Des questions linguistiques soulevées par les résultats d'alignement des mots katakana

Yayoi Nakamura-Delloye

Des questions linguistiques soulevées par les résultats d'alignement des mots *katakana*

Yayoi Nakamura-Delloye

Laboratoire LeSCLaP – Université de Picardie

Résumé : cette communication présente les réflexions que nous avons eues lors de l'évaluation des résultats de l'alignement des mots *katakana*. Dans le cadre de nos travaux antérieurs sur le développement d'un système d'alignement des phrases (Nakamura-Delloye 2005), nous avons conçu une méthode d'alignement de ces mots en *katakana* basée sur leur retranscription en alphabet latin. Afin d'approfondir nos études, nous avons réalisé une évaluation plus complète de cette technique. Mais lors de l'évaluation, nous avons été confrontés à des questions linguistiques remettant en cause l'évaluation même des résultats. En effet, bien que dans certains cas la justesse ou la fausseté de la segmentation soit évidente, dans d'autres cas nous n'arrivons pas à trancher.

1. Introduction

Le présent article présente les réflexions que nous avons eues lors de l'évaluation des résultats de l'alignement des mots *katakana*. Le syllabaire *katakana* est l'un des trois principaux systèmes d'écriture utilisés dans les textes japonais, qui est généralement employé pour les mots emprunts qu'il sert à transcrire phonétiquement. Dans le cadre de nos travaux antérieurs sur le développement d'un système d'alignement des phrases (Nakamura-Delloye 2005), nous avons conçu une méthode d'alignement de ces mots en *katakana* basée sur leur retranscription en alphabet latin. L'efficacité de ce type de méthode était déjà connue (Collier *et al*. 1997) (Brill *et al*. 2001) (Tsuji 2001) (Tsuji *et al.* 2002) mais nos travaux différaient des travaux connexes notamment par la non-utilisation d'analyseur morphologique et par la retranscription par transducteur. Afin d'approfondir nos études, nous avons réalisé une évaluation plus complète de cette technique, et nous avons alors été confrontés à des questions linguistiques remettant en cause l'évaluation même des résultats.
L'article décrit d'abord une méthode de segmentation des mots composés en *katakana* (§ 2) ainsi qu'une méthode de mise en correspondance de ces mots avec leur traduction (§ 3). Nous nous intéresserons ensuite au résultat d'expérience — corpus, évaluation et analyse —, et surtout aux questions rencontrées lors de cette évaluation (§ 4).

2. Segmentation des mots composés en *katakana* à l'aide d'un arbre lexicographique

Comme le japonais ne possède pas de signes permettant de segmenter les phrases *a priori*, un système d'analyse morphologique est généralement utilisé, système dont l'objectif est de transformer, à l'aide d'un dictionnaire et de probabilité des combinaisons, la phrase en une suite d'unités appartenant chacune à une catégorie morpho-syntaxique.

2.1. Segmentation par analyseur morphologique

L'analyse morphologique commence par la consultation d'un dictionnaire afin de trouver toutes les séquences qui peuvent constituer une unité morphologique. Par exemple, pour une séquence « *zen-koku-to-dô-fu-ken-gi-chô-kai* » (cf. Fig. 1), on peut trouver 18 candidats d'unités morphologiques composantes en consultant un dictionnaire. Ces candidats donnent ensuite lieu à différentes combinaisons possibles présentées dans la figure 1. L'opération consiste ensuite à déterminer la combinaison la plus adéquate de ces candidats, généralement avec une méthode probabiliste, pour trouver la segmentation de la séquence. Mais cette méthode a comme inconvénient d'être dépendante de l'existence et de la qualité des dictionnaires.

