielles, nous avons comparé la transcription en romaji avec la version japonaise. Nous avons d'abord converti le romaji en hiragana puis nous avons utilisé Mecab[7] comme analyseur morphologique du japonais pour générer le furigana des kanji inclus dans le japonais. Nous avons ensuite comparé le hiragana issu de la conversion du romaji avec le furigana issu de l'analyseur. La comparaison du hiragana s'effectue en générant toutes les variantes comme indiqué en 4.10. Dès qu'une différence est trouvée, celle-ci est marquée pour correction ultérieure.

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

Les variantes avec et sans petit tsu sont générées lors de la comparaison (voir 4.7.3).

6 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
7 https://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

Lors de cette étape, le furigana a été ajouté aux exemples japonais grâce à la sortie de l'analyseur.

Cette étape pourrait également être améliorée. En effet, les deux exemples indiqués précédemment ne sont en fait pas des erreurs mais proviennent du fait qu'il peut y avoir plusieurs furigana possibles pour le même kanji. Il faudrait donc utiliser la sortie d'un analyseur donnant toutes les solutions possibles.

5 Récupération des liens de Wikipedia

La couverture du dictionnaire Cesselin est déjà conséquente : plus de 82 000 articles. Cependant, sa sortie date de 1939, soit avant la deuxième guerre mondiale qui a eu pour conséquence l'occupation du Japon par les forces armées américaines de 1945 à 1952. Depuis cette époque, beaucoup de mots anglais ont été intégrés au japonais après avoir été transcrits en katakana. Un dictionnaire de japonais moderne ne peut les ignorer. Nous avons donc réutilisé deux ressources gratuitement disponibles pour compléter le premier ensemble de données provenant du Cesselin. Il s'agit de Wikipedia et du JMdict.

Nous avions prévu au départ d'utiliser Wiktionary, mais nous avons constaté que la plupart des traductions japonaises présentes dans le Wiktionary français provenaient en fait des liens de traduction entre pages Wikipedia. Nous avons préféré utiliser directement la ressource d'origine. Nous avons donc repris les liens des pages du Wikipedia japonais vers les pages françaises et anglaises.

5.1 Processus de récupération

Wikipedia propose pour chaque langue de télécharger toutes ses données sous forme d'export de la base de données au format SQL[8]. Nous avons téléchargé les informations sur chaque page telles que l'identifiant et le titre de la page (jawiki-latest-page.sql.gz) ainsi que liens vers les pages dans d'autres langues (jawiki-latest-langlinks.sql.gz). Nous avons ensuite importé ces données dans une base de données MySQL puis nous avons créé une nouvelle table « traduction » contenant au départ l'identifiant de chaque page, son titre, le titre de la page reliée pour les pages en français et en anglais puis la langue de la page reliée (français ou anglais).

5.2 Extraction du hiragana

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 Wikipedia[9], 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.

8 https://dumps.wikimedia.org/jawiki/latest/
9 https://ja.wikipedia.org/w/api.php

5.3 Génération du romaji

Comme expliqué en 3.2.2, le romaji ne peut être généré automatiquement à partir du hiragana s'il contient les suites de lettres おう ou うう. D'autre part, alors que le japonais est une langue sans séparateurs, il est d'usage d'ajouter des espaces entre les mots ainsi que des majuscules au début des noms propres dans les transcriptions en romaji, ce qui facilite grandement la lecture. Si le romaji est généré à partir du hiragana (ou même des kanji), il n'y aura donc pas d'espaces. Nous avons observé que les pages qui sont des traductions dans d'autres langues de noms propres japonais indiquent souvent dans la première phrase la graphie japonaise du nom propre ainsi que le romaji.

- Exemple : 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.

Comme pour l'extraction du hiragana, nous utilisons l'API de Wikipedia pour extraire le romaji des traductions des pages en japonais.

Dans le cas où nous n'avons pas trouvé de romaji dans les pages de traduction, nous allons chercher les lectures de chaque kanji qui compose le mot japonais dans le dictionnaire Kanjidic[10]. Plusieurs types de lecture sont associées : onyomi (d'origine chinoise), kunyomi (d'origine japonaise), okurigana (avec des hiragana supplémentaires servant de terminaison), nanomi (pour les noms propres).

Il faut ensuite prendre en compte les phénomènes compositionnels en japonais : rendaku (harmonisation consonantique, ex : か→が), sokuon (consonne géminée : つ→っ), renjō (doublement du 'n' final, ex : んあ→な).

Ensuite, la génération du romaji est effectuée sur la base de la conversion du hiragana.

