Comparaison de deux outils d'analyse de corpus japonais pour l'aide au linguiste, Sagace et MeCab

Raoul Blin

21ème Traitement Automatique des Langues Naturelles, Marseille, 2014

Comparaison de deux outils d'analyse de corpus japonais pour l'aide au linguiste, Sagace et MeCab

Raoul Blin
CNRS-CRLAO, 131 Bd St.Michel, 75006 Paris
blin@ehess.fr

Résumé. L'objectif est de comparer deux outils d'analyse de corpus de textes bruts pour l'aide à la recherche en linguistique japonaise. Nous mesurons leur précision dans la tâche de comptage de chaînes de morphes. Les deux outils représentent chacun une approche spécifique. Le premier, un dispositif basé sur l'analyseur morphologique statistique MeCab, segmente et étiquette préalablement les phrases complètes. Le second compte les occurrences de la chaîne dans le texte en l'état. Les performances de Sagace sont globalement un peu inférieures mais la différence est moins importante qu'attendu. Du fait de leur facilité de mise en œuvre, les outils comme Sagace sans analyse morphologique préalable sont donc des outils malgré tout intéressants pour le linguiste.

Abstract. Our purpose is to compare two tools used to help linguists analyze large corpora of raw Japanese text. We measure their precision while counting strings of morphs. Each tool is representative of a specific approach. The first tool is based on the statistical morphological analyzer MeCab. It first tokenizes and POS tags the whole sentence before searching and counting strings. The second tool, Sagace, searches and counts within the text as it is. In accordance with our assumptions, Sagace performed slightly worse overall but the difference is not as marked as expected. Taking into account the needs of linguists, Sagace is nevertheless useful for many tasks.

Mots-clés : Japonais, Corpus, Analyseurs morphologique, MeCab, Sagace
Keywords: Japanese, Corpus, Morphological analyzer, MeCab, Sagace

Le linguiste travaillant sur le japonais écrit contemporain dispose de nombreux outils d'analyse de corpus, que ce soit pour dénombrer des collocations ou constituer des concordances. On peut séparer ces outils en deux groupes. Le premier rassemble les outils qui s'en tiennent à chercher un patron dans du texte brut. Les outils du second groupe procèdent en deux temps : le texte est d'abord segmenté en morphes et étiqueté en parties du discours, puis a lieu la recherche et le comptage de patrons. En japonais, à cause de l'absence de marques de frontière des mots et des nombreuses ambiguïtés graphiques, la seconde approche paraît la plus fiable. Il reste cependant à mesurer l'écart réel de qualité entre les dispositifs. Si cet écart est « raisonnable », alors la légèreté et la facilité de mise en oeuvre des premiers dispositifs pourraient leur donner un avantage, face à des dispositifs à base de segmentation préalable dont la mise en oeuvre est longue. En effet, les analyseurs morphologiques sont aujourd'hui tous statistiques et nécessitent un entraînement. Dans la pratique et en particulier en TAL, de nombreuses études utilisent les lexiques existants. Mais de nombreuses recherches en linguistique nécessitent de modifier ces lexiques, que ce soit les entrées lexicales elle-même ou bien les parties du discours Pour conserver une analyse morphologique statistique optimale, il faudrait à chaque modification du lexique réentraîner l'analyseur, ce qui demande un temps considérable incompatible avec les pratiques en linguistique.

Dans cet article, nous comparons les performances d'extraction et de comptage des noms communs, avec et sans analyse morphologique (statistique) préalable. Pour cela, nous utilisons des outils déjà disponibles, Sagace (Blin, 2012) et un dispositif basé sur l'analyseur morphologique statistique MeCab (Kudo, 2006). Dans la première partie, nous présentons les caractéristiques du japonais écrit qui sont susceptibles d'affecter les résultats des comptages automatiques effectués par ces dispositifs. La deuxième partie est consacrée à la description des outils et des ressources utilisés pour le comparatif. Les résultats de la comparaison figurent dans la troisième partie.

1 Rappel sur le japonais écrit

Nous rappelons quelques propriétés du japonais écrit qui sont susceptibles d'influencer l'analyse automatique des noms communs. Ces propriétés ont été largement décrites ailleurs de façon plus globale (voir par exemple en français (Yazawa, 2006)).