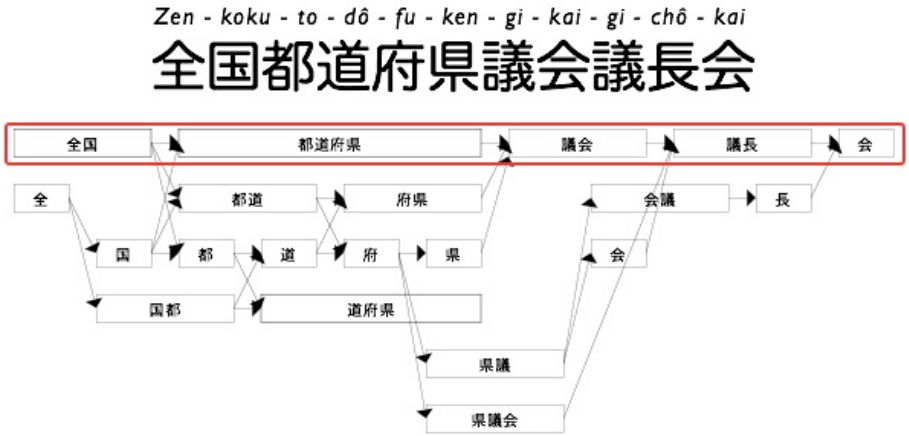

Fig. 1: Possibilités de combinaison de tous les candidats

Toutefois, il existe également une méthode classique d'analyse morphologique partielle permettant d'extraire les éléments lexicaux sans aucune connaissance extérieure, appelée segmentation par type de caractère, qui met à profit la principale particularité de l'écriture japonaise utilisant plus de trois systèmes d'écriture.

2.2.Segmentation par type de caractère et ses problèmes

En japonais, plusieurs types de caractères sont utilisés selon la nature des mots. Il existe trois principaux systèmes d'écriture : *kanji*, *hiragana* et *katakana* (cf. Fig. 2). Les *kanji* sont des idéogrammes qui sont utilisés pour représenter les mots pleins et les radicaux qui ont un sens. Les *hiragana* sont un des deux syllabaires japonais et ils sont souvent utilisés pour représenter les mots grammaticaux (notamment les particules) et la partie variable des mots variables (verbes, adjectifs). Les *katakana* sont l'autre syllabaire japonais et comme je l'ai déjà évoqué dans l'introduction, ils sont généralement employés pour transcrire phonétiquement les mots emprunts.

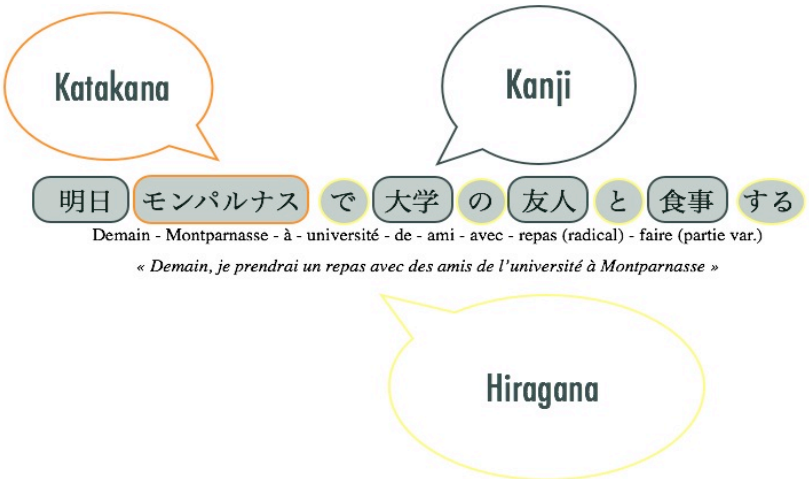

Fig. 2: Les trois systèmes d'écriture japonais

Cette méthode de segmentation possède toutefois deux grands problèmes : impossibilité de segmentation des séquences de mots composés constituées de plusieurs substantifs du même type de caractère, juxtaposés les uns derrière les autres (cf. Fig. 3 : deuxième exemple « *rin-ji-koku-kai-hei-kai* » et troisième exemple « *yu-u-za-i-n-ta-a-fu-e-i-su* ») ; segmentation fausse des mots constitués de différents types de caractères (cf. Fig. 3 : premier exemple « *i-ki* »).
Dans le cadre du développement d'un système d'alignement des phrases traitant les textes japonais sans analyseur ni dictionnaire, nous avons conçu une amélioration de la segmentation par type de caractère en résolvant le premier type de problème, à savoir la segmentation des séquences où la frontière entre les deux mots n'est pas marquée par un changement de type de caractère.

