Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais

Yayoi Nakamura-Delloye

RECITAL 2005, Dourdan, 6–10 juin 2005

Système AlALeR

Alignement au niveau phrastique des textes parallèles français-japonais

Yayoi NAKAMURA-DELLOYE
Université Paris 7 (École Doctorale de Sciences du Langage) - Lattice
30 Rue du Château des Rentiers 75013 Paris, http://www.linguist.jussieu.fr
1 rue Maurice Arnoux 92120 Montrouge, http://www.lattice.cnrs.fr
yayoi@free.fr
Date de soutenance prévue : 2006

Mots-clefs – Keywords

Alignement, corpus parallèles, analyse morphologique japonaise partielle, mémoire de traduction
Alignment, parallel corpora, partial japanese morphological analysis, translation memory

Résumé - Abstract

あられ【arare】n.
1. Perle de glace. **2.** Petit biscuit de riz. **3.** **INFORM.** *AlALeR (système d'Alignement Autonome, Léger et Robuste)* Aligneur adapté au traitement du japonais caractérisé par l'absence d'utilisation d'analyseur morphologique et de dictionnaire.

Le présent article décrit le Système AlALeR (Système d'Alignement Autonome, Léger et Robuste). Capable d'aligner au niveau phrastique un texte en français et un texte en japonais, le Système AlALeR ne recourt cependant à aucun moyen extérieur tel qu'un analyseur morphologique ou des dictionnaires, au contraire des méthodes existantes. Il est caractérisé par son analyse morphologique partielle mettant à profit des particularités du système d'écriture japonais et par la transcription des mots emprunts, à l'aide d'un transducteur.

The present paper describes the AlALeR System, an Autonomous, Robust and Light Alignment System. Capable of aligning at the sentence level a French text and a Japanese one, the AlALeR System doesn't use any external tool, such as morphological parsers or dictionaries, contrary to existing methods. This system is characterized by a partial morphological analysis taking advantage of some peculiarities of japanese writing system, and by the transcription of loan words with a transducer.

1 Introduction

La plupart des méthodes d'alignement au niveau phrastique sont caractérisées par leur simplicité de réalisation et de calcul, obtenue grâce à l'utilisation exclusive d'informations « internes », telles que la distribution de chaînes de caractères ou la longueur de chaîne.

Mais, les systèmes d'alignement du japonais intègrent tous aussi bien un analyseur morphologique que des dictionnaires bilingues pour traiter cette langue très différente des langues telles que l'anglais, l'allemand ou le français.

Cependant, l'alignement étant une opération élémentaire constituant souvent une étape préparatoire pour un autre traitement, un système léger est favorable pour le traitement du japonais également. Ainsi, nous avons conçu le système `AlALeR` adapté à l'alignement des textes japonais, qui ne recourt à aucun moyen extérieur, ni dictionnaire ni analyseur morphologique, en mettant pleinement à profit certaines particularités du système d'écriture du japonais.

Nous présentons dans cet article ce nouveau système d'alignement du japonais : après un bref parcours des techniques antérieures, nous décrirons d'abord le principe de fonctionnement de ce système avec le détail des procédures de certaines opérations, et présenterons ensuite les résultats obtenus.

2 Techniques antérieures et Système `AlALeR`

Les recherches sur la technique d'alignement ont débuté dans le cadre de travaux sur la traduction automatique. Si bien que les précurseurs ont cherché avant tout la simplicité de réalisation et de calcul, donnant ainsi naissance à des méthodes caractérisées par l'utilisation exclusive d'informations internes telles que la distribution lexicale (KAY & RÖSCHEISEN, 1993) ou la longueur des phrases (BROWN *et al.*, 1991), (GALE & CHURCH, 1993).

Les chercheurs occidentaux ont choisi pour améliorer la technique, la poursuite de la voie initiée par ces précurseurs en introduisant de nouvelles notions telles que les cognats ((SIMARD *et al.*, 1992), (LANGLAIS, 1997) et (KRAIF, 2001)), qui ne font pas appel aux informations extérieures.

