Déterminer automatiquement le genre et le style d'expressions idiomatiques en japonais

Raoul Blin

Blin R., 2015, Déterminer automatiquement le genre et le style d'expressions idiomatiques en japonais, Japon Pluriel, Piquier.

BLIN Raoul
CRLAO-CNRS, Paris

Déterminer automatiquement le genre et le style d'expressions idiomatiques en japonais

1 Introduction

A ce jour, au moins en japonais, les lexiques et dictionnaires (monolingues ou plurilingues) proposent peu d'informations relatives au genre et au style des expressions idiomatiques. Pourtant, ces informations jouent un rôle important, que ce soit pour produire ou pour interpréter un énoncé. Il serait donc intéressant de disposer de telles informations.

Les expressions se comptant par centaines, renseigner un dictionnaire entier est difficile à réaliser manuellement car cela réclame un nombre d'heures/hommes très important. L'automatisation est inévitable.

L'analyse automatique consiste à évaluer la fréquence des expressions dans des corpus considérés comme représentatifs. L'appartenance d'une expression à un genre ou un style est corrélée à la fréquence de l'expression dans le corpus représentatif de ce genre ou style. Cette approche présente l'intérêt de graduer l'appartenance, ce qui est plus réaliste que de classer de façon binaire (appartient ou n'appartient pas).

Cette méthode a été appliquée précédemment pour l'analyse du genre et du style de 16 000 noms (Blin, 2012, abr. « DFJC », Blin, 2012b). Il s'agissait de formes invariables. La question était de savoir si la méthode et les outils étaient utilisables aussi pour des structures contenant des formes fléchies, comme les expressions idiomatiques. Pour y répondre, nous avons repris les mêmes outils et les mêmes corpus (à quelques différences quantitatives près) et avons étudié quatre cent expressions comprenant des verbes conjugués. Nous présentons ici les résultats.

Désormais, pour simplifier le discours, nous utiliserons le terme de « *genre* » pour désigner le « genre et le style ».

Dans le chapitre 2, nous présentons les expressions étudiées. Dans le chapitre 3, nous présentons les *genres* étudiés et les corpus correspondants. Le chapitre 4 est consacré à la méthodologie : le logiciel utilisé, les difficultés rencontrées pour l'analyse automatique des corpus et les solutions adoptées en tenant compte des contraintes imposées par le logiciel. Enfin nous proposerons quelques résultats.

2 Présentation des expressions idiomatiques étudiées

L'étude a porté sur 400 expressions idiomatiques. Ce sont des constructions composées d'un prédicat accompagné d'au moins un argument, avec un sens non compositionnel. La structure syntaxique est figée ainsi que le choix des lexèmes. Morphologiquement, elles ne sont que partiellement figées, puisque le prédicat peut être fléchi. Le sens non compositionnel est valable pour au moins deux flexions du prédicat.

Les expressions retenues se démarquent des autres expressions parce qu'un des arguments fait référence à une partie du corps. Par exemple :

Ex.1 : *atama ni kuru*
tête p.« à » venir
[ça] [m']énerve

Dans l'exemple 1, l'expression a le sens non compositionnel de « ça énerve ». Ce sens est valable pour au moins deux conjugaisons : les formes neutre (*kuru*) et parfait (*kita*). Enfin, l'argument *atama ni* (« pied »), fait référence à une partie du corps.

Certaines de ces constructions ont un double sens, l'un compositionnel et l'autre non compositionnel. Par exemple

Ex.2 : *ashi wo arau*
pied p.O laver

a le sens non compositionnel « s'acheter une conduite » et le sens compositionnel « laver le(s)/un/des pied(s) ». Cette polysémie n'est pas sans conséquences sur les comptages et nous y reviendrons.

3 Corpus exploités

Le corpus est divisé en sous-corpus chacun considérés comme représentatifs d'un *genre* particulier de la langue écrite contemporaine. La notion de représentativité prêtant à débat, nous nous en sommes tenu des critères pratiques.

