Analyse et classification d'exemples illustratifs dans le dictionnaire "Cesselin" en utilisant Google Traduction et un dictionnaire UNL-UWs

Mutsuko Tomokiyo, Mathieu Mangeot-Nagata, Christian Boitet

LIG-GETALP, Bâtiment IMAG, 700 avenue Centrale F-38400 St. Martin d'Hères, France ;
mutsuko.tomokiyo@imag.fr, Mathieu.Mangeot@imag.fr, Christian.Boitet@imag.fr

Résumé : Savoir si un exemple dans un dictionnaire bilingue est un proverbe, une expression idiomatique, une collocation ou un simple exemple d'utilisation serait très utile aux utilisateurs, mais est rarement indiqué. Dans cet article, nous présentons une classification manuelle et une annotation d'exemples dans un dictionnaire bilingue japonais-français, le « Cesselin » (145000 entrées), dérivées de l'analyse manuelle de 500 entrées, et décrivons les étapes d'automatisation du processus d'annotation. Quant aux proverbes, nous utilisons simplement des dictionnaires de proverbes. Pour les autres cas, notre méthode consiste à comparer la traduction (automatique) en anglais d'un exemple original (japonais) et sa traduction (française) dans le dictionnaire. La comparaison est basée d'abord sur le nombre de lexèmes communs aux deux sorties (anglaises), et deuxièmement sur le nombre de lexèmes synonymes, qui peuvent être trouvés en utilisant un grand dictionnaire UNL-UW. Nous montrons notre procédure de classification en détail et son automatisation future.

Mots-clés : classification des exemples dans un dictionnaire bilingue, annotation manuelle des exemples, automatisation de la classification des exemples, Cesselin, UNL-UW dictionary, Google Traduction

1 Introduction

Nous présentons des travaux en cours sur un dictionnaire bilingue japonais-français, le "Cesselin" [1], qui a été publié en 1939 et 1957 au Japon. Notre objectif est de le rendre disponible en tant que dictionnaire en ligne pour les apprenants de japonais

1. 和仏大辞典 (Wafutsu daijiten, Grand dictionnaire japonais-français), Cesselin, Gustave Jean Baptiste Cesselin, 丸善 (Maruzen) et 三省堂 (Sanseidô), 1939, Japon.

ou de français, les chercheurs en études japonaises et les développeurs de systèmes de traduction automatique japonais-français (Mangeot, 2016).

En fait, il existe très peu de ressources lexicales électroniques bilingues de qualité, et à la fois gratuites et libres de droits [2]. Le Cesselin est libre de droits de l'auteur, et cela nous a permis de le numériser en gardant les informations langagières et culturelles qu'il contient.

Le dictionnaire Cesselin (Cesselin) inclut des connaissances sur la langue japonaise parlée et écrite [3] et de riches exemples illustratifs des mots-vedette, tels que des descriptions phonétiques aujourd'hui obsolètes, les anciens caractères chinois et l'ancienne façon d'écrire les Okurigana [4]. Il n'est cependant pas satisfaisant que les exemples soient donnés sans indication [5] concernant leur nature : sont-ils des proverbes, des exemples d'utilisation normale, des exemples de quantificateurs, ou d'autres types de collocations ? Leur introduction permettrait aux développeurs de systèmes de TA de traiter méthodiquement les ambiguïtés lexicales (Tomokiyo, Boitet & Mangeot, 2017). Cela aiderait aussi les apprenants ou les chercheurs en langues à pratiquer le japonais ou le français et à approfondir leur compréhension de la culture japonaise.

Après avoir analysé un ensemble de 500 exemples (sur 280 000 exemples estimés), nous avons défini une classification des exemples et un processus de classification. La prochaine étape envisagée est de l'automatiser pour économiser le plus de temps d'expert humain possible.

Dans la section 2, nous rapportons une expérience faite afin d'étudier quels types d'exemples sont inclus dans le Cesselin. Dans la section 3, nous proposons une classification, et dans la section 4, une procédure de classification d'un exemple donné. Après avoir expliqué le dictionnaire d'UWs dans la section 5, nous proposons une procédure de classification automatique et examinons quelques aspects techniques de son automatisation dans la section 6.