Néanmoins, du fait que le système d'écriture du japonais ne dispose pas de séparateur graphique indiquant les frontières entre les mots, les chercheurs japonais ont intégré très tôt des analyseurs morphologiques dans leurs systèmes d'alignement (MURAO, 1991). De plus, le japonais est fortement différent des langues principalement traitées dans le TAL – telles que l'anglais, le français ou l'allemand – aussi bien sur le plan syntaxique que sur le plan lexical, ce qui n'a pas permis une simple application des méthodes utilisées pour ces langues au traitement des textes japonais. Aussi, les Japonais ont-ils également dû recourir à des dictionnaires bilingues et rechercher la performance plutôt que la simplicité (HARUNO & YAMAZAKI, 1996).

Notre système a résolu le problème de segmentation par une analyse morphologique partielle basée sur une méthode traditionnelle[1] qui profite d'une particularité du système d'écriture du japonais, possédant plusieurs types de caractères différents. Par ailleurs, la transcription des mots emprunts, à l'aide d'un transducteur, a permis un meilleur alignement des mots sans recourir à un dictionnaire bilingue.

[1] Voir aussi (NAKAMURA-DELLOYE, 2003).

3 Principe de fonctionnement

3.1 Schéma général du Système

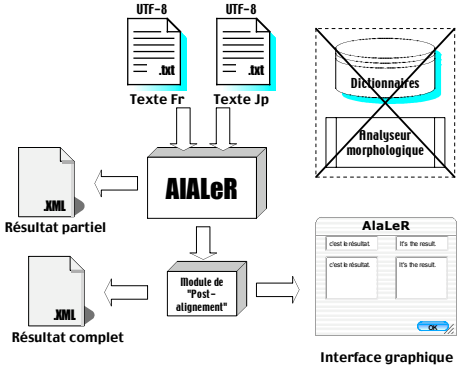

FIG. 1 – Schéma général du Système AlALer

Le système reçoit comme données une paire de textes parallèles rédigés en français et en anglais, ou plus particulièrement d'un texte en français (ou en anglais) et d'un en japonais. Afin de s'affranchir des problèmes d'encodage, fréquents lorsqu'il s'agit de traitements multilingues, AlALeR présuppose comme entrées des textes bruts au format texte, encodés en UTF-8.

Le système peut fournir comme résultat soit un alignement partiel très fiable des textes entrés, soit un alignement complet avec l'option « complet ». Lorsque cette option est choisie, le module de « post-alignement » réalise un appariement des phrases qui n'ont pas été alignées pendant le processus principal. L'appariement de ce module est réalisé selon la probabilité d'alignement de paires de phrases, calculée à partir de la corrélation de leur longueur.

Les résultats sont fournis, soit sous forme de fichier XML, soit par transfert vers l'interface graphique. Celle-ci permet non seulement de visualiser le résultat sous un format plus agréable à lire, mais aussi de faciliter la vérification et éventuellement la modification manuelle des résultats fournis par le système[2].

3.2 Procédure générale

La procédure générale est constituée de deux grandes étapes :

1. **Étape de construction de l'index du lexique**, au cours de laquelle les mots graphiques sont triés pour constituer quatre listes selon leur nature : transfuges, cognats, *katakana* et mots lexicaux ;
2. **Procédure d'alignement** :
   - **étape de préalignement**, au cours de laquelle un premier ancrage est réalisé pour limiter le nombre de possibilités d'alignement, à l'aide notamment des transfuges et des cognats ;
   - **procédure principale**, au cours de laquelle les phrases sont alignées par un calcul de similarité de la distribution des mots qu'elles contiennent. Cette procédure est itérative.

[2]Le système est implémenté en langage C++ et l'interface graphique avec les API Apple Carbon.

Nous allons maintenant présenter chaque étape. Étant donné que le système fonctionne un peu différemment selon les langues traitées, nous ne nous préoccupons ici que du cas d'un alignement de textes français et japonais, afin de mieux montrer la particularité de notre système.

3.3 Construction de l'index du lexique

Cette étape est composée elle-même de quatre étapes :
– Construction de la liste des phrases (`LPH`).
– Construction de la liste des mots graphiques (`LMOT`).
– Création de quatre listes à la suite du tri des mots graphiques : liste des transfuges (`LTRNS`), liste des cognats (`LCOG`), liste des mots en *katakana* (`LKTKN`) et liste des mots lexicaux (`LEX`).
– Création de l'index des mots lexicaux après leur lemmatisation (`ILX`).

