TALN 2008, Avignon, 9–13 juin 2008

Y a-t-il une véritable équivalence entre les propositions syntaxiques du français et du japonais ?

Yayoi NAKAMURA-DELLOYE[1, 2]
(1) Université Paris VII - Lattice (UMR 8094), 75013 Paris
(2) Université Paris X - MoDyCo (UMR 7114), 92001 Nanterre Cedex
yayoi@free.fr

Résumé. La présente contribution part de nos constats réalisés à partir des résultats d'évaluation de notre système d'alignement des propositions de textes français-japonais. La présence importante de structures fondamentalement difficiles à aligner et les résultats peu satisfaisants de différentes méthodes de mise en correspondance des mots nous ont finalement amenés à remettre en cause l'existence même d'équivalence au niveau des propositions syntaxiques entre le français et le japonais. Afin de compenser les défauts que nous avons découverts, nous proposons des opérations permettant de restaurer l'équivalence des propositions alignées et d'améliorer la qualité des corpus alignés.

Abstract. This paper is based on our observations obtained from the results of our French-Japanese clause alignment system. Structures fundamentally difficult to align were so numerous and results obtained by various word-matching methods were so unsatisfactory that we questioned the existence of equivalence at the syntactic clause level between French and Japanese. In order to compensate the defect that we discovered, we propose some operations to restore aligned clause equivalence to improve the quality of aligned corpora.

Mots-clés : Alignement, proposition syntaxique, études contrastives français-japonais, similarité lexicale.

Keywords: Alignment, syntactic clause, French-Japanese contrastive study, lexical similarty.

1 Introduction

Du fait de l'intérêt incontestable de l'alignement des unités sous-phrastiques, nous avons développé un système réalisant l'alignement au niveau des propositions syntaxiques[1] et adapté au traitement des textes français-japonais (Nakamura-Delloye, 2007). Le système n'a pas pu donner de résultat satisfaisant et la principale cause d'échec provenait d'une part de la nature même de l'unité de proposition telle que nous l'avons définie, et de l'autre de la non pertinence des similarités calculées, due notamment au mauvais résultat de la mise en correspondance des mots entraînant un mauvais calcul de la similarité lexicale des propositions.

Nous avons donc essayé cette fois une autre approche tout à fait différente de mise en corres-

[1] On entend ici par proposition l'unité syntaxique constituée d'un sujet et d'un prédicat. Nous renvoyons pour notre définition précise de la notion de proposition à nos travaux antérieurs (Nakamura-Delloye, a paraitre).

pondance des mots, avec une méthode statistique. N'ayant pu obtenir de résultat satisfaisant avec cette approche non plus, une question nous est apparue : y a-t-il équivalence entre les propositions syntaxiques du français et du japonais ? Tous les mots lexicaux français ont-ils leur correspondant dans le texte japonais ? Les propositions françaises et japonaises ont-elles une similarité lexicale suffisante pour être alignées et pour être réutilisées ailleurs ? Afin d'élucider ces questions, nous avons alors examiné le corpus pour évaluer l'alignabilité des mots lexicaux entre les phrases françaises et japonaises.

Dans cet article, nous allons tout d'abord présenter brièvement notre système d'alignement des propositions et deux éléments pénalisant l'alignement que nous avons constatés lors de son évaluation (§ 2), avant d'aborder notre expérience sur la comparaison de méthodes de la mise en correspondance des mots (§ 3) qui nous a amenés à remettre en cause l'équivalence des propositions. Nous décrirons ensuite notre étude sur les propositions effectivement alignées (§ 4). Enfin, nous terminerons notre discussion par un exposé sur certaines possibilités pour rétablir l'équivalence entre les propositions du français et du japonais (§ 5).