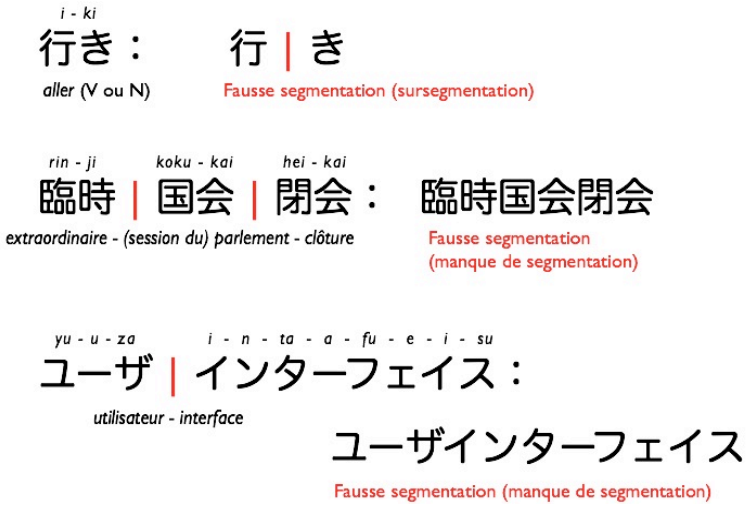

Fig. 3 : Problèmes de la méthode de segmentation par types de caractère

2.3. Amélioration de la segmentation par type de caractère

Cette amélioration est inspirée par la méthode d'analyse morphologique partielle proposée par (Kay & Röscheisen, 1993). Elle consiste à trouver les sous-chaînes préfixales ou suffixales communes à plusieurs formes effectives des mots graphiques et à déterminer les radicaux, porteurs de sens. Il s'agit donc de la recherche des sous-chaînes préfixales communes à plusieurs formes effectives. Par exemple, à partir de trois séquences « engagé », « engager » et « engageons », nous pouvons obtenir la sous-chaîne préfixale commune « engag ».

Nous avons adapté cette méthode au traitement des mots composés en *kanji* et en *katakana*. En effet, trouver les frontières de mots dans une séquence constituée d'un même type de caractère est également une recherche des sous-chaînes communes à plusieurs formes effectives. Par exemple, à partir de trois séquences « *shoku-ryô-kyô-kyû* », « *shoku-ryô-bu-soku* », « *shoku-ryô-ki-ki* » de l'exemple présenté dans la figure 4, nous pouvons obtenir la sous-chaîne préfixale commune « *shoku-ryô* ». L'adaptation de cette approche au traitement des textes japonais a nécessité quelques modifications, notamment le traitement des parties restantes considérées comme des suffixes, qui doivent, dans le cas du japonais, être conservées en tant qu'unités lexicales autonomes. On obtient donc à partir d'un mot graphique *a b*, non pas uniquement une forme de base *a*, mais deux formes de base *a* et *b*. Dans le cas de l'exemple cité précédemment, on obtient à partir des trois séquences, non pas un lemme, mais quatre lemmes « *shoku-ryô* », « *kyô-kyû* », « *bu-soku* » et « *ki-ki* ».

Nous avons également appliqué la même méthode à la segmentation des séquences en *katakana*. Mais, dans le cas des séquences de *katakana*, nous ne cherchons pas de sous-chaînes communes à plusieurs séquences de *katakana*, mais des sous-chaînes semblables à une autre séquence ou à un lemme extrait d'une séquence plus grande. Autrement dit, lorsqu'on a une séquence « *ba-i-po-o-ra-a-to-ra-n-ji-su-ta-a* », c'est uniquement quand on trouve « *to-ra-n-ji-su-ta-a* » utilisé seul qu'elle est segmentée en deux mots (cf. Fig. 5, exemple du haut). Ainsi, on empêche la segmentation des séquences telles que « *i-n-su-to-o-ru* » et « *i-n-to-ro-da-ku-shi-yo-n* » en deux parties, même si la séquence « *i-n* » est la sous-chaîne commune de ces deux formes (cf. Fig. 5, exemple de bas).