2 Études de cas utilisant le système Google Translation pour la classification

Pour classer tous les exemples dans le Cesselin en fonction de leurs usages linguistiques, nous avons traduit 500 exemples japonais et leurs traductions françaises en anglais en utilisant le système de traduction de Google ("GT" ci-dessous), et avons comparé les deux traductions.

2. Il existe quelques autres dictionnaires français-japonais en ligne : le Diko (3 000 mots d'entrée, Diko), le Dictionnaire français-japonais (15 000 mots d'entrée, Lexilogos), le Dictionnaire japonais (24 000 mots d'entrée, Assimil) et le Dictionnaire Glosbe. (mots d'entrée inconnus, Glosbe), etc. Le nombre de mots de ces dictionnaires peut changer car ils vont être édités et étendus de manière collaborative.

3. Le système d'écriture japonais a été fixé par le gouvernement japonais en 1986, où la façon d'écrire est conforme à la façon de prononcer.

4. Un mot japonais est écrit en kanji (漢字, caractère chinois) et hiragana (平仮名, caractère japonais), en kanji seulement, en hiragana seulement ou en katakana (片仮名). Les okuriganas (送り仮名) sont des suffixes qui suivent les racines écrites en kanji. Les règles pour les okuriganas ont changé plusieurs fois, et ont été standardisées par le gouvernement japonais en 1973.

5. L'indication n'est pas donnée à l'exception des indications relatives à l'utilisation d'un sens figuré, par l'étiquette {fig}, et à l'utilisation de noms de domaine des mots-vedette.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Notre approche est basée sur l'hypothèse que, si un exemple est un proverbe, ou si un mot dans un exemple est utilisé dans un usage collocatif, les lexèmes dans l'exemple traduit JE devraient différer de ceux de l'exemple traduit FE : dans les proverbes et les expressions collocatives, y compris les classificateurs, les mots semblent avoir tendance à être utilisés au sens figuré (Tomokiyo, Boitet & Mangeot, 2017), GT ne fournit pas de traductions adéquates pour eux (Tomokiyo & Boitet, 2016), mais plutôt des traductions word-to-word. Ainsi, les deux sorties contiennent des lexèmes différents, et on peut distinguer les proverbes et les collocations des exemples d'utilisation ordinaires.

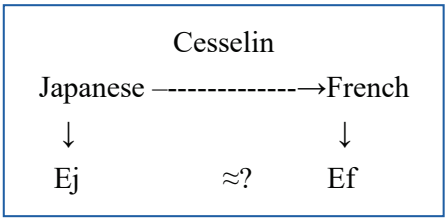


FIGURE 1: Hypothèse : r(Ej) ⊆ r(Ef) ⟹ usage collocatif, proverbial, idiomatique

2.1 Traduction japonais-anglais et français-anglais de 500 exemples par GT

La description dans le Cesselin est structurée pour une entrée comme : mot-vedette (en japonais), prononciation – donnée avec deux formes différentes d'écriture (en japonais et en translittération en caractères latins, appelée ロ-マ字 (romaji), les formes conjuguées pour les verbes et les adjectifs, la partie du discours (POS), des définitions (en français) et des exemples en japonais avec leur translittération en romaji, et une traduction ou des traductions en français.

Exemple : Extrait du verbe "nomu (drink)" du Cesselin

EXEMPLE : Extrait du verbe "nomu (drink)" du Cesselin

nomu [ma, mi, me] 飲む - 呑む【のむ】v.t.
1. Boire, avaler.
2. Aspirer, sucer, fumer.
3. Prendre.
4. .....
5. ....
- 飲まねば薬も効能なし (Nomaneba kusuri mo kônô nashi) "Qui veut la fin veut les moyens."
- 飲まぬ酒には酔はぬ (Nomanu sake ni wa yowanu) "Il n'y a point de fumée sans feu."
- 飲めや唄へや (Nome ya utae ya) "Faisons la noce et soyons joyeux!"
- … .....