2 Notre système d'alignement des propositions et problèmes généraux de cet alignement

Notre système présenté dans (Nakamura-Delloye, 2007) reçoit comme entrée la liste des propositions détectées, pour les textes français et japonais, avec leurs relations de dépendance et leur étiquette, qui permet de construire un arbre dépendanciel des propositions pour chaque phrase. Nous avons réalisé deux méthodes : l'une basée sur l'appariement des arbres de propositions avec un algorithme provenant de la théorie des graphes et l'autre basée sur la classification ascendante hiérarchique (CAH) mettant à profit les informations lexicales. Le résultat de l'évaluation[2] a montré une meilleure performance de la méthode CAH. En effet, du fait des différences considérables de structures, dans beaucoup de cas, les informations topologiques ne suffisaient pas et un alignement correct n'a été possible qu'avec la méthode à CAH basée sur la similarité lexicale. Cette dernière a amélioré considérablement le résultat, mais les propositions exactement correctement alignées restent de 50 à 70%, ce qui n'est pas, finalement, tout à fait satisfaisant.

La principale cause d'échec dans cette méthode provient de la non pertinence des similarités calculées, due notamment au mauvais résultat de la mise en correspondance des mots entraînant un mauvais calcul de la similarité lexicale des propositions. Toutefois, nous avons également constaté deux sources d'erreurs plus générales, indépendantes de la qualité de la méthode, résidant dans la nature même de l'unité de proposition.

[2]Nous avons réalisé une évaluation des méthodes proposées avec quatre corpus parallèles : (1) corpus LMD, constitués d'articles du Monde Diplomatique, (2) corpus BRVF, composés de deux brevets techniques et (3) BRVJ, composés d'un brevet technique, (4) corpus FdT, un extrait du roman « La fin des temps » de Haruki MURAKAMI. L'évaluation a été réalisée pour les paires de phrases comportant plus d'une proposition dans chaque langue, soit 526 paires de phrases.

Y a-t-il une véritable équivalence entre les propositions du français et du japonais ?

2.1 Deux éléments pénalisant l'alignement

Nous nous intéressons ici à deux problèmes généraux de l'alignement des propositions. Le premier est lié à la définition de l'unité de proposition et le second à l'existence de structures fondamentalement différentes. Nous allons d'abord examiner les problèmes de la différence du nombre des propositions, puis nous aborderons les cas d'exemples dans lesquels l'alignement des propositions est fondamentalement difficile, même manuellement.

Différence du nombre de propositions Une des causes du mauvais résultat de l'alignement des propositions est la présence importante de paires complexes difficiles à aligner, c'est-à-dire de paires constituées d'une proposition française et de plusieurs propositions japonaises ou vice-versa. C'est notamment le cas de notre corpus de brevets techniques dont le résultat est particulièrement médiocre : il comporte des paires constituées d'une proposition française et de sept jusqu'à onze propositions japonaises. Ceci est dû à la différence entre les définitions de la proposition adoptées pour le français et pour le japonais. En japonais, du fait de l'absence d'opposition sur la forme, on ne peut pas faire de distinction entre emplois fini et infini des mots variables. De plus, tout complément étant susceptible d'être omis, le repérage de la proposition dans la phrase japonaise se base essentiellement sur la présence d'un prédicat. Les propositions japonaises ainsi définies ne correspondent pas toujours aux propositions françaises définies sur la base de l'opposition sujet-prédicat. Beaucoup de propositions japonaises ont comme éléments équivalents en français des syntagmes participiaux. Dans les brevets techniques, les syntagmes participiaux sont utilisés de manière très importante, ce qui entraîne une différence considérable des nombres de propositions entre le français et le japonais. Et cette présence importante de paires complexes complique la tâche d'alignement.

Cas difficiles Dans l'évaluation des méthodes, nous avons constaté que l'échec provenait également de la différence considérable entre le français et le japonais. Cette différence réside sur différents plans – lexical, syntaxique ou encore rhétorique – si bien que nous avons rencontré des constructions très différentes de diverses natures. Ces exemples, pour lesquels un appariement même manuel est souvent très difficile, sont constatés plus particulièrement dans notre corpus littéraire. Nous examinons ici un exemple dans lequel nous avons constaté une différence sur le plan syntaxique.