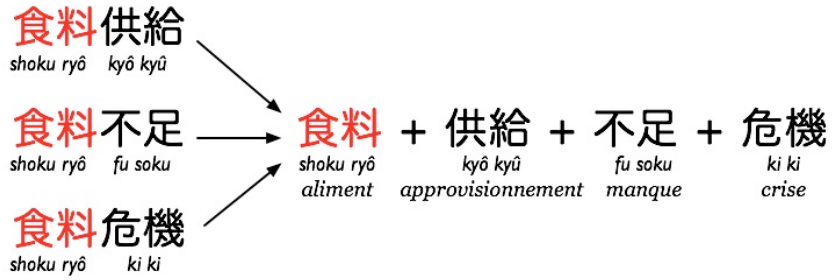

Fig. 4 : Segmentation par recherche des sous-chaînes communes

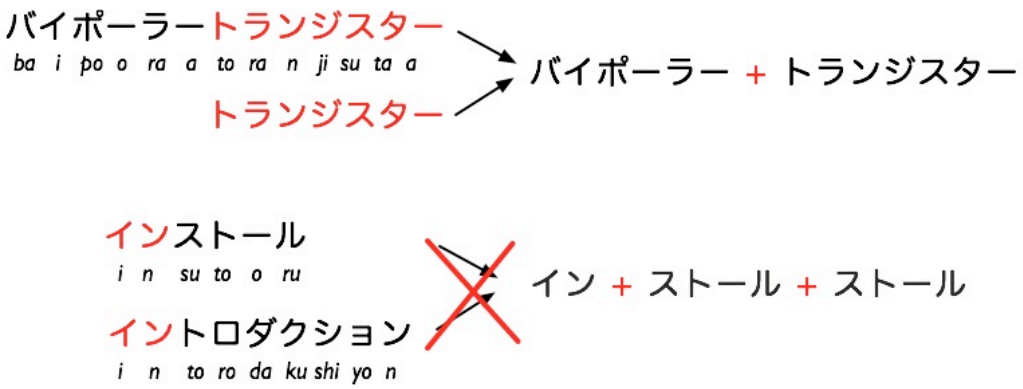

Fig. 5 : Segmentation de séquences de katakana

Cette méthode peut être implémentée efficacement à l'aide de la structure de données appelée arbre lexicographique. La figure 6 représente un exemple d'arbres qui vérifient des chaînes préfixales et suffixales. L'arbre gauche est construit avec des séquences normales et il est destiné à trouver les sous-chaînes préfixales. L'arbre droit est construit avec des séquences inversées et il est destiné à trouver les sous-chaînes suffixales.