Nous avons soumis à GT 500 exemples (contenus dans une vingtaine de mots-vedette) sélectionnés dans le Cesselin (voir les colonnes (a) et (d) dans le Tableau 1) : 飲む (nomu, boisson), 買う (kau, acheter), 走る (hashiru, ran), 居る (iru, rester, être), 来る (kuru, viens), 思う (omou, penser), 言う (iu, say), 可愛い (kawaii, jolie), 強い (tsuyoi, fort), 安い (yasui, bon marché), 直ぐ (sugu, bientôt, immédiatement), から (kara,

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

parce que, depuis, depuis), 家 (ie, maison, famille), 夜 (yoru, nuit), 足 (ashi, jambe), 車 (kuruma, voiture), 年 (nen, année), なん時 (nanji, à quelle heure), だけ (dake, seulement), ながら(nagara, while), 必要 (hitsuyou, nécessité), もう (mou, déjà), 事 (koto, chose). Nous avons utilisé GT pour produire des traductions du japonais vers l'anglais et du français vers l'anglais (voir les colonnes (b) et (c) du Tableau 1).

Tableau 1 : Extrait des résultats de l'expérience

	(a)	(b)	(c)	(d)	(e)
	Exemples japonais du Cesselin	J-E traductions par GT	F-E traductions par GT	Traductions françaises dans le Cesselin	Correspondant aux résultats pour J-E et F-E
1	飲まねば薬も効能なし。(Nomaneba kusurimo kounou nashi.)	Il n'y a pas d'indication pour un médicament si vous ne buvez pas.	Who wants the end wants the means.	Qui veut la fin veut les moyens.	Ø[a] (dic. de proverbes)[b]
2	飲まぬ酒には酔わぬ。(Nomanu sake niha yowanu.)	I'm not intoxicated with drinking alcohol.	There is no smoke without fire.	Il n'y a point de fumée sans feu.	Ø (dic. de proverbes)
3	飲めや唄へやの大騒ぎ。(Nome ya etae ya no oosawagi.)	Drinking and singing fudge fuss.	Let's do the wedding and be merry!	Faisons la noce et soyons joyeux !	Ø (collocation)
4	飲むに減らず吸うに減る。(Nomuni ni herazu suu ni heru.)	Reduce to suck without reducing to drink.	It is not by dint of drinking, but by dint of sucking that it diminishes.	Ce n'est pas à force de boire, mais bien à force de sucer que ça diminue,	Ø
5	飲むに減らず吸うに減る。(Nomuni ni herazu suu ni heru.)	Reduce to suck without being reduced to drink.	Les dépenses les plus répétées, bien que peu sensibles à chaque occasion, restent les plus préoccupantes.	{fig}[c] Les dépenses les plus répétées, quoique peu sensibles à chaque fois, restent les plus inquiétantes.	Ø (dic. de proverbes) {fig}
6	飲んだり吐出したりする。(Nondari hakidashitari suru.)	It drinks or spits out.	Swallow and turn in turn.	Avaler et rendre tour à tour,	Ø
7	飲んだり吐出したりする。(Nondari hakidashitari suru.)	It drinks or spits out.	Accept and refuse.	{fig} Accepter et refuser.	Ø {fig} (collocation)
8	飲んで吐き出した様。(Non de hakidashita you.)	Like drinking and spitting out.	As if we had made what we have just swallowed.	Comme si on avait rendu ce que l'on vient d'avaler	Ø
9	飲んで吐き出した様。(Non de hakidashita you.)	Like drinking and spitting out	avec a pale face and suddenly swollen.	{fig} Avoir une figure pâle et comme subitement boursouflée.	Ø {fig} (collocation)
10	飲んで悲しみを忘れる。(Non de kanashimi wo wasureru.)	Drink and forget sorrow.	To drown his sorrows, to leave his sorrows at the bottom of the glass.	Noyer ses chagrins, laissez ses peines au fond du verre.	Ø

a. Zéro correspondance au niveau des mots.

b. L'exemple figure dans le dictionnaire de proverbes.

c. {fig} : étiquette indiquant l'usage d'un sens abstrait, donnée dans les traductions en français d'exemples du Cesselin, même si la liste n'est pas exhaustive. Les étiquettes {botanique}, {zoologie}, etc. sont données pour montrer le domaine d'utilisation d'un mot.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

2.2 Comparaison et contraste des deux résultats de traduction

La colonne (e) du tableau 1 montre une comparaison manuelle des sorties de la traduction J-E (colonne (b)) avec les traductions F-E (colonne (d)) du point de vue des lexèmes dans un exemple. Le tableau 2 montre le contraste des deux sorties de traduction.