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

5.4 Sélection des vedettes à importer

À ce stade, il s'agit de sélectionner les entrées que nous allons importer dans le dictionnaire Cesselin. Il ne s'agit pas d'importer toutes les entrées pour « faire du chiffre » mais uniquement celles dont les vedettes sont attestées dans d'autres ressources. Dans un premier temps, nous allons sélectionner les traductions françaises. Les traductions anglaises seront importées après la récupération du dictionnaire JMdict. L'algorithme de sélection est le suivant :

- Vérifier si le titre de la page est dans le dictionnaire Daijirin ;

    - Si oui, vérifier si le titre de la page (vedette en japonais) est déjà dans le Cesselin ;

        - Si oui, ajouter le lien vers les entrées de Wikipedia dans l'article du Cesselin,

        - Si non, vérifier si le romaji est dans le Cesselin ;

            - Si oui, vérifier si les vedettes en japonais des entrées du Cesselin ont été vérifiées à l'étape 4.10 ;

10 http://www.csse.monash.edu.au/~jwb/kanjidic.html

- Si toutes les vedettes sont vérifiées alors importer l'entrée ;

- Si le romaji n'est pas dans le Cesselin alors importer l'entrée.

Cet algorithme ne garantit pas une sélection sans problèmes. En effet, à cause des erreurs potentielles de lecture optique, il est possible qu'une vedette soit importée alors qu'elle est déjà dans le Cesselin si le kanji est vide et si le romaji comporte une erreur. Dans ce cas, une fois que le romaji et le kanji de l'article original du Cesselin seront corrigés, une recherche de doublons vedette japonais + vedette hiragana permettra de supprimer les articles dupliqués. Un autre problème possible consiste à ne pas importer un article car le romaji est déjà dans le Cesselin mais pas le kanji. Dans ce cas, une fois que toutes les vedettes japonais avec kanji correspondant à ce romaji seront vérifiées, il sera possible de réimporter l'article manquant.

Au total, 23 456 articles ont été générés à partir des liens de Wikipedia et importés dans le Cesselin. Parmi ceux-ci, 20 825 articles sont traduits en français et 2 631 articles traduits en anglais.

6 Récupération du dictionnaire JMdict

JMdict est un dictionnaire japonais → anglais (voir figure 4). Afin que notre dictionnaire ait dès le départ une large couverture, nous avons décidé d'importer les entrées du JMdict même si la plupart des traductions sont en anglais et pas en français. D'une part, l'anglais étant proche du français et beaucoup étudié, la plupart des francophones peuvent comprendre l'anglais écrit et d'autre part, du fait du manque de ressources lexicales français-japonais, beaucoup d'apprenants du japonais utilisent de toutes façons des dictionnaires anglais-japonais. Par contre, même si nous avions la possibilité technique de croiser le JMdict avec un dictionnaire français-anglais, nous avons préféré laisser la traduction en anglais pour éviter les contresens. Les contributeurs pourront eux-mêmes proposer en ligne une traduction en français lorsqu'ils consulteront ces entrées.

Chaque article du JMdict contient, pour la vedette, une liste de mots en kanji (k_ele) et une liste de mots en hiragana ou katakana (r_ele). Chaque mot en kana est parfois suivi d'une liste de restrictions pour indiquer à quel mot en kanji il correspond. D'autre part, si, dans les textes japonais, le mot ne s'écrit qu'en hiragana ou katakana, la liste de mots en kanji peut être vide. Par conséquent, la lecture des vedettes (lecture en hiragana + kanji) n'est pas immédiate et demande de calculer les bonnes correspondances. Il n'y a pas non plus de transcription en romaji.

Pour toutes les vedettes, nous avons donc clarifié les informations et généré une liste de vedettes conforme à la structure que nous avons choisie pour le Cesselin (voir 3.2.2) : si la liste k_ele est vide, on y copie les éléments de la liste r_ele. Si un élément r_ele est écrit en katakana, il est convertit en hiragana. Ensuite, chaque élément de la liste k_ele est une vedette japonaise à laquelle est associée une vedette en hiragana. S'il y a plusieurs lectures pour le même mot en kanji, la vedette japonaise est dupliquée pour chaque hiragana correspondant. Puis, l'algorithme de génération du romaji de la section 5.3 est repris pour générer le romaji à partir du hiragana et des kanji.

Ensuite, les kanjis des vedettes en japonais ont été simplifiés comme dans la section 4.7.2 puis les vedettes en doublon supprimés.

Enfin, l'algorithme de la section 5.4 a été réutilisé pour sélectionner les articles à importer. Lors de

l'import des articles dans le Cesselin, l'identifiant de l'article dans le JMdict est gardé pour référence ultérieure.

Une liste d'articles à traduire en français en priorité a été également générée à partir de la liste de fréquences JapFreqList_5109_Novels (voir 4.6.4).

Au total, 47 810 articles du JMdict dont 2 521 traduits en français et 45 289 traduits en anglais ont été importés dans le Cesselin.