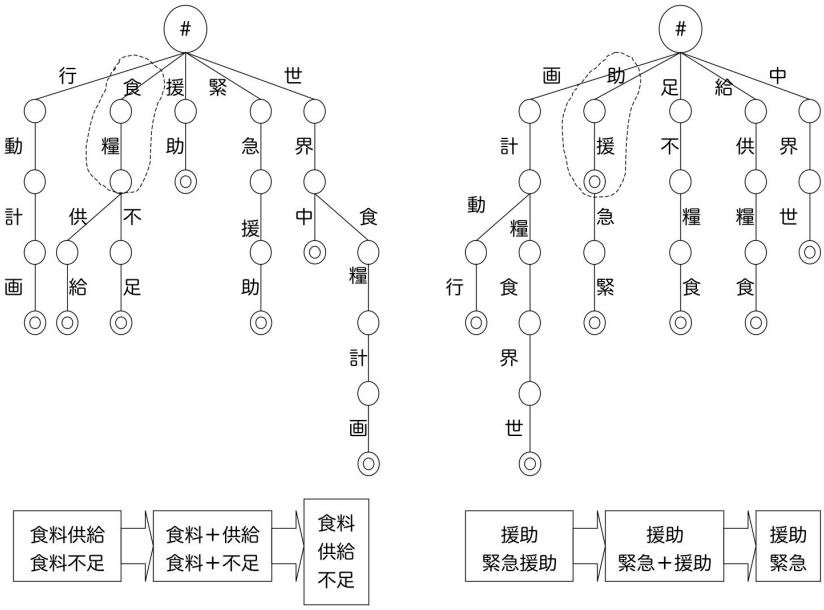

Fig. 6 : Recherche des sous-chaînes communes avec des arbres lexicographiques

3. Alignement des mots *katakana* par la retranscription avec transducteur

Nous nous intéressons à présent à la méthode dans laquelle les mots katakana ainsi extraits sont alignés avec leur mot d'origine à partir d'un corpus parallèle non aligné. Leur alignement consiste en leur retranscription en alphabet latin à l'aide d'un transducteur spécifiquement conçu.

La figure 7 représente la procédure d'appariement d'un mot en *katakana*. Les mots en *katakana* extraits avec la méthode de segmentation constituent la liste des mots *katakana* (cf. étape 1 Fig. 7). Notre transducteur retranscrit chaque mot en *katakana* en plusieurs formes en alphabet latin (cf. étape 2 Fig. 7). Avec l'autre corpus, on constitue également une liste des mots (cf. étape 3 Fig. 7). Chaque mot de cette liste est comparé avec des retranscriptions des mots *katakana* (cf. étape 4 Fig. 7) et si la similarité entre le mot considéré et une séquence de retranscription d'un mot en *katakana* atteint un seuil prédéfini, ce mot est stocké dans la liste des candidats (cf. étape 5 Fig. 7). Une fois l'examen des mots terminé, on cherche, parmi les candidats extraits, le mot d'origine le

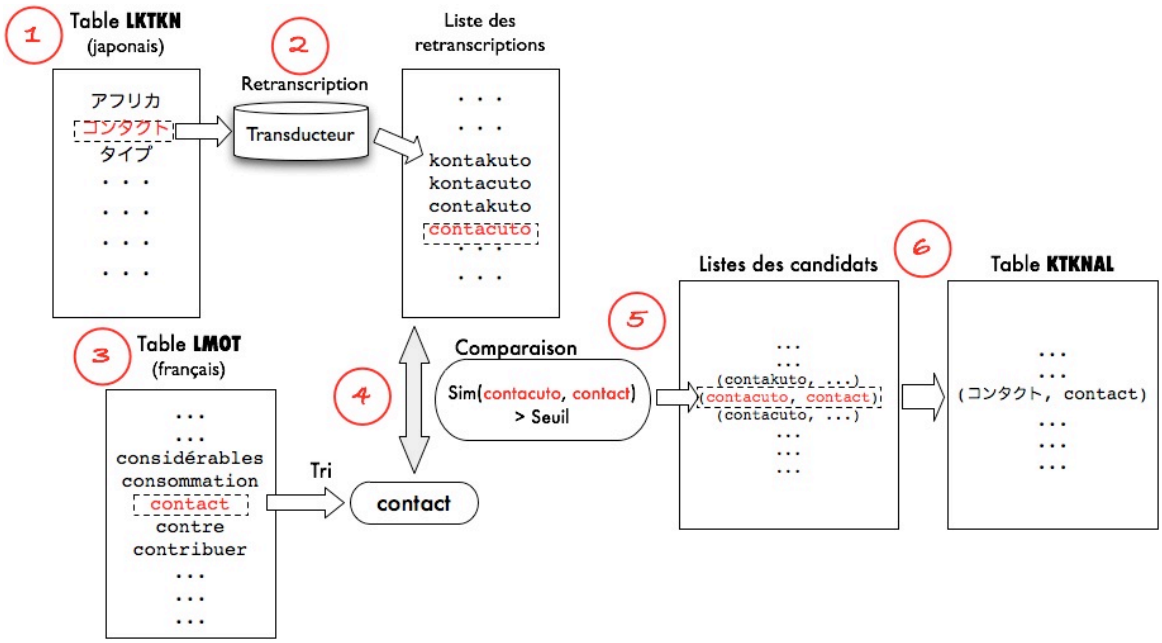

Fig. 7 : Procédure d'alignement des mots katakana

plus probable pour chaque mot japonais en *katakana*, et on constitue la liste des mots en *katakana* alignés (cf. étape 6 Fig. 7).