Nous considérons que des textes relèvent d'un même « genre/style » si d'une part ils émanent d'une même source ou bien de sources qui partagent des points communs explicitement énonçables, et d'autre part si ils ont une structure textuelle similaire (structuration en titre, paragraphes etc.). A chaque sous-corpus nous a été attribué une étiquette qui rendait compte au mieux du *genre*. Ce choix comporte évidemment une part d'arbitraire.

Par exemple, nous avons constitué un corpus de « déclarations de brevets », constitués des textes disponibles sur l'unique site www.patentjp.com et structurés de la même manière. Ce corpus a été étiqueté « brevets ». Le corpus représentatif du *genre* « journalistique » est constitué de textes émanant de plusieurs sources, mais qui toutes ont en commun d'être les sites web de journaux

papier, et de présenter en première page les informations du moment. Seuls les textes constitués d'un titre suivi d'un ou plusieurs paragraphes ont été sélectionnés (il s'agit *de facto* des informations mises en ligne).

Pour garantir une plus grande représentativité, les corpus ne sont pas échantillonnés, contrairement par exemple au corpus de référence BCCWJ (Maruyama, 2009). On dispose donc de collections complètes de textes.

Le corpus est à peu près le même que celui utilisé pour le DFJC, à deux différences près. La première est la présence d'un nouveau sous-corpus, constitué de déclaration de brevets. La seconde est que la taille de plusieurs des sous-corpus a sensiblement augmenté suite à l'ajout de textes produits en 2012.

Voici une présentation synthétique de l'ensemble des sept sous-corpus. Y sont donnés la taille des sous-corpus (en nombre d'énoncés), la périodicité de renouvellement et quelques caractéristiques relatives au mode de production et au lectorat. Une description corpus par corpus est fournie dans Blin R. (Blin, 2012a).

(conventions : + oui ; - non ; +/- pas de généralisation possible pour ce corpus)

		nb énoncés	périodicité du renouvellement	relecture	spécialisation thématique	restriction de producteur	restriction de lectorat
monologue	livres blancs	105 520	annuel, complet	+	-	+	+
	dictionnaire Daijirin	176 809	partiel, pluriannuel	+	-	+	-
	journaux	167 819	quotidien, complet	+	-	+	-
	textes juridiques	34 688	annuel, complet	+	+	+	+
	déclarations de brevets	7 270 517	(inconnu)	+	+/-	+	+
dialogue	QR gouv.	54 901	annuel, complet	+	-	+	+
	QR divers	136 946	quotidien, partiel	+/-	-	-	-
	Tchats	23 547	quotidien, complet	+/-	-	-	-

4 Outils et méthode d'analyse

Le comptage des occurrences des expressions dans les corpus a été effectué automatiquement. Nous présentons ici les avantages et inconvénients de l'outil, compte tenu de la nature des structures cherchées, et la façon dont été traitées les difficultés.

4.1 Logiciel

Le dénombrement des occurrences des expressions dans les corpus a été effectué à l'aide du logiciel SAGACE version 4.2.0 (Blin, 2012c). SAGACE est un moteur de recherche de motifs. Il n'effectue pas d'analyse morphosyntaxique ou sémantique complète des énoncés.
Ce logiciel a été retenu pour sa facilité de mise en oeuvre.
Contrairement aux analyseurs morphosyntaxiques symboliques, il ne nécessite pas de développer des lexiques-grammaires dont la création et la maintenance sont lourdes. Parmi les ressources *disponibles pour le public*, il n'existe d'ailleurs pas aujourd'hui de lexique-grammaire complet pour le japonais. Les analyseurs morphosyntaxiques à vaste couverture sont en effet tous statistiques.
Par rapport aux analyseurs statistiques type Mecab [1] etc., l'intérêt de SAGACE est de ne pas nécessiter d'apprentissage. Pour améliorer les performances des logiciels statistiques, un apprentissage serait nécessaire pour chaque sous-corpus. Il faudrait pour cela disposer d'autant de corpus d'apprentissage que de *genres* à étudier. Il existe bien des dispositifs entraînables sur des corpus partiellement annotés (Flannery, Miyao, Neubig, & Mori, 2011) mais cela n'implique pas moins un entraînement lourd à mettre en place.
Par ailleurs, SAGACE est autonome et ne fait pas appel à d'autres logiciels. Il assure à lui seul toutes les fonctions nécessaires pour une analyse : interface pour la requête, moteur de recherche de motifs, affichage des résultats.
Enfin, comme SAGACE exploite du texte brut, il ne nécessite pas de traitement préalable des corpus. De surcroît, les ressources lexicales nécessaires pour le japonais sont déjà disponibles dans un format compatible avec cet outil.