7 Mise en ligne sur la plate-forme Jibiki

7.1 Description du site

Le site Web du projet[11] est construit autour de la plate-forme Jibiki (Mangeot, 2006) de gestion de ressources lexicales hétérogènes en ligne. Celle-ci est programmée à l'aide de Enhydra, un serveur d'objets java basé sur une architecture 3/tiers. La base de données utilisée pour la couche données est Postgres. Cette plate-forme, développée et constamment améliorée depuis 2001 est disponible en licence libre (LGPL) sur la forge du laboratoire LIG[12].

Outre les fonctions de gestion des ressources (consultation et édition des dictionnaires), nous avons ajouté au site :

- un blogue en page d'accueil ;
- une interface de consultation de corpus bilingue aligné (concordancier) ;
- un module d'aide à la lecture ;
- une page permettant de télécharger l'intégralité les données présentes sur le site et de téléverser de nouvelles données (dictionnaires et corpus).

Le site est disponible en trois langues : français, anglais et japonais.

7.1.1 Page d'accueil : blogue

La page d'accueil est composée d'une interface de consultation simple (voir 7.1.4), d'un texte présentant brièvement le projet, de listes d'articles à traiter en priorité et d'un blogue programmé à l'aide de WordPress.

Deux listes d'articles sont affichées : celle des articles du dictionnaire Cesselin dont les kanjis du mot-vedette n'ont pas été reconnus lors de la lecture optique (voir 4.6.4) et celle des articles issus du JMdict dont la traduction est en anglais à traduire en français (voir 6). Pour chaque liste, les dix mots les plus fréquents sont affichés. Un bouton de calcul des listes est disponible afin de recalculer les listes si des corrections ont été effectuées.

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.

11 http://jibiki.fr
12 https://jibiki.ligforge.imag.fr

7.1.2 Corpus bilingues alignés

Le module de consultation des corpus bilingues alignés français-japonais est programmé en perl à l'aide de la plate-forme IMS Open Corpus Workbench[13] et de son outil de consultation Corpus Query Processor (CQP). Celui-ci permet d'utiliser un langage d'expressions régulières pour établir des requêtes complexes. Par exemple, la requête "intére(t|ss)(e|é) (r|e)?s?" obtiendra les mots intérêt, intérêts, intéressé, intéresser, intéressée, intéressées. Il est également possible de combiner les niveaux d'annotations. Par exemple, la requête [(lemme="sous.+") & (cat="V.*")] obtiendra toutes les formes conjuguées des verbes commençant par le préfixe « sous ». Ce module est dérivé d'une première version utilisée dans le projet GDEF pour un corpus estonien-français[14].

Les données proviennent d'une part du projet OPUS[15] pour les logiciels (KDE, OpenOffice), le Coran et les sous-titres de films (OpenSubtitles) et d'autres part de corpus que nous avons constitué nous-mêmes avec des textes trouvés sur le Web (Le Monde Diplomatique, la Bible, une convention fiscale franco-japonaise et la déclaration universelle des droits de l'Homme).

Ces données, alignées au niveau du paragraphe, ont été ensuite étiquetées pour le français en utilisant Tree tagger (voir 4.11) et pour le japonais en utilisant Mecab (voir 4.12). Le furigana a été ajouté au texte japonais, ce qui permet d'effectuer des recherches au niveau de la lecture des kanji. Par exemple, une recherche du mot « はな » (haha) obtiendra les mots 花 (fleur), 鼻 (nez), etc.

La figure 11 montre le résultat de la recherche du mot japonais « 配置 » (haichi) dans le corpus du Monde Diplomatique. Le corpus étant parallèle, il est également possible de rechercher des mots français.

À l'heure actuelle, la totalité des corpus atteint environ 6 millions de mots dont :

- Journaux : Le Monde Diplomatique (288 745 mots) ;
- Textes légaux : Convention fiscale franco-japonaise (17 443 mots) et Déclaration universelle des droits de l'Homme (2 208 mots) ;
- Logiciels : KDE 4 (1,179 millions de mots) et OpenOffice 3 (569 903 mots) ;
- Textes religieux : la Bible (904 914 mots) et le Coran (Tanzil) (192 905 mots) ;
- Sous-titres de films (OpenSubtitles) (4,714 millions de mots).