Afin de permettre la retranscription, une grammaire a d'abord été définie. Cette grammaire a ensuite été transformée en un ensemble de règles de transition et de sortie pour constituer un transducteur de retranscription. Les règles de transition indiquent la transition d'un état à un autre provoquée par chaque symbole d'entrée. Les règles de sortie indiquent un ou plusieurs symboles de sortie liés à chaque état. À l'aide de ce transducteur, un mot *katakana* est retranscrit en plusieurs formes possibles.

Ces formes retranscrites sont ensuite comparées avec tous les mots lexicaux du texte de l'autre langue pour calculer leur similarité de forme. La similarité doit être calculée tout en tenant compte d'éventuelles divergences de forme dues à la différence de système phonétique / phonologique entre le japonais et l'autre langue. À cet effet, nous recourons aux méthodes de calcul utilisées pour la mise en correspondance des cognats largement étudiées dans le cadre de travaux sur l'alignement entre les textes de langues apparentées. Notre formule est inspirée notamment de celle de la sous-chaîne maximale parallèle proposée par Kraif (2001). Toutefois, du fait des besoins particuliers de la retranscription des *katakana*, elle diffère de cette dernière par le fait qu'elle tient compte non seulement de la sous-chaîne maximale mais aussi des consonnes communes. Le nombre de consonnes communes est pris en compte pour favoriser les deux chaînes ayant le plus de caractères consonantiques communs plutôt que celles dont les caractères vocaliques coïncident le plus.

À cette similarité, nous combinons également la similarité de distribution afin d'exclure les correspondances hasardeuses et déterminons pour chaque mot *katakana* le mot de l'autre texte ayant la similarité la plus élevée et supposé en relation de traduction.

4. Expérience : corpus, évaluation, analyse

Afin d'approfondir nos études antérieures, nous avons réalisé à nouveau une évaluation plus complète de ces techniques de segmentation et de mise en correspondance des mots *katakana*.

4.1. Corpus

Nous avons utilisé des manuels (anglais/français/japonais) de produits électroniques de la société Apple, disponibles sur le site de la société. Pour l'alignement français-japonais, nous avons utilisé en plus un corpus constitué d'articles du Monde Diplomatique (LMD ci-dessous). Ces corpus sont caractérisés par les occurrences importantes des *katakana*. À titre d'exemple (cf. Tab. 1), l'extrait du

	Roman	Journal	LMD	Apple
Nb de mots	20 150	22 367	24 343	63 638
Nb de mots différents (A)	2 857	3 052	4 290	2 272
Nb de mots katakana (B)	201	120	529	384
B / A	**7 %**	**4 %**	**12 %**	**17 %**

Tab. 0 : Caractéristiques des corpus

roman japonais « *La fin des temps* » de Haruki Murakami (Shinchosha, 1985) et une série d'articles du quotidien japonais « *Yomiuri* » contenaient respectivement 7 et 4% de mots *katakana* dans l'ensemble des mots utilisés, tandis que les articles traduits du Monde Diplomatique en comportaient 12%, et les manuels Apple, eux, 17%.
Le corpus Apple — comme tous les textes de ce domaine — est caractérisé par l'importance des mots emprunts, souvent des néologismes, appartenant notamment au vocabulaire informatique. En revanche, le corpus LMD contient plus de noms propres transcrits par ce syllabaire.