La contrepartie à cette souplesse est que le logiciel ne fait pas d'analyse des phrases complètes, ce qui accroît le risque d'erreurs d'analyse. Pour le comptage des expressions, nous avons opté pour limiter ce risque, au dépend de l'exhaustivité. Nous nous en sommes tenu pour cela au comptage des occurrences des expressions dans des distributions très spécifiques, dont la segmentation est relativement sûre. Il en résulte qu'en principe, les nombres d'occurrences et les fréquences obtenus sont inférieurs à la réalité. Nous n'avons pas de moyen d'évaluer la différence entre valeurs réelles et valeurs obtenues.

4.2 Requête

La description du patron à chercher tient compte de l'insécabilité des expressions étudiées. En effet, l'interprétation non compositionnelle des expressions n'est possible que si aucun syntagme, quel que soit son statut syntaxique, n'est introduit dans la chaîne. Par exemple l'interprétation figée

1 http://code.google.com/p/mecab/

« s'acheter une conduite » n'est plus possible si l'on introduit un adverbial de temps (*kinou*, « hier ») dans la chaîne des arguments et du radical verbal :

Ex.3 : *ashi wo <u>kinou</u> aratta*

pied p.O <u>hier</u> avoir_lavé

[je me suis] lavé les pieds <u>hier</u>.

L'autre particularité des expressions est que les arguments ne peuvent être élidés sans affecter l'interprétation non compositionnelle. Ainsi, l'interprétation non compositionnelle n'est plus possible si l'argument *ashi* (« pied ») est élidé :

Ex.4 : *ore ga aratta*
moi p.S avoir_lavé

C'est moi qui [l']ai lavé

De ce fait, la recherche des expressions peut être limitée à la recherche de la chaîne des arguments et du radical verbal « complet » (sans élision) et en l'état (sans inversement d'arguments et sans ajout).

Compte tenu du figement syntaxique et lexical des expressions, et de l'impossibilité pour le logiciel de procéder à une analyse morphosyntaxique, pour décrire les expressions figées dans le lexique et dans la requête pour SAGACE, nous avons recouru a l'artifice suivant. Le radical verbal et sa chaîne d'arguments ont été considérés comme constituant un « radical verbal » insécable, auquel est adjointe la conjugaison.

< *ashi wo ara* > + *u*
<N particule radical_verbal> « rad.v » + conj

La conjugaison de ce radical verbal *ad hoc* est la même que celle du verbe qui le compose. Ainsi <*ashi wo ara*> se conjugue comme *arau* (« laver »).

Les expressions ont donc été lexicalisées dans le même format que les verbes :

ashi wo ara [[vRad & distrib:transmet:w]]

Il est inutile d'insister sur le fait que cette analyse n'a aucun fondement linguistique et qu'elle n'est motivée que par des aspects pratiques liés au logiciel utilisé. Cela n'a de toute façon aucune incidence sur le comptage.

SAGACE n'est pas concu pour analyser les flexions. Les conjugaisons sont donc enregistrées d'un seul tenant dans les dictionnaires. Par exemple, la conjugaison « désidératif, non poli, parfait, négatif » est enregistrée d'un seul tenant comme suit :

```
cat:conj & conjDesideratif & conjNeutre & conjTa


takunakatta    // itidan
takunakatta    // kuru
```

itakunakatta // 5 dan - a
kitakunakatta // 5 dan - ka
kitakunakatta // *iku*
gitakunakatta // 5 dan - ga
sitakunakatta // 5 dan - sa
chitakunakatta // 5 dan - ta
nitakunakatta // 5 dan - na
bitakunakatta // 5 dan - ba
mitakunakatta // 5 dan - ma
ritakunakatta // 5 dan - ra
sitakunakatta // *suru*

Pour chaque classe morphologique de verbe (itidan, kagyoudan, gagyoudan etc.) il y a une entrée.