13 http://cwb.sourceforge.net
14 http://corpus.estfra.ee
15 http://opus.lingfil.uu.se

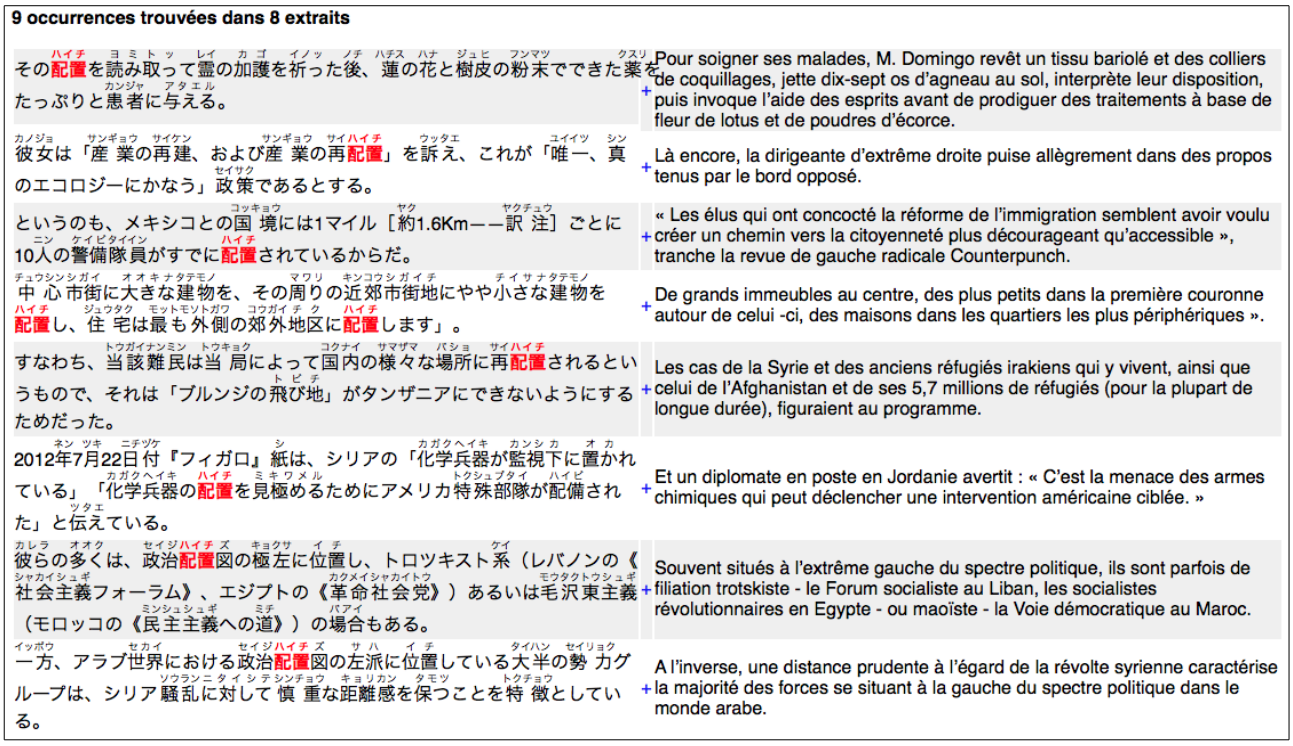


Figure 11 : Résultat de la recherche du mot « 配置 » (haichi) dans le corpus du Monde Diplomatique

7.1.3 Module de lecture active

Le module de lecture active propose une aide à la lecture pour un utilisateur qui connait une langue mais ne la maîtrise pas. Celui-ci entre un texte dans cette langue puis d'afficher une traduction mot-à-mot de ce texte. Dans notre cas, l'utilisateur francophone peut entrer un texte japonais et le japonophone un texte français. Ensuite, le module ajoute la prononciation des mots ou le furigana dans le cas du japonais ainsi que les traductions de chaque mot. Pour ne pas gêner la lecture, celles-ci ne sont affichées que lorsque l'utilisateur pointe un mot avec sa souris. Voir figure 12 pour un exemple d'affichage sur un texte japonais avec la souris pointée sur le mot « 時代 » (jidai).

Le module envoie d'abord le texte à un analyseur morphologique : Tree tagger pour le français (voir 4.11) et Mecab pour le japonais (voir 4.12) puis récupère les lemmes pour chaque mot. Il utilise ensuite l'interface de programmation (API) REST[16] de la plate-forme Jibiki[17] pour consulter différentes ressources. Pour les textes japonais, le furigana est obtenu avec l'analyseur morphologique et les traductions avec le dictionnaire Cesselin. Pour les textes français, dans l'attente de la construction d'un dictionnaire français→japonais, les traductions sont pour l'instant proposées en anglais. La prononciation et la traduction sont obtenues en utilisant le dictionnaire FeM.