3.3.1 Construction de la liste des phrases

Comme il a déjà été mentionné dans (SIMARD & PLAMONDON, 1998), la reconnaissance des phrases représente à elle-seule, malgré l'impression de trivialité que l'on a généralement, une question à part entière. La segmentation en phrases de textes français ou anglais n'est pas évidente à cause du caractère polysémique du séparateur graphique principal de phrase, le point final. Il est donc nécessaire de définir des règles assez détaillées permettant de segmenter correctement les séquences contenant des abréviations ou des sigles (« U.S.A »), des séquences symboliques (« abc@cdf.fr ») ou encore des nombres décimaux (1.5 en anglais).

Le point final japonais est beaucoup moins polysémique, facilitant ainsi la tâche de découpage.

3.3.2 Extraction des mots graphiques

Lors de la deuxième étape, consacrée à la construction de la liste des mots graphiques, la liste pour le texte français est construite par extraction des séquences entourées de séparateurs graphiques des mots – préalablement définis.

Pour le texte japonais – dans lequel il n'existe pas de séparateur graphique indiquant les frontières entre les mots –, une étape lourde de segmentation, réalisée généralement par analyse morphologique, est nécessaire. Cependant, il est possible de reconnaître la plupart des mots lexicaux sans analyse morphologique complète, en profitant d'une particularité du système d'écriture du japonais qui utilise trois types de caractères différents selon la nature des mots : *hiragana*, *katakana* et *kanji*.

– ***hiragana*** : premier syllabaire japonais souvent utilisé pour représenter la partie variable des mots variants et les mots grammaticaux ;
– ***kanji*** : idéogrammes utilisés pour représenter les mots pleins et les radicaux ayant un sens ;
– ***katakana*** : second syllabaire japonais employé pour la transcription des mots emprunts des langues étrangères (sauf le chinois).

Quoiqu'il soit impossible de segmenter totalement de manière correcte une phrase en mots uniquement avec cette méthode, il est possible de reconnaître la plupart des mots lexicaux en extrayant les séquences de *kanji* ou de *katakana*.

La liste ainsi obtenue ne contient néanmoins pas de mots grammaticaux. Nous supprimons donc les mots grammaticaux de la liste LMOT du texte français à l'aide d'une liste de mots grammaticaux préalablement définie.

3.3.3 Tri des mots

Le tri est ensuite réalisé aussi bien pour la liste LMOT du texte français que pour celle obtenue à partir du texte japonais afin de construire quatre nouvelles listes : la liste des transfuges (LTRNS), la liste des cognats (LCOG), la liste des mots en *katakana* (LKTKN) et la liste des mots lexicaux (LEX).

Pour les mots des trois premières listes, leur équivalence traductionnelle est calculable simplement par leur forme. Qui plus est, le résultat de ce calcul est beaucoup plus sûr que le résultat obtenu par la similarité des distributions.

Les « **cognats** », mots apparentés, sont des chaînes de caractères identiques ou proches graphiquement se trouvant dans les lexiques de langues ayant une relation historique plus ou moins étroite, tels que les paires anglais-français *generation/génération* et *error/erreur*. La notion de cognats améliore de manière simple et économique les méthodes statistiques qui n'utilisent aucune information lexicale, encore que son efficacité soit limitée aux langues appartenant à une même famille. Cependant, le japonais intégrant également dans son système d'écriture l'alphabet latin (*rôma-ji*), la possibilité d'obtention d'un résultat a été signalée très tôt dans (CHURCH *et al.*, 1993).

Le système AlALeR ne considère comme cognats que les chaînes alphabétiques totalement identiques apparaissant dans les deux textes entrés. Le système constitue d'abord la liste LCOG du texte japonais en extrayant les mots écrits en alphabet latin. Ensuite, en se référant à la liste japonaise, il construit une liste française en recherchant les séquences identiques aux éléments de la liste japonaise.