Fondamentalement, le patron de la chaîne cherchée est de la forme :

<vRad> + <conjugaison>

Ayant posé en plus des contraintes à gauche et à droite pour limiter les risques d'erreur de segmentation, le patron à chercher est décrit dans la requête comme suit :

```
>0 cat : particule | ponctuation | marqeDebEnonce
=0 cat:vRad
=0 cat:conj
=0 cat:ponctuation
```

4.3 Le traitement des cas difficiles

Trois difficultés notoires ont été rencontrées, qui sont susceptibles d'affecter les résultats.

La première difficulté tient à la polygraphie des composants des expressions. Les mots en japonais peuvent en effet apparaître sous plusieurs graphies dans les textes. Ils peuvent être transcrits en hiragana, katakana, en sinogrammes, ou dans un mélange relativement libre des trois. Il peut y avoir aussi plusieurs choix de sinogrammes. La plupart des dictionnaires destinés au traitement automatique choisissent de lexicaliser les variantes les plus probables. Mais cela ne recouvre pas toutes les variantes possibles. Pour les expressions, cela aboutirait à la production d'un lexique de taille démesurée.

Nous n'avions pas de moyen de régler élégamment le problème. Nous nous en sommes donc tenu à une seule graphie, la plus standard. Le nombre d'occurrences relevées sur les corpus pour chaque expression peut donc être inférieur au nombre d'occurrences réel puisque toutes les variantes graphiques n'ont pas été prises en compte.

La deuxième difficulté rencontrée lors du comptage était de distinguer, pour une expression donnée, les occurrences s'interprétant compositionnellement et les occurrences ayant l'interprétation non compositionnelle. En principe, les occurrences avec interprétation compositionnelle ne devraient pas être comptées puisque nous nous intéressons au sens non compositionnel. Mais l'exclusion nécessiterait une analyse sémantique, ce que ne permet pas SAGACE, ni aucun autre analyseur disponible à ce jour. Il faudrait donc recourrir à l'analyse par un expert humain. A ce stade du travail, cette expertise n'a pas été effectuée. En comptant aussi les occurrences interprétées compositionnellement, le nombre d'occurrences des expressions polysémiques (disposant d'une interprétation non compositionnelle et d'une interprétation compositionnelle) peut avoir été surestimé.

La troisième difficulté est liée aux restrictions sur les environnements gauche et droit des expressions. Pour limiter les erreurs de segmentation, nous avons défini les contextes gauche et droit immédiats (voir section précédente). Cette contrainte peut entraîner une sous-estimation du nombre d'occurrences puisque les expressions ne sont pas comptées si elles apparaissent dans d'autres environnements (notamment si le verbe est la forme en *-te* et non suivi d'une particule ou d'une ponctuation).

Les différents problèmes ainsi rencontrés entraînent tantôt une surestimation du nombre d'occurrences, tantôt à une sous-estimations. Nous n'avons pas moyen de connaître la marge d'erreur globale et de pondérer les résultats en conséquence.

