Interlots:LexInnova/Ressources à récupérer

Mathieu Mangeot

Interlots:LexInnova/Ressources à récupérer

Auteur : Mathieu Mangeot
Date : 7 juin 2016

Introduction

Ce document recense les ressources lexicales libres de droits ou sous licence ouverte de type Creative Commons qu'il est possible et intéressant de réutiliser pour construire les dictionnaires de référence pour chaque langue du projet ainsi que la base lexicale multilingue à structure pivot.

Le but était, dans un premier temps, de construire effectivement les dictionnaires de référence. Il nous est apparu prématuré de lancer un tel chantier alors que la décision d'intégrer le prototype LexInnova à la plate-forme Claroline Connect de façon à ce qu'il soit disponible dans l'ENPA n'a pas encore été prise. Nous indiquerons cependant dans ce document où se procurer chaque ressource et quelle est la marche à suivre pour l'intégrer à notre base lexicale.

Wiktionary

Les dictionnaires du projet Wiktionary [1] sont disponible sous licence libre [2].

Le principal défaut du projet est que l'information n'est pas structurée de la même manière à l'intérieur d'un wiktionary d'une même langue, ni a fortiori entre wiktionary de différentes langues. Les choix lexicographiques sont faits de manière indépendante pour chaque langue.

Attention, les chiffres du projet wiktionary sont à prendre avec précautions. Par exemple, pour le français, il est indiqué 2,9 millions d'entrées mais parmi elles, la majorité sont des mots d'une langue étrangère traduits en français. Le nombre de vocables français est de 367 000 (voir plus bas).

DBnary et extraction avec un outil java

Pour remédier au problème de manque de structuration des données, Gilles Sérasset a lancé le projet DBnary [3]. Celui-ci consiste à extraire automatiquement les informations de divers wiktionary et à produire du contenu structuré sous forme de linked data (Linguistic Linked Open Data).

Les données disponibles actuellement concernent les langues suivantes : bulgare, néerlandais, anglais, finnois, français, allemand, grec, italien, japonais, polonais, portugais, russe, serbo-croate, espagnol, suédois et turc.

Voici le tableau du nombre de vocables disponibles pour chaque langue du projet Innovalangues :

langue	nombre de vocables
allemand	147 000
anglais	689 000
espagnol	85 000
français	367 000
italien	46 000
japonais	33 000

Attention, la taille varie beaucoup entre les différentes langues. Pour un dictionnaire complet, une taille de 200 000 articles est considérée comme nécessaire. Les wiktionary en italien, espagnol et japonais sont à ce titre de taille insuffisante. Il faudra trouver d'autres ressources.

D'autre part, la qualité varie également entre les wiktionary. Nous avons par exemple analysé de près les liens de traduction français-japonais tirés des wiktionary et il ressort que la majorité de ces liens proviennent en fait de

wikipedia. La plupart des entrées de wikipedia proviennent elles-mêmes d'un dictionnaire japonais-français dico-fj lui-même obtenu à partir de la traduction française des traductions anglaises du dictionnaire japonais-anglais JMdict.

Le contenu de la base peut ensuite être extrait de la base DBnary en utilisant des requêtes SPARQL. Il existe également un outil programmé en java par Andon Tchetchmediev dans le cadre de sa thèse permettant d'extraire des données de DBnary. Les sources de cet outil sont disponibles ici [4].

Pivax 2

Pivax 2 [5] est une base lexicale multilingue comprenant principalement des dictionnaires du projet Universal Network Language [6].

langue	nombre de vocables
anglais	45 471
espagnol	7 080
français	27 538
hindi	31255
malais	37 342
russe	28 475
UNL	82 804
vietnamien	6 585

Chinois

CEdict

CEdict et un dictionnaire anglais-chinois provenant d'un système de traduction automatique. Il contient 107 712 entrées anglais->chinois et 215 424 entrées chinois->anglais. Il est consultable sur le site du projet Papillon [7].

Français

TLFi light

Le Trésor de la Langue Française informatisé est disponible en ligne [8]. Une version XML libre de droits car expurgée de ses exemples (qui sont sous droit d'auteur) peut être utilisée. L'archive est disponible ici [9]. Le dictionnaire contient 54 280 articles.

Morphalou 3

Morphalou est un lexique de formes fléchies du français disponible en ligne : https://www.ortolang.fr/market/lexicons/morphalou [10]. La version 3 contient 159 271 lemmes. Le lexique Morphalou3 comprend également la trancription phonétique de 93 681 lemmes et 504 898 formes fléchies.

La version 3 de Morphalou a été obtenue par la fusion de quatre lexiques :

- Morphalou 2 (version de décembre 2013)
- DELA (version de décembre 2011)
- Dicollecte (version 4.3)
- LGLex et LGLexLefff (version 3.4)
- Lefff (version 2.1 avril 2006)

Glawi

http://redac.univ-tlse2.fr/lexiques/glawi.html

Japonais

Jibiki/Cesselin

Le dictionnaire Jibiki/Cesselin est un dictionnaire japonais->français disponible sous licence CC0 (domaine public). Il contient 154 000 entrées. Les données peuvent être téléchargée directement sur le site Web [11] du projet.

JMdict et Kanjidic

Le JMdict est un dictionnaire japonais-anglais sous licence libre contenant 173 128 entrées. Le Kanjidic est un dictionnaire de kanjis contenant 13 108 caractères. Ils peuvent être téléchargés sur le site du projet [12].

WaDokuJiTen

WaDokuJiTen est un dictionnaire japonais-allemand de plus de 228 000 entrées sous licence libre. Le site du projet est disponible ici [13].

Conclusion

Afin d'obtenir des ressources plus complètes que wiktionary, comme celles indiquées pour el français, il faut pour chaque langue du projet enquêter auprès de chercheurs en traitement automatique des langues (TAL) de la langue concernée.

Références

[1] http://wiktionary.org
[2] https://wikimediafoundation.org/wiki/Terms_of_Use/fr
[3] http://kaiko.getalp.org/about-dbnary/
[4] http://totoro.imag.fr/Lexinnova/Ressources/dbnary-api-eclipse-sample-project.tgz
[5] http://www.getalp.org/pivax/
[6] http://www.vai.dia.fi.upm.es/ing/projects/unl/index.htm
[7] http://www.papillon-dictionary.org/
[8] http://atilf.atilf.fr/tlf.htm
[9] http://totoro.imag.fr/Lexinnova/Ressources/TLF_light.tar.gz
[10] https://www.ortolang.fr/market/lexicons/morphalou
[11] http://jibiki.fr
[12] http://www.edrdg.org/jmdict/edict_doc.html
[13] http://wadoku.de

Sources et contributeurs de l'article

Interlots:LexInnova/Ressources à récupérer *Source*: http://wiki.innovalangues.net/index.php?oldid=17872 *Contributeurs*: Eggerse, Mangeot

Licence