Tableau 2 : Comparaison entre les traductions J-E et les traductions F-E pour 500 exemples

100% de concordance (les résultats de la traduction J-E et de la traduction F-E par GT sur cet exemple incluent exactement les mêmes mots)	63/500	12,6%
100% * (les résultats de la traduction J-E et de la traduction F-E par GT sur cet exemple incluent des mots synonymes)	79/500	15,8%
Il y a concordance des vocables, avec ⊆ ou ⊇ (la traduction J-E est plus longue que la traduction F-E, mais les deux traductions comprennent presque les mêmes mots. De plus, la traduction F-E est plus longue que la traduction J-E, mais les deux traductions contiennent presque les mêmes mots.	26/500	5,2%
Il y a concordance au niveau de l'analyse morphosyntaxique (les résultats de la traduction J-E et de la traduction F-E par GT sur cet exemple incluent les mêmes mots pour le verbe prédicat et ses actants).	15/500	3,0%
Pas de concordance ni d'expression collocative	108/500	21,6%
Ø {fig} : Pas de concordance ni d'expression collocative avec le label {fig.}	8/500	1,6%
Ø : Pas de concordance et pas d'expression collocative (la traduction J_E et la traduction F_E par GT sur cet exemple n'incluent aucun mot commun.)	161/500	32.2%
Pas de concordance et il s'agit de proverbes (la traduction J-E et la traduction F-E par GT sur cet exemple n'incluent aucun mot commun et l'exemple japonais apparaît dans un dictionnaire de proverbes japonais[a])	32/500	6,4%
D : Pas de concordance avec le domaine d'utilisation	4/500	0,8%
X : Zéro sortie par GT	4/500	0,8%

D'après nos analyses, les exemples du Cesselin peuvent être classés en 5 classes : exemples d'usage ordinaires, proverbes, expressions collocatives, quantificateurs et exemples spécifiques au domaine.

3 Classer les exemples du Cesselin

Dans les exemples suivants pour le verbe "飲む(nomu, boire)" (a) est un proverbe (b) est une phrase incluant "飲む" dans un sens figuratif, et (c) est une phrase qui n'est pas un proverbe ni une collocation, mais un usage ordinaire. Quand on compare les lexèmes dans les traductions J-E et F-E par GT, alors que (a) et (b) ne contiennent aucun mot commun, en (c), les deux traductions ont deux mots ("boire" et "eau") en commun.

Cela confirme que les exemples montrant des usages ordinaires ont tendance à se traduire par la composition de mots communs. Nous avons pensé que ce phénomène pourrait être un atout pour classer tous les exemples du Cesselin, car les traduc-

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

tions japonais-anglais par GT sont presque des traductions mot à mot et, quand un mot dans les exemples japonais est utilisé dans le sens figuratif dans les contextes, les résultats de traduction ne correspondent jamais aux résultats de la traduction français-anglais.

Tableau 3 : Exemples d'un proverbe, d'une expression collocative et d'un usage habituel

	Exemples japonais	Traductions J-E	Traductions F-E	Traduction d'exemples japonais
(a) Proverbe	飲まぬ酒には酔わぬ [a] (nomanu sake ni-ha yowanu)	I'm not intoxicated with drinking alcohol.	There is no smoke without fire.	II n'y a point de fumée sans feu.
(b) Collocation	彼は妻くんに飲まれている (Kare wa saikun ni nomareteiru)	He is being drunk by his wife.	is wife is bowing him.	Sa femme le berne.
(c) Exemple d'utilisation ordinaire	水を飲む (Mizu wo nomu)	Drink water.	Drink water.	Boire de l'eau.

a. Traduction mot-à-mot : On ne se saoule pas avec le Sake que l'on ne boit pas.