16 https://fr.wikipedia.org/wiki/Representational_State_Transfer
17 http://jibiki.fr/jibiki/Api.po

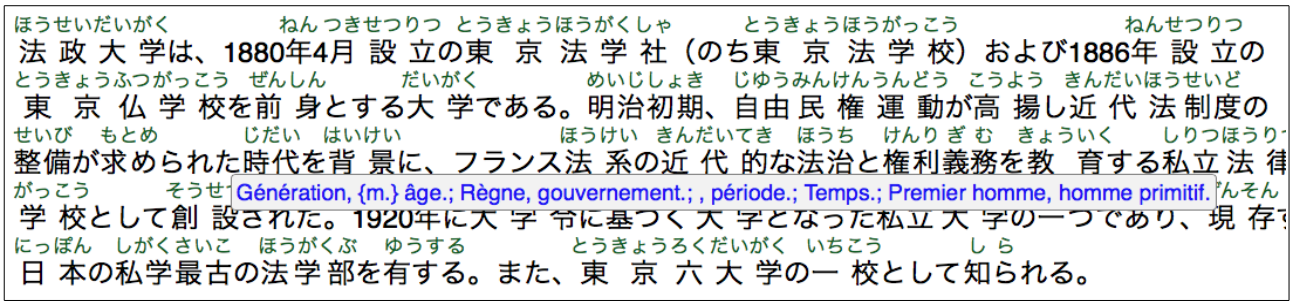


Figure 12 : Affichage du module pour un texte japonais avec la souris sur le mot « 時代 » (jidai)

7.1.4 Consultation simple

La consultation simple permet de rechercher un mot japonais tapé en romaji, kana ou kanji.

L'interface de consultation simple affiche une vue évoluée du dictionnaire imprimé : la partie gauche affiche les vedettes du voisinage immédiat du mot recherché classées par ordre alphabétique. Elle est équipée d'un ascenseur infini permettant de parcourir toutes les vedettes du dictionnaire selon l'ordre alphabétique. Lorsqu'on clique sur une vedette de la partie gauche, l'article complet s'affiche dans la partie droite.

Un menu est affiché en haut à droite de chaque article. Celui-ci comprend des liens vers le formulaire d'édition, l'historique des modifications, la vue source XML de l'article, la page originale scannée du dictionnaire au format PDF et le résultat de la recherche de la vedette dans le corpus bilingue.

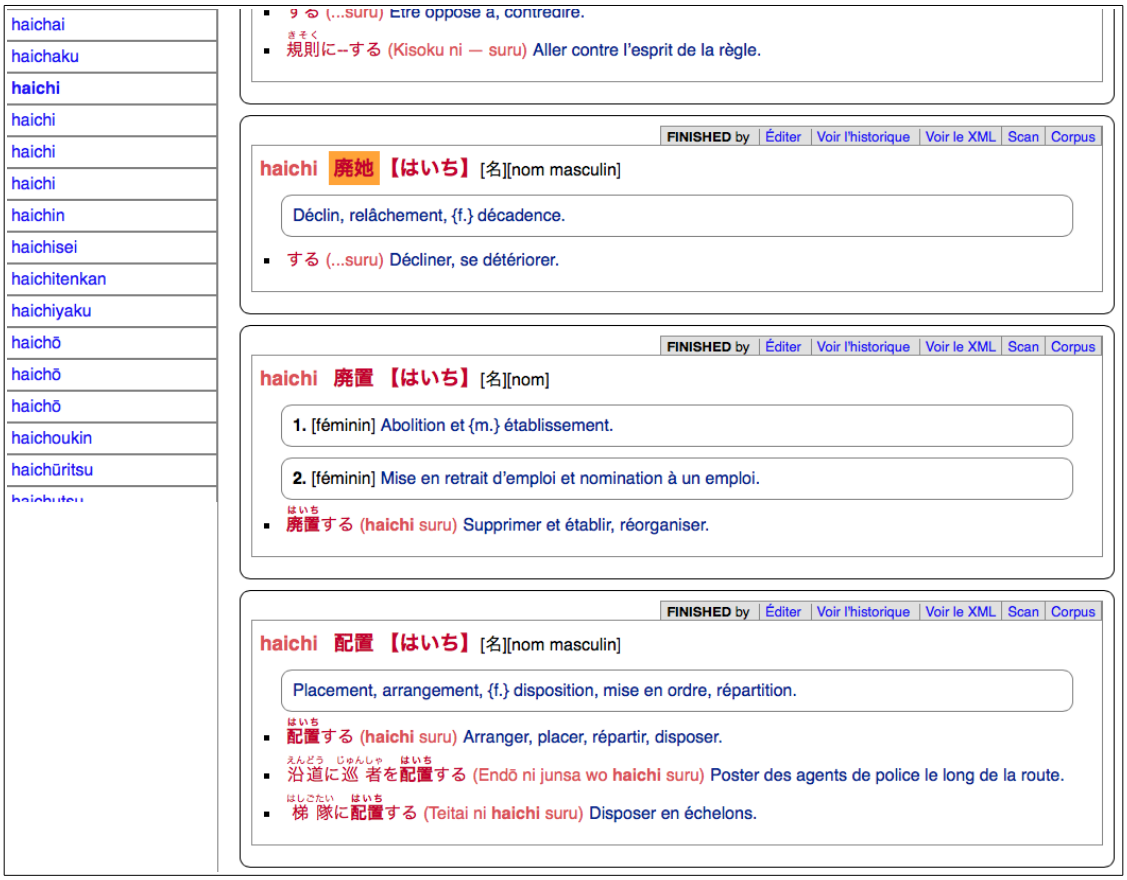


Figure 13 : résultat de la recherche du mot « haichi » via l'interface de consultation simple

Les erreurs ou anomalies détectées plus haut sont affichées avec un fond de couleur spéciale. Les vedettes non attestées et la non correspondance du romaji et du japonais sont en fond orange comme la vedette « 廃妣 » (haichi) de la figure 13. Les erreurs de français sont en fond jaune (voir figure 14). Les traductions anglaises provenant du JMdict ou de Wikipedia non traduites en français sont en fond vert.