Les « **transfuges** » sont des chaînes invariantes à la traduction telles que les chiffres ou les symboles, inclus au début dans les cognats par les définitions du domaine de l'alignement, et regroupés plus tard par (LANGÉ & GAUSSIER, 1995) pour constituer une nouvelle catégorie. Les listes de transfuges LTRANS sont constituées séparément dans les deux langues par simple extraction des séquences de symboles ou de chiffres.

La troisième liste contient les mots du texte japonais écrits en *katakana*. La figure 2 représente la procédure d'appariement d'un mot en *katakana*. Ces transcriptions des mots emprunts sont retranscrites par le système à l'aide d'un transducteur – que nous avons développé spécifiquement – en une ou éventuellement en plusieurs formes en alphabet latin. Au cours du tri des mots français, si la similarité entre le mot français considéré et une séquence de retranscription d'un mot en *katakana* atteint un seuil prédéfini, le mot français est stocké dans la liste LKTKN. Les mots japonais en *katakana* qui n'ont pas trouvé d'équivalent une fois le tri des mots français terminé, sont stockés dans la liste des mots lexicaux pour leur laisser à nouveau une chance d'être finalement alignés par la similarité de distribution.

Le calcul de similarité entre une séquence retranscrite et un mot français est proche de la méthode de la sous-chaîne maximale parallèle utilisée dans (KRAIF, 2001) pour la reconnaissance des cognats. Notre formule, adaptée aux besoins particuliers de la retranscription des *katakana*, diffère de celle de ce dernier par le fait qu'elle tient compte non seulement de la sous-chaîne

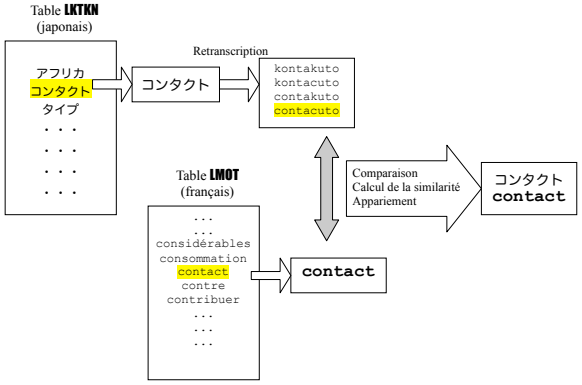

FIG. 2 – Appariement des mots en *katakana*

maximale parallèle 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.

Les paires constituées de deux mots appartenant à l'une de ces trois listes constituent ensuite des listes de paires de mots alignés, appelées table des « Transfuges alignés » (TRAL), table des « Cognats alignés » (COGAL) et table des « Katakana alignés » (KTKNAL).

3.3.4 Lemmatisation

Lemmatisation des mots français Nous avons eu recours à la méthode utilisée à l'étape morphologique dans (KAY & RÖSCHEISEN, 1993). Elle consiste à trouver les sous-chaînes préfixales ou suffixales communes à plusieurs formes effectives des mots graphiques et à trouver ensuite leurs radicaux, porteurs de sens. Ce traitement est implémenté efficacement grâce à l'utilisation d'une structure de données appelée *trie*.

Lemmatisation des mots japonais Après la segmentation par type de caractère, il subsiste encore un cas de segmentation à réaliser : les séquences de mots composés constitués de plusieurs substantifs juxtaposés les uns derrières les autres. Dans ce type de séquence, généralement entièrement en *kanji*, la frontière entre les deux mots composants n'est pas marquée par un changement de type de caractère.

Il s'agit donc également de la recherche des sous-chaînes communes à plusieurs formes effectives. Si bien que nous avons adopté pour le japonais une lemmatisation reposant sur la structure de données *trie*.

La différence dans le cas du japonais est que les parties restantes ne sont pas des suffixes mais un ou même plusieurs autres mots portant eux-mêmes un sens propre. On obtient donc à partir d'un mot graphique ab, non pas un lemme a, mais deux lemmes a et b. La figure 3 représente un exemple d'arbres vérifiant des chaînes préfixales et suffixales, créés à partir de sept entrées. Elle montre comment segmenter les mots japonais à l'aide de ces arbres. En réalisant de manière itérative cette opération, on peut obtenir plus de deux lemmes si la séquence en contient.