Parmi les exemples d'usage courant, les proverbes, les expressions collocatives, les quantificateurs et les exemples spécifiques au domaine, nous pouvons distinguer les proverbes des autres types d'exemples en utilisant un dictionnaire de proverbes et en utilisant des labels tels que {fig} et {botanic}, {zoology. . . }, qui sont attachés à certains exemples dans le Cesselin. Nous pouvons aussi distinguer les expressions collocatives et les exemples spécifiques au domaine des autres types d'exemples, respectivement.

4 Procédure de classification des exemples du Cesselin

Nous proposons les étapes suivantes pour la classification des exemples.

4.1 Différenciation des proverbes des expressions collocatives

Étape 1) Nous distinguons les proverbes des expressions collocatives, en utilisant un dictionnaire de proverbes. Si un exemple apparaît dans le dictionnaire, nous le classons comme un proverbe.

Etape 2) Pour les expressions collocatives, si la traduction d'un exemple dans le Cesselin est marquée comme utilisation figurative ({fig}), nous la classons comme une expression collocative.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Tableau 4 : Exemples pour un mot ayant un sens figuré, avec ou sans l'étiquette {fig}

	Exemples japonais	Traductions J-E	Traductions F-E	Traduction française d'exemples japonais
Collocation sans indication	人を飲んでかかる (Hito wo non de kakaru)[a]	I'm drinking people and hanging.	Miss someone, look down.	Manquer à quelqu'un, regarder de haut.
	痛い上に貼る (Itai ue ni haru)[b]	Painfully stick it on	misfortune on misfortune.	malheur sur malheur.
Collocation avec indication {fig}	飲んだり吐き出したりする(Nondari hakidashitari suru)[c]	To drink or to spit.	Accept and refuse.	{fig} Accepter et refuser.

a. Traduction mot-à-mot : menacer quelqu'un en avalant.
b. Traduction mot-à-mot : coller quelque chose à la partie souffrante.
c. Traduction mot-à-mot : boire et cracher alternativement.

4.2 Évaluation d'expressions collocatives sans étiquette dans le Cesselin

Étape 3) Il y a des cas où les traductions françaises dans le Cesselin n'ont pas d'étiquette {fig} et, en l'état, elles sont utilisées au sens figuré. Si leurs traductions en anglais n'ont pas de lexème en commun ($r(Ej) \cap r(Ef) = \emptyset$), nous les classons en tant qu'expressions collocatives.

Tableau 5 : Exemples avec sens figuré sans aucune étiquette

Mots-vedette	Exemples japonais	Traductions J-E	Traductions F-E	Traduction en français pour les exemples japonais
言う (iu, say, tell)	言うに言われないほど (iu ni iware nai hodo)[a]	Not to be told to say	Beyond any expression.	Au-delà de toute expression.
足 (ashi, jambe, pied)	足を洗う (Ashi wo arau)[b]	Wash your feet	Rise from a lower class.	S'élever d'une classe inférieure.

a. Traduction mot-à-mot : de façon à ne pas pouvoir dire si je dis.
b. Traduction mot-à-mot : se laver les pieds. L'expression 足を洗う a du sens en usage concret et en en même temps en usage figuré. Nous n'avons pas encore résolu ce problème.

Les deux traductions du tableau 5 n'ont pas de lexème en commun. Dans ce cas, la comparaison des exemples traduits est effectuée avec une liste KWIC (KeyWords In Context).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

4.3 Exemples traduits ayant une relation de synonymie avec des mots existants dans chaque exemple

Étape 4) Dans le tableau 6, la traduction J-E pour l'exemple 必要費 est différente de la traduction F-E pour "cost" et "expenses", et la traduction J-E pour l'exemple 車 est différente de la traduction F-E pour "ride" et "get in "のる", qui sont les verbes prédicatifs des phrases. Dans ces cas, la relation de synonymie entre les deux mots est examinée, et si les deux mots sont synonymes, l'exemple est classé comme exemple d'utilisation ordinaire.

Tableau 6 : Les deux exemples traduits qui ont des lexèmes synonymes

Mots-vedette	Exemples japonais	traduction J-E	Traductions F-E	Traductions en français	Comparaison
必要 (hitsuyou)[a]	必要費[b] (hitsuyou hi)	Coût nécessaire	Necessary expenses.	Dépenses nécessaires.	100%*[c]
車 (Kuruma)[d]	車に乗る (kuruma ni noru)[e]	I ride a car.	Get in the car	Monter en voiture.	100%*

a. Nécessaire.
b. Coût nécessaire.
c. 100%* signifie que la correspondance des mots est 100% en utilisant la synonymie entre mots.
d. Voiture.
e. Monter en voiture.