5 Résultats

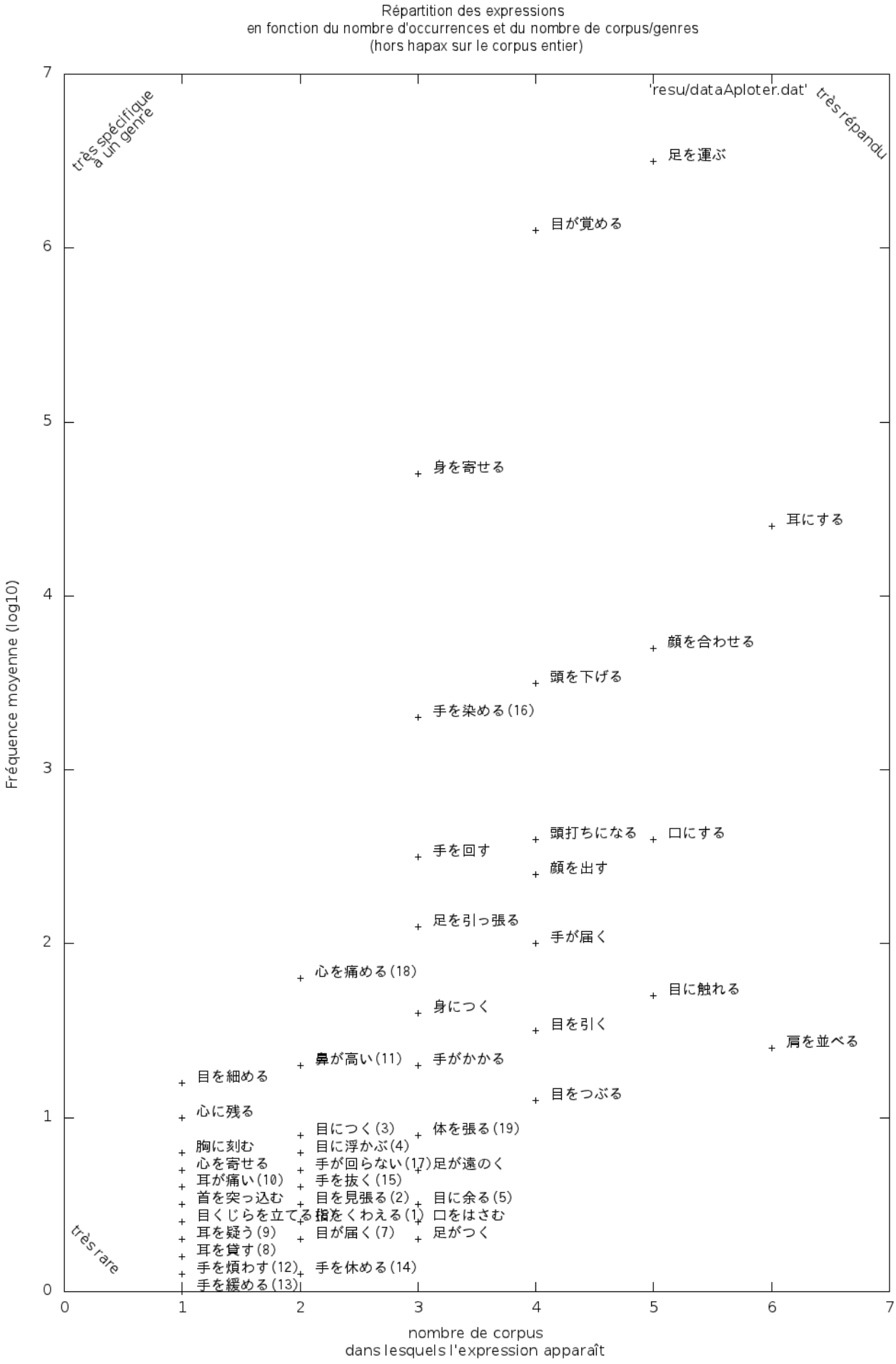
Voici les résultats pour quelques entrées, sur le modèle de présentation du DFJC.

Le tableau fait apparaître la fréquence des expressions dans chaque corpus ainsi que différents chiffres qui permettent de décrire la variété des *genres* couverts. Il s'agit, pour chaque expression, de la fréquence moyenne, de l'écart entre la fréquence minimale et la fréquence maximale, et enfin du nombre de corpus dans lequel l'expression apparaît au moins une fois.