Le japonais utilise trois écritures : les hiragana et katakana que l'on peut qualifier de phonétiques (50 caractères de base pour chaque série) et un ensemble d'environ 2000 sinogrammes. Les noms communs empruntés au XXè S. aux langues étrangères sont écrits en katakana. Les autres peuvent indifféremment être transcrits avec l'une des trois écritures, voire avec un mélange. Il existe des « recommandations » officielles d'emplois des écritures. Elles sont plus ou moins respectées selon les milieux sociolinguistiques. La presse s'y conforme assez strictement. Nous estimons qu'il en va de même pour les brevets. Par contre, les blogs n'ont pas de contraintes mais nous estimons que dans leur immense majorité, les pratiques sont proches de la norme officielle (une raison est que les textes sont saisis au clavier à l'aide de logiciel de transcription qui par défaut appliquent les recommandations officielles).

Dans les lexiques associés aux analyseurs morphologiques automatiques et utilisés ici, les graphies les plus courantes d'un mot font en général l'objet d'une entrée propre. Par exemple le mot *rôka* (« couloir ») est enregistré trois fois dans le lexique UniDic (voir plus bas), avec trois graphies différentes : sinogrammes 廊下, hiragana ろうか, mélange des deux ろう下. D'autres variantes sont possibles, notamment en utilisant les katakana. Ces variantes n'étant pas répertoriées, elles ne seront pas détectées dans les textes par les analyseurs que nous utilisons. Il s'ensuit que le nombre de mots reconnus peut être minimisé. Cependant, les variantes non lexicalisées sont celles qui apparaissent rarement, sauf dans des textes très spécifiques. Leur omission a donc très peu d'impact dans les comptages de textes « ordinaires » comme ceux utilisés pour le présent travail.

Comme on le voit dans l'exemple précédent, le nombre de caractères qui composent un mot peut changer en fonction du choix de l'écriture. C'est ce qui explique qu'à nombre égal de mots, la longueur de deux textes mesurée en nombre de caractères peut varier si ils utilisent des écritures différentes. Sauf pour quelques noms, l'écriture en sinogrammes est systématiquement une des plus courtes.

Les noms communs japonais sont invariables morphologiquement. Seule leur graphie varie. En conséquence, pour cette catégorie, une entrée du lexique correspond à la fois à un mot et à un morphe. Dans les comptages, il serait possible de regrouper les variantes graphiques d'un mot, de sorte à les considérer comme les occurrences d'un seul mot. Par exemple les trois graphies de *rôka* seraient considérées comme des occurrence du mot « couloir ». Nous ne l'avons pas fait car nous ne savons pas quel type d'erreur peut survenir dans la reconnaissance du mot ni l'impact que cela peut avoir sur les comptages, compte tenu des différences entre les lexiques. Nous mesurons donc les occurrences des morphes/entrées lexicales indépendamment les uns des autres. Nous utiliserons désormais le terme de morphe.

Il existe de nombreux homographes. En particulier, phénomène peu souligné, de très nombreux noms communs peuvent être utilisés comme patronymes. Aucune marque graphique ne distingue les deux emplois. Par exemple 森 (*mori*, « forêt ») est un patronyme très répandu. La phrase *mori*$_{N}$ *wo*$_{O}$ *miru*$_{V}$ peut tout aussi bien signifier « Regarder$_{V}$ Mori$_{N}$. » que « Regarder$_{V}$ la forêt$_{N}$. ». Des ambiguïtés surviennent aussi avec les homophones lorsqu'ils sont transcrits phonétiquement, comme par exemple はし (*hashi*), « pont » ou « baguette(s) » (couvert). Avec les trois lexiques de la présente étude, MeCab propose à tort comme première interprétation « baguette » pour l'occurrence de *hasi* dans la phrase *kuruma*$_{N}$ *de*$_{en}$, *hashi wo*$_{O}$ *watatta*$_{V}$ (« En voiture$_{N}$, [j']ai traversé [le] *hashi* »). Des techniques existent pour lever l'ambigüité (Tokunaga, 2012) mais ne sont pas intégrées dans les analyseurs morphologiques. Des erreurs de comptage peuvent s'ensuivre de la confusion des morphes, aussi bien dans le sens d'un plus grand bruit que d'un plus grand silence. L'ajout d'une couche d'analyse, notamment des relations de dépendances, serait possible et utile pour désambiguïser, mais alourdirait considérablement le dispositif tout en risquant d'introduire d'autres erreurs.