4.2. Résultat : segmentation des séquences en katakana

Les corpus ont d'abord été segmentés en mots par deux méthodes différentes : avec notre méthode de segmentation et par un analyseur morphologique du japonais ChaSen, analyseur morphologique utilisé très largement au Japon et développé par le Nara Institute of Science and Technology.
ChaSen a tendance à segmenter plus, fournissant 10% de mots en plus. En effet, ChaSen découpe très souvent les mots absents dans son dictionnaire en sous-chaînes qui y figurent. Ainsi beaucoup de noms propres et de néologismes ont été sursegmentés de manière erronée.
Dans le résultat de notre méthode de segmentation, il y a beaucoup moins de sursegmentations erronées. En revanche, on constate plus l'absence de segmentation. Par exemple, ChaSen a segmenté les séquences telles que « *su-te-e-ta-su-ra-m-pu* » (status light) et « *ka-a-ki-t-to* » (car kit) en deux mots alors que notre méthode n'a effectué aucun découpage.
Il y a également des cas inverses. Par exemple, les séquences telles que « *pu-re-i-ri-su-to* » (play list) et « *de-e-ta-ro-o-mi-n-gu* » (data roaming) ont été segmentées en deux mots par notre méthode, mais pas par l'analyseur ChaSen.
Néanmoins, notre plus grand problème résidait dans l'évaluation de ce résultat de segmentation. En effet, bien que dans certains cas la justesse ou la fausseté de la segmentation soit évidente, dans d'autres cas nous n'arrivons pas à trancher : « *su-ra-i-do-sho-o* » (slideshow) n'est-il qu'un seul mot ? Et « *sa-i-n-a-p-pu* » (sign up) ou « *ba-k-ku-a-p-pu* » (backup/back up) ? Le figement d'un mot est difficile à déterminer d'autant plus sans doute lorsqu'il s'agit d'un mot emprunt d'une autre langue.
Ainsi nous sommes retombés dans la question classique de la définition de mot. Nous avons donc décidé de ne pas évaluer à cette étape et de regarder l'influence de ces segmentations différentes dans le résultat de l'alignement.

	ANG / JP		FR / JP	
	Rappel	Précision	Rappel	Précision
Avec analyseur	0,49	0,76	0,23	0,41
Sans analyseur	0,55	0,82	0,26	0,46

Tab. 2 : Résultat d'alignement des mots katakana

Mots d'origine	Segmentation par analyseur	Segmentation par recherche des sous-chaînes
Car kit	☺ カー | キット	カーキット
Broadband	ブロード | バンド	☺ ブロードバンド
Data roaming	データローミング	☺ データ | ローミング
Playlist	☺ プレイリスト	プレイ | リスト

☺ = segmentation ayant permis l'alignement des mots

Tab. 3 : Exemples de résultat de segmentation des séquences en katakana

4.1. Résultat : appariement des mots katakana

Le nombre de paires correctement alignées est *grosso modo* semblable pour les deux méthodes (cf. Tab. 2). En effet, la segmentation (ou la non segmentation) peut à la fois favoriser et défavoriser l'alignement. Dans certains cas, la segmentation en plus petites unités favorise l'alignement, et dans d'autres elle l'empêche. Par exemple (cf. Tab. 3), « *ka-a-ki-t-to* » (car kit) et « *bu-ro-o-do-ba-n-do* » (broadband) sont segmentés en deux mots par l'analyseur, mais la segmentation a empêché l'alignement du deuxième exemple contrairement au premier qui a été aligné avec les deux mots anglais correspondants. « *pu-re-i-ri-su-to* » (playlist) et « *de-e-ta-ro-o-mi-n-gu* » (data roaming) sont segmentés en deux mots par notre méthode et non par l'analyseur. Là encore, la segmentation de « *pu-re-i-ri-su-to* » a empêché l'alignement alors que la segmentation en « *de-e-ta* » et « *ro-o-mi-n-gu* » a permis l'alignement avec le mot anglais correspondant.
La conséquence intéressante de ce résultat disparate est que malgré leur résultat comparable, chaque méthode contient plus de 10% de paires qui ne sont pas alignées par l'autre méthode.
Pour l'alignement également, nous avons comparé deux méthodes, l'une avec la similarité calculée par la similarité des formes et des distributions et l'autre sans prise en compte des distributions. Dans des travaux antérieurs de mise en correspondance des mots *katakana*, la distribution n'est pas prise en compte, mais elle améliore sensiblement la précision. Dans le cas de l'alignement anglais-japonais, nous avons obtenu une précision de 76-82% avec un taux de rappel de 50-55%. Avec la prise en compte des distributions, ce taux de précision remonte à 96-98%, mais le rappel tombe à moins de 25%. Le résultat d'alignement français-japonais du corpus Apple est décevant avec une précision de 45% et un taux de rappel de 25%. Mais avec la prise en compte des distributions, la précision remonte jusqu'à 88% avec un rappel de 16%. Le résultat d'alignement du corpus LMD est tout de même encourageant avec une précision de 93% et un taux de rappel de 34%.

