Contributions et corrections dans le dictionnaire japonais-français jibiki.fr

Mathieu Mangeot & Mutsuko Tomokiyo
Laboratoire LIG, équipe GETALP
Bâtiment IMAG CS 40700
38058 GRENOBLE CEDEX 9, FRANCE
{mathieu.mangeot,mutsuko.tomokiyo}@imag.fr

Introduction

Bien que le français et le japonais soient considérés comme des langues bien dotées concernant les outils et les ressources linguistiques, le couple franco-japonais est considéré comme une paire de langues peu dotées en ce qui concerne sa disponibilité sur le Web. En effet, il existe peu de ressources lexicales électroniques bilingues de qualité et qui sont à la fois gratuites et libres de droit. Les corpus bilingues franco-japonais et systèmes de traduction automatique sont logiquement également rares.

Heureusement, il existe des dictionnaires imprimés français-japonais de bonne qualité et suffisamment anciens pour être libres de droits tels que le dictionnaire japonais-français de Gustave Cesselin (Cesselin, 1940). Nous avons réutilisé cette ressources pour construire un dictionnaire de bonne qualité et large couverture disponible sur le Web (Mangeot, 2016). Le dictionnaire Cesselin a d'abord été scanné, puis lu optiquement et analysé pour détecter les mots-clés et les articles. Ensuite, plusieurs corrections d'erreurs ont été effectuées sur le français et le japonais. Afin de mettre à jour ces données dont le vocabulaire est parfois ancien, nous avons réutilisé des ressources électroniques existantes telles que Wikipedia et le dictionnaire électronique japonais-anglais Jmdict (Breen, 2002). La ressource résultante a été ensuite mise en ligne sur le Web[1] pour consultation et correction par des contributeurs volontaires. Les données sont disponibles dans le domaine public.

Cette méthodologie pourrait être appliquée à d'autres couples linguistiques dans une situation similaire avec de bons dictionnaires imprimés mais peu de ressources électroniques.

Dans cet article, après deux ans et demi de vie du projet, nous effectuons un bilan des contributions manuelles et automatiques.

1 Corrections automatiques

Les corrections automatiques s'effectuent à l'aide de scripts programmés en Perl qui utilisent l'interface de programmation (API) de la plateforme jibiki (voir http://jibiki.fr/jibiki/Api.po).

1.1 Transcription du kanji 大

En japonais, le o long est noté ō ou ô en romaji et おう [ou] en hiragana sauf lorsqu'il transcrit le kanji 大. Dans ce cas, il est transcrit おお [oo]. Par exemple, le nom 王族 (famille royale) sera transcrit ōzoku en romaji et おうぞく [ouzoku] en hiragana ; l'adjectif 大き (grand) sera transcrit ōki en romaji et おおき [ooki] en hiragana.

Avant la mise en ligne du dictionnaire, lorsque pour chaque mot-vedette, nous avons généré le hiragana à partir du romaji présent dans le dictionnaire, nous n'avions pas pensé à intégrer cette exception. Il a donc fallu la corriger à l'aide d'un script automatique une fois le dictionnaire mis en ligne. 571 mots-vedettes ont été corrigés.

1 http://jibiki.fr/

1.2 Okurigana

Les okurigana (送り仮名) sont des suffixes qui suivent la racine en kanji (漢字, caractères chinois) des verbes et des adjectifs. Le système d'écriture d'okurigana a été changé à plusieurs reprises, et il a été standardisé par le gouvernement japonais en 1973. Comme le Cesselin (publié en 1939) a pris un ancien système d'okurigana, nous l'avons l'a remplacé de manière automatique par le système en cours en utilisant le dictionnaire Super Daijirin inclus dans MacOs. Par exemple :

- 考へる →考える (kangaeru, penser),
- 空騒 (karasawagi, du bruit pour rien) →空騒ぎ

Pour chaque mot-vedette du dictionnaire Cesselin, le dictionnaire Super Daijirin est consulté avec le furigana (prononciation) du mot-vedette (からさわぎ dans le deuxième exemple). Ensuite, les hiragana sont supprimés du mot-vedette du Cesselin ainsi que des mots-vedette trouvés dans le Super Daijirin (空騒ぎ→ 空騒 dans le deuxième exemple). Si des mots-vedette correspondent (空騒=空騒), alors le mot-vedette du Cesselin est remplacé (空騒→空騒ぎ) et l'ancienne version est gardée dans l'article avec une mention (vieux okurigana). Cela permet de retrouver l'article 空騒ぎ même si la recherche est effectuée avec l'ancienne version (空騒).

Nous avons corrigé de cette manière environ 6500 articles.

1.3 Mots-vedette

Les données provenant de lecture optique, il est nécessaire de vérifier chaque mot-vedette pour s'assurer qu'il n'y a pas eu d'erreur de reconnaissance de caractères.
Plusieurs cas ont été distingués :

- Le mot-vedette et le furigana apparaissent dans un autre dictionnaire : celui-ci est marqué vérifié. Cela concerne 42 219 mots-vedette, soit 50 % du total.
- Le mot-vedette n'apparait pas dans d'autres dictionnaires mais il se prononce comme le furigana : celui-ci est marqué à vérifier. Cela concerne 12 922 mots-vedettes soit 16 % du total. Il reste encore 7 765 mots-vedette à vérifier.
- Le mot-vedette n'apparait pas dans d'autres dictionnaires et il ne se prononce pas comme le furigana. Il doit être corrigé à la main. Cela concerne 17 233 mots-vedette, soit 21 % du total. Il reste encore 12 872 mots-vedette à corriger.
- Le mot-vedette est vide. Il faut le rajouter à la main. Cela a concerné 10 329 mots-vedette, soit 12 % du total des mots-vedette.