Nous obtenons ainsi l'ensemble des données nécessaires à la mise en correspondance des mots permettant d'associer ensuite les phrases à aligner.

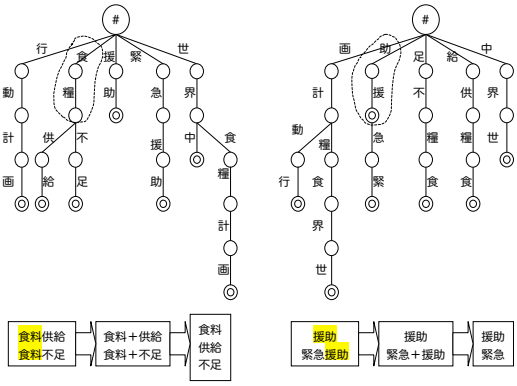

FIG. 3 – Arbres vérifiant des chaînes préfixales (à gauche) et suffixales (à droite)

3.4 Procédure d'alignement

Notre système utilise une technique basée sur les informations des distributions lexicales, présentée par (KAY & RÖSCHEISEN, 1993). Cette méthode, reposant sur l'hypothèse que les phrases correspondantes comprennent des éléments correspondants, est constituée d'un appariement grossier des mots, qui permet ensuite l'alignement des phrases contenant les mots appariés.

Les deux textes sont représentés par une matrice dont les lignes correspondent à chacune des phrases du texte français et les colonnes à celles du texte japonais. L'étape d'alignement est précédée par le préalignement et composée de trois opérations correspondant chacune à la construction d'une structure de données particulière : la table « Candidats des paires de phrases à aligner » (CPR), la table « Mots alignés » (MAL) et la table « Résultat d'alignement » (RAL).

3.4.1 Préalignement

Le préalignement consiste à trouver des ancrages sûrs permettant de réduire la zone de recherche. Le préalignement de notre système, inspiré de la méthode proposée dans (KRAIF, 2001), est réalisé à l'aide des tables TRAL, COGAL et KTKNAL. Il se fait via deux parcours de ces tables.

3.4.2 Table « Candidats des paires de phrases à aligner »

La table CPR est une matrice indiquant les paires de phrases susceptibles d'être alignées. Basée sur l'hypothèse de diagonalité de l'alignement, la zone constituée des cases correspondant aux paires candidates forme une ellipse avec pour axe principal la diagonale de la matrice.

3.4.3 Table « Mots alignés »

La table MAL contient l'ensemble des paires de mots supposés être traductions l'un de l'autre.

L'appariement des mots est réalisé selon la similarité de la distribution de chaque mot. Tous les mots appartenant à un même candidat paire de phrases sont comparés, et leur est attribuée

une similarité basée sur leur distribution. De nombreuses formules ont été proposées jusqu'aujourd'hui pour le calcul de cette similarité de distribution lexicale. Notre méthode est inspirée de l'amélioration par (KITAMURA & MATSUMOTO, 1997) du coefficient de Dice : en plus de la différence de fréquences, elle tient également compte de la fréquence elle-même, donnée contrôlée séparément dans les algorithmes antérieurs. La nouveauté apportée par notre formule est la prise en compte du nombre de phrases où les mots considérés apparaissent. Cette modification améliore les résultats lorsque deux paires ont une similarité identique, situation entraînant des conflits avec les méthodes précédentes.

3.4.4 Table « Résultat d'alignement » et module de post-alignement

La table RAL contient l'ensemble des paires de phrases supposées être traductions l'une de l'autre.

L'appariement des phrases utilise la table MAL obtenue précédemment, en plus des tables TRAL, COGAL et KTKNAL, pour calculer combien de couples de mots de ces tables contient chaque paire de phrases appartenant à la CPR. Si une paire de phrases comporte plus de paires de mots alignés que le seuil défini – en fonction de la taille du texte –, ces phrases sont considérées comme correspondantes traductionnelles. Ces nouvelles paires servant de nouvelles ancres, on crée une nouvelle CPR pour réaliser de manière itérative ces opérations d'alignement.