5. Conclusion et perspectives

Le résultat de segmentation n'a pas montré l'avantage d'une des deux méthodes comparée à l'autre. La segmentation (ou la non-segmentation) peut à la fois favoriser et défavoriser l'alignement. Cette expérience nous a amené surtout à remettre en cause la définition même de mot, que nous avons choisi comme unité d'alignement. Nous avons donc mis de côté l'évaluation à cette étape pour regarder plutôt l'influence de ces segmentations sur le résultat de l'alignement.
Le résultat d'alignement n'a pas non plus montré l'avantage d'une de ces deux méthodes de segmentation. La segmentation (ou la non segmentation) peut à la fois favoriser et défavoriser l'alignement. La solution serait peut-être de combiner les deux méthodes de segmentation pour profiter entièrement des paires de mots alignés par au moins une des deux méthodes.
Quant à l'opération d'alignement, notre résultat n'est pas encore totalement satisfaisant. En effet,

notre grammaire de retranscription définie manuellement est incomplète. En revanche, les travaux existants (Tsuji 2001) (Tsuji *et al.* 2002) proposent une méthode pour définir automatiquement une grammaire à partir de ressources terminologiques bilingues. Sur le modèle de ces travaux et par la prise en compte de la probabilité de chaque transcription, nous pourrions certainement améliorer nos résultats. Par ailleurs, l'exploitation de corpus comparables, présentant différents avantages par rapport aux corpus parallèles — notamment leur importance et leur diversité —, a attiré depuis toujours les chercheurs travaillant dans le domaine de la constitution automatique de ressources multilingues et plusieurs travaux portant sur l'extraction de terminologies bilingues ont été présentés, et ce non seulement avec des approches pour les langues parentes (Dejean et Gaussier 2002) (Morin et Daille 2004), mais aussi celles adaptées pour le japonais (Kageura *et al.* 2000). L'intérêt de l'utilisation de notre méthode d'alignement des mots *katakana* dans ce type de travaux exploitant des corpus comparables est également incontestable.

Références

Brill E., Kacmarcik G. et Brockett C. (2001). « Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs », Asia Federation of Natural Language Processing. p. 393-399.

Collier N., Hirakawa H. et Kumano A. (1997). « Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using KATAKANA matching », Proceedings of the Natural Language Processing Pacific Rim Symposium 1997. p. 309-314.

Dejean H. et Gaussier E. (2002) « Une nouvelle approche à l'extraction de lexique bilingues à partir de corpus comparables », dans : Teubert W. et Krishnamurthy R. (ed.), Corpus Linguistics : Critical Concepts in Linguistics, Routledge Publishers.

Morin E. et Daille B. (2004). « Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé », Vol. 45/3. p.103-122

Kageura K., Tsuji K. et Aizawa N. (2000). « Automatic Thesaurus Generation through Multiple Filtering », Proceedings of the 18th Conference on Computational linguistics, Vol. 1. p. 397-403.

Kay M. et Röscheisen M. (1993). « Text-translation alignment », Computational Linguistics, 19 (1). p. 121–142.

Kraif O. (2001). « Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation », TAL, 42 (3).

Nakamura-Delloye Y. (2005). « Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais », Actes de la conférence TALN/RECITAL 2005. p. 585-594.

Tsuji K. (2001). « Automatic extraction of translational japanese-katakana and english word pairs from bilingual corpora », Proceedings of the 19th International Conference on Computer Processing of Oriental Langages (ICCPOL'01). p. 245–250.

Tsuji K. *et al*. (2002). « Extracting french-japanese word pairs from bilingual corpora based on translitération rules », Proceedings of the third International Conference on Language Ressources and Evaluation (LREC 2002).