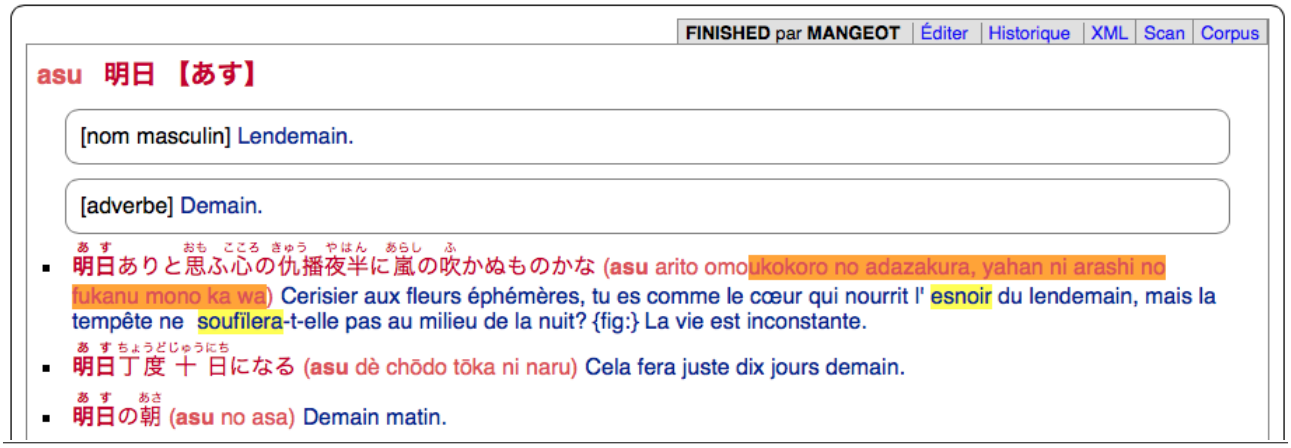


Figure 14 : Article « 明日 » (asu) avec des erreurs en français (fond jaune) et en japonais (fond orange).

7.1.5 Consultation avancée

L'interface de consultation avancée permet d'effectuer des recherches multi-critères sur toutes les ressources lexicales installées sur la plate-forme. Les critères sont applicables directement sur les données (vedette, prononciation, catégorie grammaticale, traductions, exemples, etc.) mais également sur les méta-données (auteur, statut, identifiant de l'article, identifiant de la contribution, etc.). Ceux-ci peuvent être combinés dans une seule recherche (voir figure 15). Les résultats de recherche sont affichés par liste alphabétique sur la partie gauche de la fenêtre. Si le nombre de résultat est supérieur à 100, une requête AJAX est effectuée pour aller chercher sur le serveur les résultats suivants par ordre alphabétique, comme pour l'interface de consultation simple.

Figure 15 : Interface de consultation avancée avec combinaison de critères

7.2 Édition en ligne

Pour éditer un article en ligne, l'utilisateur doit être enregistré au préalable sur la plate-forme Jibiki.

7.2.1 Édition rapide

Lors de la consultation d'un article, il est possible d'effectuer de petites modifications directement à sur l'article affiché. Pour cela, il suffit de double-cliquer sur le segment à modifier et celui-ci se transforme en champ de texte avec un bouton « ok » sur la droite pour valider la saisie (voir figure 16).

Cet éditeur est programmé à l'aide de la technologie AJAX[18]. Il utilise l'interface de programmation (API) REST[16] de la plate-forme Jibiki[17] pour dialoguer avec le serveur. Lors de la validation de la saisie (clic sur le bouton « ok »), la nouvelle chaîne de caractères est envoyée au serveur avec le pointeur XPath[19] du segment édité et l'identifiant unique de l'article.