Ce premier résultat partiel peut être complété par une procédure de post-alignement. Le module de post-alignement extrait les sous-matrices constituées des phrases non alignées par le noyau AlALeR et calcule la probabilité d'alignement de toutes les paires possibles de phrases. Il réalise ensuite l'appariement de ces phrases avec une méthode de programmation dynamique, de manière à mettre en relation toutes les phrases avec au moins une phrase de l'autre texte.

4 Résultat

Nous avons testé les performances de notre système (sur PowerMac G5) avec cinq textes parallèles français-japonais et deux anglais-japonais : deux articles de *Label France*[3] (désignés ci-après « Bio » et « FIV »), un texte du sommet G8 (« G8 »), *How to Unicode* (« Unicode »[4]), un texte de l'EU (« EUJP ») et deux œuvres littéraires (un extrait de *Zadig* de Voltaire (« Zadig ») et *Balthasar* (« Balth ») d'Anatole France).

Pour chaque texte, nous avons analysé deux résultats : le résultat partiel sans opération de post-alignement (noyau AlALeR) et le résultat complet obtenu grâce au post-alignement.

	Bio		FIV		G8		Unicode		Zadig		EUJP		Balth	
Lang	Fr	Jp	Fr	Jp	Fr	Jp	Fr	Jp	Fr	Jp	Ang	Jp	Ang	Jp
Phr	69	75	54	52	53	47	274	268	1206	1376	252	238	321	423
M/C	1418	3615	1176	2597	1398	3077	4224	14155	17912	43426	3881	14308	4835	11491

TAB. 1 – Caractéristiques des textes

[3] Magazine du Ministère des affaires étrangères

[4] Sites internet : (VF) http://www.freenix.fr/unix/linux/HOWTO/Unicode-HOWTO.html ; (VJ) http://www.linux.or.jp/JF/JFdocs/Unicode-HOWTO.html

	Modèles de traduction													partiel		complet
	0-1	1-0	1-1	1-2	1-3	1-4+	2-1	2-2	2-3	2-4+	3-1	3-2	4+ -1	rap	pré	pré
Bio	0	0	55	7	1	0	3	0	0	0	0	0	0	0.81	1	0,98
FIV	0	0	43	3	0	0	2	0	0	0	0	0	1	0,66	1	0,92
G8	0	0	38	1	0	0	7	0	0	0	0	0	0	0,95	1	0,98
Unicode	1	0	195	22	1	0	19	2	0	0	1	1	1	0,90	0,99	0,96
Zadig	7	5	773	188	29	3	69	9	6	1	10	0	2	0,34	0,97	0,78
EUJP	0	4	208	5	1	0	17	0	0	0	0	0	0	0,87	0,98	0,92
Balth	1	2	185	68	16	4	9	13	1	0	0	0	0	0,49	0,97	0,86

TAB. 2 – Modèles de traduction et résultats : rappel et précision

Le tableau 1 montre les nombres de phrases et de mots (textes français) ou de caractères (textes japonais) de chaque texte.

Le tableau 2 présente la répartition par modèle de traduction de chaque paire de textes. La colonne 1-1 montre le nombre de paires en relation traductionnelle, constituées d'une phrase du premier texte (français ou anglais) et d'une du second texte (japonais), la colonne 1-2 le nombre de paires constituées d'une phrase du texte 1 et de deux phrases du texte 2, et ainsi de suite.

Le tableau 2 présente également le résultat d'alignement avec les taux de précision et de rappel, dans le cas du résultat partiel[5].

On peut déduire de l'analyse de ce tableau que le système supporte bien les modèles complexes – i.e. les modèles constitués de plus d'une phrase, tels que 1-3 –, qui perturbaient les systèmes d'alignement basés sur des méthodes probabilistes uniquement, au point de fausser tous les alignements effectués après l'analyse d'un modèle complexe. Cette robustesse est due au résultat partiel extrêmement précis, qui sert d'ancrage fiable pour le post-alignement plus robuste. Le point faible de cette méthode par rapport aux méthodes probabilistes est, comme déjà critiqué par plusieurs auteurs, son utilisation importante de mémoire.