A l'exclusion des textes pour enfants, le japonais ne sépare pas graphiquement les « mots » ni les morphes. A cause de leur faible nombre d'occurrences, les signes de ponctuation n'aident que très peu. Quant à l'alternance des écritures, elle est trop aléatoire pour être utilisée massivement. En l'absence de séparation, un morphe peut être détecté à tort à cheval sur deux morphes. Ainsi, dans l'exemple 1 ci-dessous, [*butugo*] (« parole de Buddha ») est détecté à tort à cheval sur les morphes <*niti hutu*> (« franco-japonais ») et <*gogaku gakkai*> (« Société de langues »). Un contrôle sur le contexte des noms permet aisément d'éviter ce type d'erreurs. Une analyse morphologique statistique effectue un tel contrôle. Sagace limitera les erreurs grâce à la stratégie de priorité donnée au morphe le plus long : par préférence du morphe le

plus long, on privilégiera le mot de deux caractères *nitihutu* comme mot de tête, plutôt que celui à un caractère *niti* (« jour »). Cela exclut alors le découpage *niti+hutugo*. La longueur du morphe étant mesurée en nombre de caractères, on comprend que son succès dera tributaire des conventions d'écritures adoptées dans le texte. Un morphe peut aussi être détecté à l'intérieur d'un autre morphe comme dans l'exemple 2 : le morphe *nihonjin* (« Japonais ») « contient » le mot *hi* (« jour »). Dans cet exemple, deux analyses sont possibles et prendre en compte le contexte immédiat ne permet pas de détecter la meilleure analyse. Il faut recourir à d'autres techniques, basées sur les associations sémantiques de « mots », mais qui ne sont pas intégrées aux analyseurs morphologiques. On voit ici que la stratégie de préférence du mot le plus long privilégiera systématiquement l'interprétation « *nihonjin* ». Sagace ne détectera donc pas l'interprétation 1.

Ex 1 : <日 [仏 > < 語] 学 > <学 会>
<*niti* [*hutu*> < *go*] *gaku*> <*gak kai*>
Société franco-japonaise de langues

Ex 2 : 1) その/日 // 本人 が 来た。
sono / hi // hon'nin ga kita
« Ce jour là, la personne elle même est venue. »

2) その/日本人// が 来た 。
sono / nihonjin // ga kita
« Ce Japonais est venu. »

2 Dispositifs d'analyse et ressources utilisés pour la comparaison

Pour comparer les deux approches, avec ou sans analyse morphologique (statistique) préalable du texte, nous avons choisi Sagace et un dispositif basé sur l'analyseur morphologique MeCab. Les deux ont en commun de prendre du texte brut en entrée. Ils sont en ligne de commande, et sont open source et libres. Des chercheurs peuvent ainsi aisément les adapter pour des nouveaux besoins et intégrés à d'autres dispositifs. Dans cette section, nous présentons leurs principales caractéristiques. Comme les performances des dispositifs sont très dépendantes du contenu des lexiques associés, nous consacrerons une sous-section aux lexiques utilisés pour les tests. Aucun des analyseurs utilisés ne comptabilise les variantes graphiques de morphes non répertoriées dans le lexique. Le silence s'en trouve donc augmenté.

2.1 Sagace

Parmi les outils qui ne procèdent pas à une analyse morphologique préalable, ceux qui utilisent les caractères comme unités (`grep` etc.) sont d'un intérêt très limité pour le linguiste. Il existe d'autres outils, qui prennent comme unité de travail les morphes listés dans des lexiques. Nous avons choisi comme représentant de cette approche Sagace car il est disponible avec un lexique au format adéquat et qu'il est utilisé de longue date pour le japonais. D'autres outils peuvent faire des recherches comparables (par exemple Unitex) mais ils auraient nécessité de construire un lexique. Dans la limite de nos investigations, nous n'avons pas trouvé d'outil commercial équivalent.