	Fréquences									Occ total	Nb Corpus non nuls
	Livres blancs	Brevets	Journ.	Txt jurid.	QR gouv.	QR divers	Tchats	moyenne	max-min		
ashi wo hakobu	3.0	0.4	10.1	0	14.60	11.0	0	6.5	14.6	67	5
me ga sameru	0	0.1	0.8	0	0	11.0	24.80	6.1	24.8	30	4
mi wo yoseru	0	0	22.8	0	1.80	0	3.50	4.7	22.8	29	3
mimi ni suru	3.0	0	1.7	0	3.60	11.0	7.10	4.4	10.9	24	6

Les résultats permettent d'élaborer un graphique montrant la couverture des genres. Le graphique prend en abscisse le nombre de corpus où les expressions apparaissent au moins une fois, et en ordonnée la fréquence moyenne sur une échelle logarithmique. Pour des questions de lisibilité,

seule une partie des expressions a été reportée. Les expressions qui n'apparaissent qu'une seule fois sur la totalité du corpus (tous sous-corpus confondus) ne sont pas comptées.


Répartition des expressions
en fonction du nombre d'occurrences et du nombre de corpus/genres
(hors hapax sur le corpus entier)
'resu/dataAploter.dat'
très spécifique à un genre
très répandu
très rare
Fréquence moyenne (log10)
nombre de corpus dans lesquels l'expression apparaît
0
1
2
3
4
5
6
7
足を運ぶ
目が覚める
身を寄せる
耳にする
顔を合わせる
頭を下げる
手を染める(16)
頭打ちになる
口にする
手を回す
顔を出す
足を引っ張る
手が届く
心を痛める(18)
目に触れる
身につく
目を引く
肩を並べる
鼻が高い(11)
手がかかる
目を細める
目をつぶる
心に残る
目につく(3)
体を張る(19)
胸に刻む
目に浮かぶ(4)
心を寄せる
手が回らない(17)
足が遠のく
耳が痛い(10)
手を抜く(15)
首を突っ込む
目を見張る(2)
目に余る(5)
目くじらを立てる
指をくわえる(1)
口をはさむ
耳を疑う(9)
目が届く(7)
足がつく
耳を貸す(8)
手を煩わす(12)
手を休める(14)
手を緩める(13)

6 Conclusion et perspectives

Nous avons présenté la méthode et les outils pour décrire le genre et style des expressions idiomatiques avec prédicat fléchi, en japonais. Ce travail exploratoire a été limité aux expressions contenant au moins un mot faisant référence à une partie du corps.

Il en ressort que les outils et la procédure utilisés dans un travail précédent sur les noms communs, qui ne comprennent pas de flexions, était aussi utilisables pour des constructions avec flexion. Les erreurs d'analyse provoquées par l'outil sont partiellement différentes de celles rencontrées pour les noms communs, mais ne sont pas le fait de la présence de la flexion. Les principales difficultés rencontrées, polygraphie et ambiguïté sémantique, ne sont pas non plus propres à l'outil utilisé puisque les autres dispositifs d'analyse existants ne les résoudraient pas non plus.

Suivant la même procédure, nous menons actuellement une analyse de plus grande envergure sur la totalité des expressions japonaises. A la différence de la présente étude exploratoire, nous effectuerons en plus une distinction entre les expressions sémantiquement complètement figées (aucun sens compositionnel) et les autres expressions qui ont un sens compositionnel et un sens non compositionnel.

7 Références

BLIN Raoul. *Dictionnaire de fréquence du japonais contemporain - 16 000 noms.* Paris, You Feng, 2012a.

BLIN Raoul. Automatic Addition of Genre Information in a Japanese Dictionary. *Acta Linguistica Asiatica*, *2*-2, 2012b : 83–96.

BLIN Raoul. SAGACE v4.2.0. Blin Raoul. 2012c.

FLANNERY Daniel, MIYAO Yusuke., NEUBIG Graham, MORI Shinsuke. *Training Dependency Parsers from Partially Annotated Corpora (BibTex, Code)*. Presented at the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand. 2011, November.

Maruyama Takehiko . “Gendai nihongo kakikotoba keikin koopasu” monitaa kaihatu deeta (2009nendoban) sanpuring houhou ni tuite [“Balanced Corpus of Contemporary Written Japanese” (v.2009)] about the method of sampling]. National Institute for Japanese Language and Linguistics. 2009.