Par ailleurs, le taux de rappel très bas de certains textes est dû non seulement à la présence faible voire l'absence de cognats ou de transfuges, mais aussi à la présence importante des mots de fréquence faible, notamment 1. C'est justement un autre point faible des méthodes basées sur la similarité de distribution. Afin de compenser cet inconvénient, notre système adopte un appariement final basé sur la corrélation des longueurs.

5 Conclusion et perspectives des travaux futurs

Les résultats d'alignement fournis par notre système AlALeR ont montré la possibilité de conception d'un aligneur traitant les textes japonais qui ne recourt à aucun dictionnaire ni analyseur morphologique. Ce résultat est d'abord dû à la stratégie d'appariement des mots japonais en *katakana*. Ceux-ci étant très nombreux dans les textes traduits, la retranscription de mots japonais en *katakana* pour trouver leur mot d'origine a été d'autant plus efficace qu'ils sont souvent absents des dictionnaires. En effet, ce sont très souvent des néologismes ou des noms propres. Cette stratégie s'est montrée extrêmement robuste, ce que nous n'aurions pas pu constater si

[5]Le taux de rappel représente la proportion de phrases appariées avec au moins une phrase de l'autre texte et le taux de précision, la proportion parmi ces phrases appariées de celles l'étant correctement, avec au moins une phrase de l'autre texte.

nous avions dépendu d'un dictionnaire.

Nous avons également testé le système avec quelques traductions de brevets techniques et nous avons obtenu de très bons résultats grâce à la présence très importante de transfuges. Néanmoins, les phrases de ce type de document sont si longues que l'alignement au niveau phrastique ressemble plutôt à un alignement de paragraphes. Comme Simard le fait remarquer dans (SIMARD, 2003), l'alignement à un niveau sous-phrastique est plus bénéfique que celui réalisé au niveau phrastique, notamment en vue de la constitution de mémoires de traduction. Nous allons désormais nous attacher à réaliser un alignement de propositions, qui permettra très certainement de fournir une base de données plus intéressante, aussi bien pour la conception des mémoires de traduction que pour l'étude des linguistiques contrastives.

Références

BROWN P. F., LAI J. C. & MERCER R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, p. 169 – 176.

CHURCH K., DAGAN I., GALE W., FUNG P. & J. HELFMAN B. S. (1993). Aligning parallel texts : do methodes developed for english-french generalize to asian languages ? In *Proceedings of the Pascific Asia Conference on Formal and Computational Linguistics*.

GALE W. A. & CHURCH K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**(3), 75–102.

HARUNO M. & YAMAZAKI T. (1996). Bilingual text alignment using statistical and dictionary information. *IPSJ SIG Notes*, **NL 112**(4), 23–30. en japonais.

KAY M. & RÖSCHEISEN M. (1993). Text-translation alignment. *Computational Linguistics*, **19**(1), 121–142.

KITAMURA M. & MATSUMOTO Y. (1997). Automatic extraction of translation patterns in parallel corpora. *IPSJ Journal*, **38**(4), 727– 736. en japonais.

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

LANGÉ J.-M. & GAUSSIER E. (1995). Alignement de corpus multilingues au niveau des phrases. *TAL*, **36**(1–2).

LANGLAIS P. (1997). Alignement de corpus bilingues : intérêt, algorithmes et évaluation. In *Bulletin de Linguistique Appliquée et Générale, numéro Hors Série*, p. 245–254. Université de Franche-Comté.

MURAO H. (1991). Studies on bilingual text alignment. Bachelor thesis, Kyoto University. en japonais.

NAKAMURA-DELLOYE Y. (2003). Analyse syntaxique du japonais. Mémoire de D.E.A., Institut National des Langues et Civilisations Orientales.

SIMARD M. (2003). *Mémoire de traduction sous-phrastique*. Thèse de doctorat en informatique, Université de Montréal.

SIMARD M., FOSTER G. & ISABALLE P. (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Thoretical and Methodological Issues in Machine Translation (TMI)*, p. 67 –81.

SIMARD M. & PLAMONDON P. (1998). Bilingual sentence alignment : Balancing robustness and accuracy. *Machine Translation*, **13**(1), 59–80.