Sagace est conçu dès l'origine comme un outil d'aide à la recherche en linguistique. C'est un outil générique exploitable pour n'importe quelle langue. Il peut produire des concordances ou lister et compter des collocations. Il se veut plus adapté au linguiste en permettant une manipulation très rapide et souple des lexiques grâce à un langage (de type propositionnel) de description des catégories avec lequel il est possible de créer « à la volée » des nouvelles catégories, sur la base des catégories existantes dans le lexique, et sans intervenir dans celui-ci. Autre aspect pensé pour le linguiste : il est tout en un. Il contient une interface d'interrogation, un moteur de recherches et de comptage, et un compilateur de lexique. Il ne nécessite pas de logiciels externes (contrairement aux dispositifs avec analyse morphologique préalable), ni de segmentation préalable (comme c'est le cas de Himawari (Yamaguchi, 2011)). Mais surtout, il ne nécessite pas non plus d'entraînement comme les logiciels statistiques.

Lors de l'analyse, Sagace découpe le texte en « segments », délimités par un morphe *ad hoc* choisi par l'utilisateur. Pour la présente comparaison, nous prenons comme segment la phrase. Les segments dépassant une certaine longueur sont rejetés pour éviter des problèmes de mémoire. Les occurrences de patrons recherchés qui sont présents dans les segments rejetés ne sont donc pas pris en compte par Sagace alors qu'ils seront traités par les autres dispositifs. De ce fait, le silence dans les résultats de Sagace pourrait être supérieur. Lorsque sur une position Sagace détecte plusieurs morphes possibles, il privilégie le morphe le plus long. Cette méthode usuelle en analyse morphologique du japonais est souvent gagnante, mais peut provoquer des erreurs.

2.2 Dispositif basé sur MeCab

Les dispositifs qui procèdent à une analyse morphologique préalable (segmentation et étiquetage des parties du discours) sont des assemblages composés d'un analyseur morphologique et d'un outil de comptage qui fait aussi interface pour l'interrogation. A notre connaissance, il n'existe pas de dispositif « tout en un ». Pour des questions pratiques, nous ne pouvions pas ici évaluer les performances de toutes les combinaisons possibles de tous les analyseurs morphologiques et outils de comptage. Pour des raisons de disponibilité et de réputation, nous avons choisi l'analyseur morphologique statistique MeCab, parmi les plus utilisés du moment JUMAN (Kurohashi et al., 1994), KyTea (Neubig et al., 2011). Il est disponible sous forme de paquets pour les systèmes de type Unix. Le logiciel est en plus distribué avec plusieurs lexiques japonais libres eux aussi (voir section suivante). Pour simplifier la présentation, désormais, nous désignerons la combinaison de MeCab, d'un lexique et d'un outil de comptage par le terme « dispositif MeCab ». MeCab utilise un modèle de Markov d'ordre 1. Les paramètres du modèle, inscrits dans le lexique, sont estimés par apprentissage. MeCab prend comme segment de travail toute chaîne de caractères finissant par un saut à la ligne. Dans les textes bruts il arrive que des sauts à la ligne apparaissent à l'intérieur d'une phrase. Les analyses peuvent être perturbées. Nous estimons que l'erreur induite augmente le silence mais pas au point de modifier sensiblement les résultats.

2.3 Les lexiques

Les performances des dispositifs d'analyse varient en fonction des lexiques. C'est la raison pour laquelle nous avons mené les tests avec les lexiques disponibles pour MeCab. UniDic[1] contient 756 463 entrées lexicales, noms propres compris, ce qui en fait le lexique le plus volumineux parmi les trois utilisés ici. Contrairement aux deux autres lexiques, le choix des entrées lexicales/morphes a été argumenté et explicité (The UniDic Consortium, 2012). Le principe est de lexicaliser les plus petites unités morphosémantiques possibles. A coté de cela, les concepteurs et mainteneurs de MeCab recommandent[2] un second lexique, IPAdic[3] (392 126 entrées). Enfin, il faut compter avec un troisième lexique, Jumandic[4] (515 996 entrées). Dans ces trois lexiques figurent diverses informations : lecture, écriture usuelle du morphe si il s'agit d'une variante graphique, et partie du discours. Figurent aussi les paramètres du modèle statistique utilisé. Les paramètres ont été estimés par apprentissage sur des corpus différents, respectivement le BCCWJ data core (voir section suivante), le corpus IPA (Hashimoto, 1995) et le corpus de Kyodai (Kawahara et al., 2002). Avec Sagace, nous avons utilisé une version modifiée du naist-jdic[5] (485 000 entrées), qui est très proche du lexique IPAdic. Nous estimons que les lexiques possèdent un nombre de noms communs à peu près identique, soit environ 90 000 si l'on se base sur le naist-jdic.