2 Corrections manuelles

2.1 Mots-vedette

Le remplissage de mots-vedette non-détectés a été effectué de manière automatique avec 65 % de succès. Le reste a été effectué manuellement par des contributeurs volontaires, qui sont étudiants français ou japonais, japonisants, chercheurs, linguistes, etc. Sur la page d'accueil du projet, une liste des 20 mots-vedette non détectés classés par fréquence de leur prononciation est affichée pour motiver les contributeurs potentiels. La liste est mise à jour automatiquement tous les soirs. Elle peut également être mise à jour manuellement par un contributeur s'il a corrigé tous les mots-vedettes affichés. Dans la plupart de cas, ce sont des noms anciens de plantes ou d'animaux, des noms d'outils désuets, de vieux kanjis, etc. Le travail prend environ 30 minutes pour 20 mots-vedettes. Par exemple : 薊 (azami, chardon), 掾所(yoridokoro, fondement).
Le remplissage se fait d'abord par un copier-coller depuis la page PDF scannée du Cesselin vers l'interface d'édition en ligne de Jibiki. En effet, sur MacOs, le lecteur de PDF inclut un outil de reconnaissance optique de caractères. Lorsque le copier-coller échoue à cause de Kanjis trop vieux ou d'une mauvaise reconnaissance de caractères, il faut consulter des dictionnaires de kanji, qui

permettent de trouver et copier des kanjis par des composants (偏 (hen), 旁 (tsukuri)) ou le nombre de traits d'un kanji (kakusû, 画数)[2] lorsque l'on ne connaît ni leur prononciation ni le sens.
Cette opération a concerné 10 329 mots-vedettes et a duré 2 ans et demi. Elle a été effectuée e grande majorité par 3 contributeurs.

2.2 Compteurs

En japonais, il existe des lexèmes qui indiquent la classe de noms, lorsqu'ils apparaissent dans une expression quantitative. Ils dépendent du type de référents ou de l'observation de référents par celui qui parle. Nous avons annoté tous les lexèmes du Cesselin apparaissant comme classificateurs / quantificateurs dans nos listes (Tomokiyo, Boitet, 2016), (Tomokiyo, Boitet & Mangeot, 2017).
Par exemple :

- 家畜３０頭 (kachiku 30 tou, trente têtes de bétail)
- 一枚のＴシャツ (ichimai no ti-shatsu, un T-shirt)
- 山のような問題 (yama no youna mondai, un tas de problèmes)

Cela concerne 253 articles.

2.3 Classification et annotation des exemples

Le dictionnaire Cesselin contient des exemples divers pour aider à la compréhension de mots. Mais il manque l'indication d'usage tels que proverbes, expressions figées, quantificateurs pour chaque exemple. Nous avons donc décidé de classifier tous les exemples et annoter leur usage (Tomokiyo, Mangeot & Boitet, 2018 soumis à COLING 2018).
Par exemple, pour le verbe 飲む (nomu, boire) :
(a) un proverbe : 飲まぬ酒には酔わぬ[3](Nomanu sake ni wa yowanu, II n'y a point de fumée sans feu)
(b) une expression figée : 彼は妻君に飲まれている[4](Kare wa saikun ni nomareteiru, Sa femme le berne)
(c) un exemple d'usage : 水を飲む (Mizu wo nomu, Boire de l'eau)

Bibliographie

Ulrich Apel (2002) *WaDokuJT - A Japanese-German Dictionary Database.* Papillon 2002 Seminar, 16–18 July 2002, NII, Tokyo, Japan.
Jim W. Breen (2004) *JMDict: a Japanese-multilingual dictionary.* In: Coling 2004 workshop on multilingual linguistic resources, Geneva, Switzerland, pp. 71–78.
Gustave Cesselin (1940) *Dictionnaire japonais-français.* Maruzen, Tokyo, juillet 1940, 2340 p.
Jean-Marc Desperrier (2002) *Analyze [sic] of the results of a collaborative project for the creation of a Japanese-French dictionary*. In: Proceedings of Papillon 2002 Seminar, 16–18 July 2002, NII, Tokyo, Japan.
Mathieu Mangeot (2016) *Collaborative Construction of a Good Quality, Broad Coverage and Copyright Free Japanese-French Dictionary.* International Journal of Lexicography, Volume 31, Issue 1, 1 March 2018, Pages 78–112; doi: 10.1093/ijl/ecw035; 35 p.
Mathieu Mangeot (2015) *Construction collaborative d'un dictionnaire japonais-français de qualité, à large couverture et libre de droits.* Rapport interne LIG, 31 p.
Mutsuko Tomokiyo, Mathieu Mangeot-Nagata & Christian Boitet (2017) *Development of a classifiers/quantifiers dictionary towards French-Japanese MT.* MT Summit 2017, 18-22 September 2017, Nagoya, Japan.
Mutsuko Tomokiyo & Christian Boitet (2016) *Corpus and dictionary development for classifiers/quantifiers towards a French-Japanese machine translation.* 5th Workshop on Cognitive Aspects of the Lexicon CogAlex@COLING 2016, 12 December 2016, Osaka, Japan.

2 http://kanji.jitenon.jp/
http://kanjitisiki.com/
https://www.sanseido.biz/main/dictionary/hanrei/daijirin_v3.aspx
3 Une traduction mot-à-mot: On ne s'enivre pas de Sake quand on ne l'a pas bu.
4 Une traduction mot-à-mot : Il est avalé par sa femme.