4.4 Exemples traduits ayant une longueur différente

Étape 5) Les exemples traduits en anglais dans le tableau 7 ont une longueur différente en raison de la nominalisation du verbe prédicatif, du degré différent de politesse des énoncés, etc. Dans ce cas, l'exemple est analysé de la manière suivante : si le verbe prédicatif et son sujet (ou son objet) sont les mêmes, les deux exemples sont considérés comme ayant le même sens, et ils sont classés comme des exemples ordinaires.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Tableau 7 : Deux exemples traduits de longueurs différentes

Mot-vedette	Exemples japonais	Traductions J-E	Traductions F-E	Traduction en français pour les exemples japonais
もう (mou) (J-F)	もう休みましょう (Mou yasumi mashô) [a]	Let's have a rest now.	Let's rest now!	Reposons-nous. maintenant!
ながら (nagara) (J-F)	失礼ながら山田さんではございませんか (Shitsurei nagara Yamada san de-ha gozaimasen ka) [b]	Excuse me, aren't you Mr. Yamada?	Please excuse me, but aren't you Mr. Yamada?	Veuillez bien m'excuser, mais n'êtes-vous pas monsieur Yamada?

a. Reposons-nous maintenant.
b. Veuillez bien m'excuser, mais n'êtes-vous pas monsieur Yamada ?

5 Utilisation du dictionnaire de Mots Universels (UW) comme un pivot de lexèmes interlingues

Afin d'obtenir des informations sur la synonymie et les relations morphosyntaxiques entre les mots dans les deux exemples traduits, nous avons utilisé un dictionnaire d'UWs du système de traduction automatique UNL [6] comme un langage pivot interlingue. Un UW est une chaîne de caractères qui représente un sens de mot (renvoyant à un concept). Une représentation sémantique en UNL a la forme d'un hypergraphe, où les nœuds sont des UW (appelés Universal Words) ayant des attributs sémantiques, et les arcs portent des relations sémantiques [7] entre deux UW. Voici des exemples d'UW.

```
expense(icl>cost),
start(icl>begin),
look_for(agt>thing, equ>search, icl>examine(icl>do, obj>thing)),
樽 (taru, cask):cask(icl>wine, equ>2200 litres)
```

"agt" indique que l'agent de l'UW ("look*for[. . .]") est de type "thing", "equ>search" dit que ce sens a un lien "equ" dans la "préontologie" UNL [8] vers un UW de mot-vedette "search", et est un sous-sens de "examine" (

"examiner"), l'UW [de mot-vedette] "examine" dénotant le sens d'un verbe d'action dont l'objet (profond) est de type nominal et signifie "chose" (Uchida, Zhu et Della Senta, 2006).

6. Le projet UNL (Universal Networking Language) a été fondé à l'Institut IAS de l'Université des Nations Unies (UNU) à Tokyo en 1996, avec le soutien de l'UNU et le soutien financier de la société ASCII et de l'UNL-IAS, au titre du développement de la technologie de pointe. http ://www.undl.org/unlsys/unl/unl2005/attribute.htm (Uchida, Zhu et Della Senta, 2006)

7. Les lexèmes interlingues (UW) sont créés par des membres d'UNL, qui stockent toute l'information apparentée dans une structure de treillis, où les UW sont reliés par des relations comprenant des relations hiérarchiques (10 niveaux) comme "icl" (une sorte de) et "iof" (une instance de), et signifient la sous-signification et la quantité équivalente du mot-clé, respectivement. http ://www.undl.org/unlexp/.

8. Le dictionnaire UNL-UW contient à l'heure actuelle, 1 269 421 mots-vedette pour le japonais, 520 305 mots-vedette pour le français et 1 458 686 mots-vedette pour l'anglais. Les attributs sémantiques sont constitués de 58 étiquettes et de 39 étiquettes de relations sémantiques.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

6 Vers une classification automatique

Pour effectuer une classification automatique des exemples, nous avons besoin d'un dispositif qui associe des phrases, des mots ou des lexèmes et analyse les exemples. Il passe par les étapes suivantes, suivant l'analyse manuelle précédente.