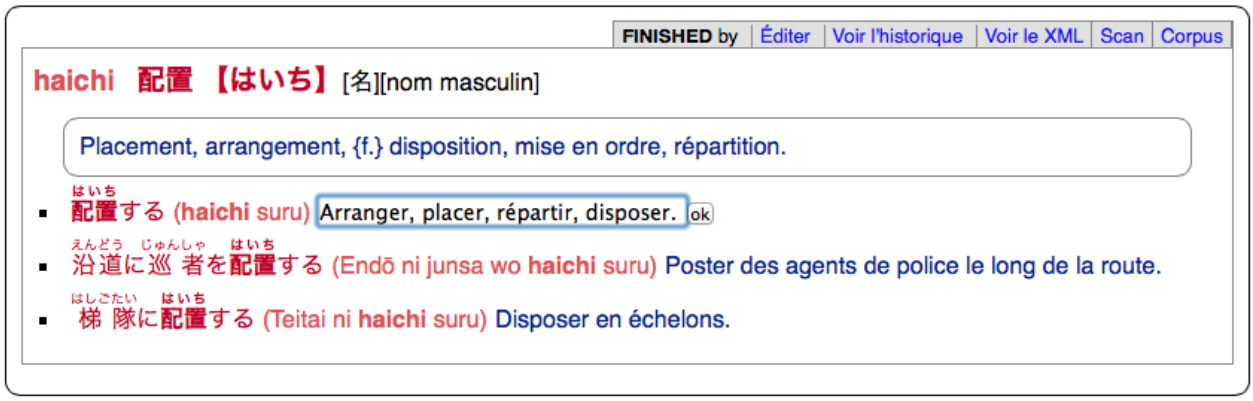


Figure 16 : Édition rapide de la traduction française du premier exemple de l'article «配置 » (haichi)

7.2.2 Formulaire d'édition complète

L'interface d'édition complète est générée automatiquement à partir de la structure des articles notées sous forme de schéma XML[20] (Mangeot & Thevenin, 2004). Celle-ci se compose d'un formulaire HTML avec interacteurs classiques pour les différents types de données (champ texte pour du texte libre, menu déroulant pour des listes fermées de valeurs, cases à cocher pour des booléens, boutons radio pour des choix de valeur, etc.) ainsi que des interacteurs élaborés pour gérer les listes d'objets (par exemple, ajout ou suppression d'un exemple dans une liste d'exemples en appuyant sur les boutons “+” et “-” dans la figure 17).

18 https://fr.wikipedia.org/wiki/Ajax_(informatique)
19 http://www.w3.org/TR/xpath/
20 http://www.w3.org/XML/Schema