2.4 Corpus

Les tests ont été effectués sur le Balanced Corpus of Written Japanese - Data Core, version 2009 (Maekawa, 2009). Il s'agit d'un corpus segmenté manuellement en morphes et étiqueté (au format XML) en reprenant les morphes et parties du discours du lexique UniDic 1.3.12 . Nous en avons tiré un texte brut sans les balises. Ce corpus s'est imposé ici car il est devenu aujourd'hui une référence en linguistique. Il est aussi utilisé en traitement automatique (voir des références dans (Mori et al., 2014)). D'autres corpus, dont celui de Kyodai ou YACIS (Ptaszynski et al., 2012), sont disponibles mais leur usage reste limité au traitement automatique. Le corpus comprend des extraits de blogs, de journaux économiques, de brevets et des phrases exemples de dictionnaires. Nous estimons que les conventions d'écriture standard sont globalement respectées. En particulier, la majorité des noms communs sont transcrits en sinogrammes lorsqu'ils peuvent l'être. Le recours à la transcription phonétique et les risques d'ambiguïtés dus à l'homophonie en sont diminués. Le corpus brut ne contient pas de phrases d'une longueur excédant les limites imposées par Sagace. Aucune phrase n'est donc rejetée par ce dernier. Certaines phrases contiennent des sauts de lignes. Cela peut affecter l'analyse de MeCab qui n'a pas été entrainé sur des corpus contenant de tels sauts de lignes.

---

1 unidic-mecab 2.1.2

2 mecab.googlecode.com/svn/trunk/mecab/doc/index.html

3 mecab-ipadic.x86_64 ; 2.7.0.20070801-6.fc18.1 .

4 mecab-jumandic.x86_64, ref ; 5.1.20070304-7.fc18 .

5 mecab-naist-jdic-0.4.3-20080812

3 Taux d'accord

Nous comparons les taux d'accord entre les analyses automatiques et l'analyse manuelle dans une tâche d'extraction d'un morphe isolé (nom commun) et dans une tâche d'extraction d'une chaîne de trois morphes (un nom commun et son contexte gauche et droite). L'extraction et le comptage sont effectués avec les lexiques en l'état, donc avec les 90 000 noms communs. Mais la comparaison des résultats n'est effectuée que sur les noms communs présents à la fois dans les trois lexiques et dans le texte segmenté manuellement. Les erreurs sur les autres noms ne sont donc pas prises en compte, ce qui pourrait minimiser artificiellement le bruit. Il en va de même avec les signes de ponctuation et les particules qui sont utilisées pour définir le contexte du nom dans le deuxième test.

Nous comptons tout d'abord les occurrences de 4 000 noms communs, sans poser de contraintes sur le contexte. Les résultats (table 1) sont conformes aux attentes. Mecab est meilleur, avec bien sûr un avantage pour l'analyse avec le lexique Unidic. Les écarts types montrent que les variations sont plus importantes avec Sagace. Nous faisons l'hypothèse que Sagace fait sensiblement plus d'erreurs sur les morphes courts et rédigés en caractères phonétiques que sur les morphes longs, en sinogrammes. La différence entre Sagace et les dispositifs MeCab, si elle n'est pas négligeable, n'est cependant pas considérable, en particulier par rapport aux deux dispositifs MeCab qui n'ont pas été entraînés sur le corpus étudié et dont les résultats sont de ce fait plus représentatifs des performances réelles des dispositifs.

	RAPPEL		PRECISION		F-MESURE
	moyenne	écart type	moyenne	écart type	
Sagace	89.34	.24	83.86	.39	.865
MeCab+ IPAdic	96.85	0.11	96.03	0.13	.964
MeCab + Jumandic	95.52	0.14	97.37	0.11	.964
MeCab + UniDic	99.09	0.06	98.78	0.08	.989

TABLE 1 : Taux d'accord des analyses de Sagace et MeCab par rapport à une analyse manuelle basée sur le lexique UniDic ; comptage des noms communs, sans contrainte sur le contexte immédiat.

Nous comparons cette fois-ci les taux d'accord dans la tâche d'extraction des noms communs mais en posant des contraintes sur leur contexte immédiat. Nous imposons la présence de particules ou signes de ponctuation (44 morphes et symboles en tout) de chaque coté du nom commun.