1. Si des exemples japonais apparaissent dans un dictionnaire de proverbes japonais, ils sont annotés comme {prov}, pour "proverbe".
2. Si la traduction française dans le Cesselin est étiquetée par {fig}, l'exemple est annoté comme {colexp}, pour "collocation", et si la traduction française dans Cesselin est étiquetée par {botanic}, zoology},..., l'exemple est annoté {domexp}, pour "utilisation spécifique au domaine" indiquée.
3. Si les lexèmes de deux sorties de traduction correspondent à 100%, l'exemple est annoté comme {ordusg}, pour "exemple d'utilisation ordinaire".
4. Si les deux sorties de traduction ne sont pas dans les cas 1, 2, 3 ci-dessus, elles sont examinées avec une liste KWIC en utilisant un logiciel, appelé Sketch Engine [9]. Si un exemple apparaît avec certaines occurrences, il est considéré comme une collocation ou un quantificateur (Alda, 2011).
5. Contrairement à la liste KWIC, si un exemple est un syntagme nominal composé de "nombre nom", alors l'exemple est annoté comme {quant} pour "quantificateur" (Tomokiyo, Boitet & Mangeot, 2017) (Miyagawa 1989).
6. Si les exemples dans les deux résultats de traduction n'apparaissent pas dans la liste KWIC avec certaines occurrences, la synonymie entre les mots dans les deux résultats de traduction est examinée en utilisant un dictionnaire UNL-UW.
7. Si le verbe prédicatif et ses actants [10] dans les deux traductions de l'exemple correspondent, et si les mots du reste des exemples traduits correspondent au niveau de la synonymie, alors l'exemple est annoté comme {ordusg} pour "exemple d'usage ordinaire".
8. S'il n'y a pas de correspondance entre les lexèmes contenus dans les deux traductions de l'exemple, il est annoté {colexp}, pour "utilisation en collocation", sauf s'ils sont déjà classés comme proverbes, quantificateurs, ou usage spécifique au domaine.
9. Si un exemple n'appartient pas à un des cas 1, 2, 3, 4, 5, 6 7, 8, mentionnés ci-dessus, l'exemple est annoté comme {ordusg} pour "exemple d'utilisation ordinaire".

Remarque

Les exemples dans les dictionnaires ne sont en général pas très longs (environ 10 mots en moyenne), et comprennent en certaines proportions des proverbes, des groupes

9. Sketch Engine est un outil de gestion de corpus ; il contient 400 corpus prêts à l'emploi. Nous y ajoutons le corpus de Mathieu Mangeot.

10. Afin de connaître le verbe prédicatif et les actants de chaque exemple traité, nous avons utilisé un analyseur appelé MeCab.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

nominaux, des groupes adjectivaux, ou des phrases incomplètes. Donc, si nous cherchons à appliquer notre méthode de classification/annotation à d'autres documents, nous ne sommes pas sûrs de pouvoir obtenir les mêmes résultats que dans le cas d'un dictionnaire.

Bibliographie

Alda Mari (2011) *Quantificateurs polysémiques*, Université Paris-Sorbonne, vol. 23, France.

Mathieu Mangeot-Nagata (2016) *Collaborative construction of a good quality, broad coverage and copyright free Japanese-French dictionary*. HAL-01294566.

Mutsuko Tomokiyo ; Christian Boitet ; Mathieu Mangeot (2017) *Development of a classifiers/quantifiers dictionary towards French-Japanese MT*, MTSummit 2017, Japan

Mutsuko Tomokiyo ; Christian Boitet (2016) *Corpus and dictionary development for classifiers/quantifiers towards French-Japanese machine translation*, COLING, Cog-Alex 2016, Japan

Shigeru Miyagawa (1989) *Structure and case marking in Japanese*, Syntax and Semantics, vol. 22, New York.

Uchida, H. ; Zhu, M. ; Della Senta, T.G. (2006) *Universal Networking Language*. UNDL Foundation, Japan.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454