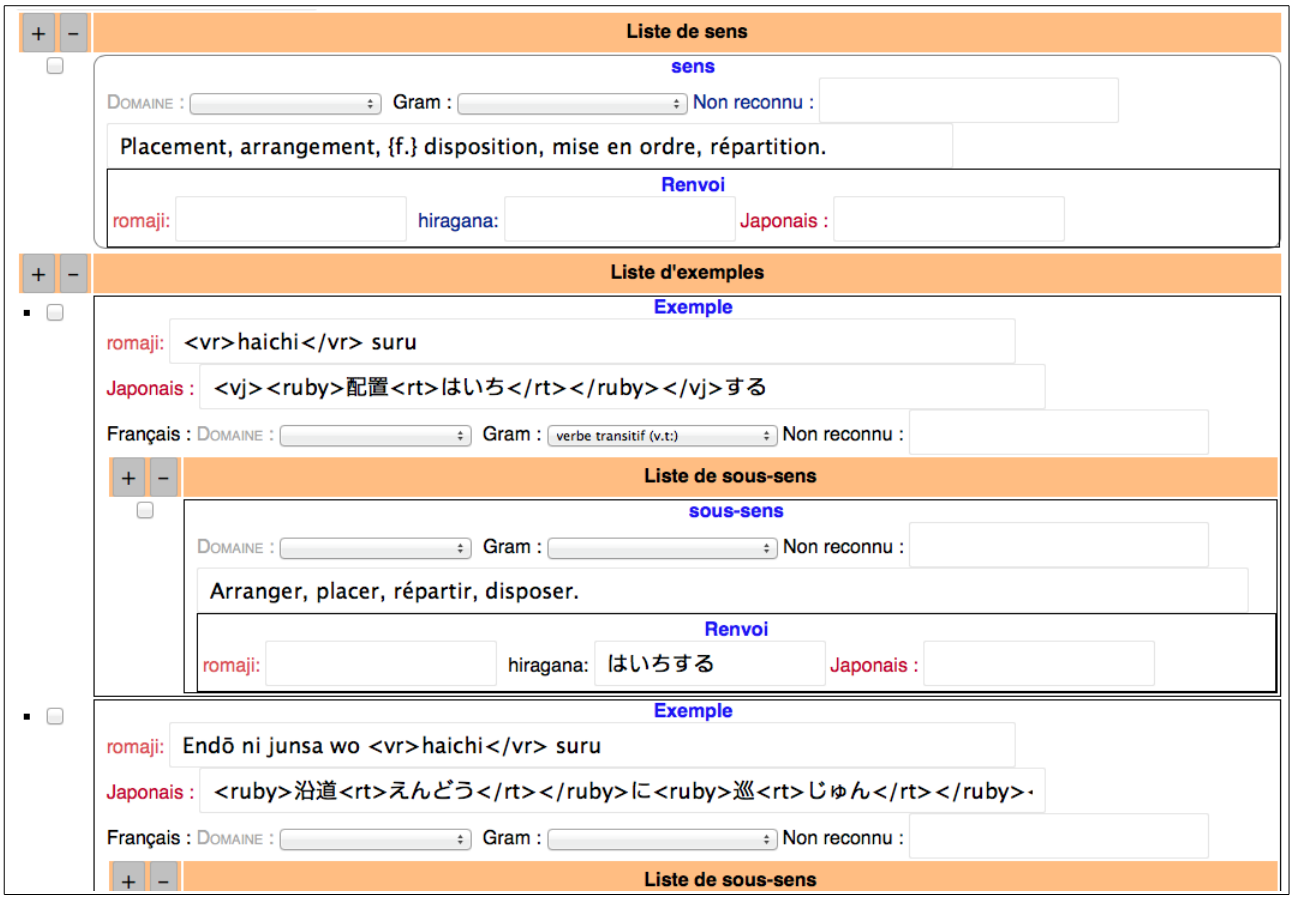


Figure 17 : Formulaire d'édition complète pour les exemples de l'article « 配置 » (haichi)

Pour accéder à l'édition complète, il suffit de cliquer sur le lien « Édition » du menu en haut à droite de chaque article. L'utilisateur peut ensuite éditer l'article concerné. À la fin de son travail, il prévisualise ses changements. Il peut ensuite soit sauvegarder temporairement ses modifications en leur affectant un statut « brouillon », soit annuler ses modifications, soit les enregistrer définitivement dans la base. Les versions précédentes sont gardées dans la base. Ce qui permet de revenir en arrière si des erreurs systématiques sont détectées chez un contributeur par exemple.

7.3 Statistiques

7.3.1 Nombre d'entrées

Le dictionnaire contient 153 897 articles au total. Parmi ceux-ci,

- 82 663 articles proviennent du dictionnaire Cesselin dont :
  - 10 243 articles (12,39 %) dont les mots-vedette en kanji sont « ?? » (qui n'ont pas été correctement reconnus par la lecture optique) ;
  - 40 259 articles (48,70 %) dont les mots-vedette n'ont pas encore été vérifiés (automatiquement ou manuellement) ;
- 47 721 articles proviennent du dictionnaire JMdict ;
- 23 512 articles proviennent des liens entre pages Wikipedia.

Sur les 153 897 articles, on en trouve 47 813 articles (31,07 %) dont les traductions sont en anglais (donc à traduire en français).

7.3.2 Nombre de contributions

Après 3 mois de fréquentation, au 25 octobre 2015, le site a enregistré 2 639 modifications d'articles dont :

- 86 mots-vedette en kanji ajoutés,
- 225 mots-vedette en kanji vérifiés,
- 132 traductions en français

7.3.3 Nombre de visites

3 mois après son ouverture le 22 juillet 2015, le site a enregistré 664 visites au 22 octobre 2015.

8 Conclusion

Nous avons montré dans cette article qu'il est possible de lancer un projet de construction collaborative de dictionnaires sur le Web en réutilisant des ressources libres de droits ce qui permet d'obtenir un dictionnaire utilisable immédiatement. Le site n'est ouvert que depuis 3 mois mais le nombre élevé de contribution permet déjà de montrer que l'expérience s'avère concluante.

La méthodologie décrite dans cet article pour récupérer un dictionnaire peut être réappliquée à toute ressource imprimée libre de droits (et il y en a beaucoup !).

La constitution de cette ressource constitue un point de départ pour des recherches futures.

Concernant la production de données, nous prévoyons de lancer un processus identique pour récupérer un dictionnaire français→japonais. Nous envisageons également d'enrichir la ressource actuelle en y ajoutant de nouvelles informations (compteurs, quantificateurs, fréquences d'apparition dans des corpus, etc.).

Concernant l'exploitation de données, l'obtention d'une ressource français→japonais permettra d'expérimenter la convergence vers une macrostructure pivot (Mangeot et al., 2004). Les exemples et leur traduction peuvent être réutilisés pour construire un corpus bilingue aligné qui peut servir par exemple à entraîner un système de traduction statistique type Moses.

9 Remerciements

Ce projet a été réalisé grâce au programme Hosei International Fund (HIF) qui nous a permis d'être accueilli à l'Université Hosei, Tokyo d'octobre 2014 à août 2015.

10 Bibliographie

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) *Analyze [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.

Griolet Pascal (2008) *Plus de « cent cinquante ans » d'histoire de l'enseignement du japonais.* Le japonais au XXIe 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 Hakusuisha 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, Genova, 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*, Hakusuisha, Tokyo, 1467 p.