	RAPPEL		PRECISION		F-MESURE	
	moyenne	écart type	moyenne	écart type		Diff. avec Table 1
Sagace	85.89	.28	89,28	.31	.875	+ 0.010
MeCab+ IPAdic	83.98	.31	96.06	.17	.896	- 0.068
MeCab + Jumandic	95.07	.20	89.60	.26	.922	- 0.041
MeCab + UniDic	88.42	.26	99.19	.08	.934	- 0.054

TABLE 2 : Taux d'accord des analyses de Sagace et MeCab par rapport à une analyse manuelle basée sur le lexique UniDic ; comptage des noms communs, avec contraintes sur le contexte immédiat.

Au regard des f-mesures (Table 2) l'écart de performances entre Sagace et les dispositifs MeCab s'est réduit, mais du fait de mouvements inverses : Sagace est meilleur que pour une recherche de morphe isolé tandis que les dispositifs MeCab sont moins performants. Pour Sagace, l'amélioration est due à l'allongement du patron cherché, et à la réduction des risques d'ambiguïtés qui s'ensuit. Sagace rejoint même MeCab sur certains résultats. Pour MeCab, la

baisse est plus sensible avec le lexique IPAdic pourtant conseillé par les auteurs du logiciel. Une explication est peut-être dans le choix des parties du discours ou du corpus d'entraînement. Une étude plus approfondie reste à faire.

4 Conclusion

Dans cette étude, nous avons comparé les performances de deux dispositifs d'analyse de corpus à destination des linguistes du japonais, pour l'exécution de tâches tout à fait usuelle en linguistique : le dénombrage de mots ou de morphes. Ces dispositifs ont été choisis comme représentatifs de deux types de traitement des corpus : une approche « légère » qui s'en tient à la recherche de patrons sans analyse générale des phrases, et une approche complète, plus lourde avec analyse morphologique du texte analysé. Les résultats montrent que, à condition d'éviter certains types de recherches, les performances du premier dispositif sont raisonnables. Si l'on tient en plus compte de l'ergonomie et en particulier la facilité de mise en oeuvre et la vitesse de traitement, il s'agit d'un outil tout à fait suffisant pour le débroussaillage de corpus et l'aide à l'analyse linguistique.

Références

BLIN R. (2012). SAGACE v4.2.0.

HASHIMOTO M. (1995). Building a Corpus at IPA. 全国大会講演論文集 *51*, 35–36.

KAWAHARA D., KUROHASHI S., and HASIDA K. (2002). Construction of a Japanese Relevance-tagged Corpus. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, pp. 2008–2013.

KUDO T. (2006). MeCab: yet another part-of-speech and morphological analyzer.

KUROHASHI S., NAKAMURA T., MATSUMOTO Y., and NAGAO M. (1994). Improvements of Japanese Morphological Analyzer JUMAN. In Proceedings of The International Workshop on Sharable Natural Language Resources, pp. 22–28.

MAEKAWA K. (2009). Daiyousei wo yû suru daikibo nihongo kakikotoba kôpasu (<tokushyû> nihongo kôpasu) [Compilation d'un corpus équilibré de textes contemporains en japonais. J. Jpn. Soc. Artif. Intell. *24*, 612–622.

MORI S., OGURA H., and SASADA T. (2014). A Japanese Word Dependency Corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), (Reykjavik, Iceland: European Language Resources Association (ELRA)), pp. 753–758.

NEUBIG G., NAKATA Y., and MORI S. (2011). Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 529–533.

PTASZYNSKI M., RZEPKA R., ARAKI K., and MOMOUCHI Y. (2012). Automatically Annotating a Five-billion-word Corpus of Japanese Blogs for Affect and Sentiment Analysis. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 89–98.

THE UNIDIC CONSORTIUM (2012). unidic-mecab ver. 2.1.1 Users Manual.

TOKUNAGA T. (2012). Nihongo nyuuryoku wo sasaeru gijutu [Technologies behind Japanese input systems].

YAMAGUCHI M. (2011). zenbun kensaku sisutemu "Himawari" riyousha manyuaru vers.1.3 [Système d'extraction "Himawari"; Manuel de l'utilisateur vers.1.3].

YAZAWA M. (2006). Traitement de texte en japonais et enseignement des kanji. In Langue, Lecture et École Au Japon, (Arles: Christian Galan et Jacques Fijalkow), pp. 97–137.