Considérons les phrases parallèles suivantes :

> **Phrase française :** [F1$_{racine}$ c'est notamment lors des débats sur les programmes d'aide aux pays du sud [F2$_{postN}$ que les questions de la contraception et du statut de la famille sont abordées]]
> **Phrase japonaise :** [J1$_{theme}$ *hinin - to - kazoku - no - chii - toiu - mondai - wa*]
> (contraception - et - famille - de - statut - tel que - problème - [marqueur de thème])
> « les questions de la contraception et du statut de la famille [thème] »
> [J2$_{racine}$ *tokuni - kaihatsu tojô koku - enjo - puroguramu - wo - meguru - giron*
> (notamment - pays en voie de développement - aide - programme - [COD] - concerner - débat
> *- no naka de - ôkiku - toriage rare ta*]
> (à l'intérieur de - être abordé [passé])
> « être abordé, notamment lors des débats sur les programmes d'aide aux pays en voie de développement »

La phrase japonaise est constituée du thème (J1) et de la proposition racine (J2). Le thème

correspond au sujet de F2 « les questions de la contraception et du statut de la famille », et la racine comporte l'élément mis en relief « notamment lors des débats sur les programmes d'aide aux pays du sud » et le prédicat de F2 « sont abordées ». En effet, la mise en focus d'un syntagme du type « c'est ... que » peut être réalisée en japonais par la simple utilisation de particules dites casuelles. Il faut donc fusionner les deux propositions dans les deux langues pour établir la correspondance. Ce genre d'alignement, impossible à réaliser automatiquement avec l'utilisation seuls des relations de dépendance et types de propositions, reste très complexe même avec l'utilisation d'informations lexicales.

3 Étude de méthodes de la mise en correspondance des mots

Suite à l'évaluation de nos méthodes d'alignement, nous avons pensé à deux solutions possibles pour améliorer la mise en correspondance des mots, dont le mauvais résultat était considéré comme cause principale de notre échec dans la méthode basée sur la similarité lexicale. La première consiste en une amélioration du dictionnaire et la seconde, en la recherche d'un autre moyen de mise en correspondance.

Méthode utilisée dans notre système d'alignement des propositions L'inadéquation des dictionnaires existants à la mise en correspondance avait été, en fait, ressentie déjà au cours du développement de notre système. Une des raisons principales de cette difficulté est la différence de catégorie entre les mots correspondants en français et en japonais. Par exemple, l'adjectif français « économique » trouve généralement son correspondant dans les dictionnaires en l'adjectif japonais « *keizai-tekina* ». Lorsque nous avons « une raison économique » dans le texte français et « *keizai-tekina riyû* » dans le texte japonais, on peut obtenir la mise en relation de ces unités par consultation du dictionnaire. Mais, en japonais, l'expansion d'un nom est aussi réalisable par la simple juxtaposition de substantifs. Si bien qu'on peut trouver par exemple « croissance économique » dans le texte français et comme traduction « *keizai seichô* » dans le texte japonais. Dans le dictionnaire, le substantif « *keizai* » n'est couplé qu'avec le substantif « économie », et la mise en correspondance de « *keizai* » avec l'adjectif français « économique » est impossible. Nous avions donc réalisé une amélioration par l'introduction du calcul de la similarité des chaînes au moment de la mise en correspondance à l'aide du dictionnaire. Elle consiste à obtenir d'abord les candidats de traduction en français à l'aide du dictionnaire, puis à chercher la chaîne la plus proche d'un des candidats par des calculs de similarité des chaînes. Cette modification nous a apporté une amélioration de résultat : les paires alignées grâce au calcul de similarité représentaient 20% du résultat total. N'ayant pas pu obtenir de résultat satisfaisant d'alignement des propositions en dépit de cette amélioration, nous avons cette fois essayé une autre approche tout à fait différente de mise en correspondance : une méthode statistique.

Mise en correspondance des mots par une méthode probabiliste Nous avons réalisé l'alignement au niveau des mots d'un corpus parallèle[3] contenant 1691 phrases françaises et 1829 phrases japonaises à l'aide du système GIZA++ (Och & Ney, 2003) qui est une implémentation

[3]Le corpus est constitué de 15 articles du Monde Diplomatique tirés des numéros de janvier, février et mars 2004 (édition informatique), que nous avons utilisé lors de l'évaluation de notre système d'alignement des propositions.

Y a-t-il une véritable équivalence entre les propositions du français et du japonais ?

		A	B	C	D	E
		Avec (F–>J)	Avec (F<–J)	Avec (F<–>J)	Sans (F<–>J)	Dictionnaire
1	Nb. lemmes*	40,9			21,6	
2	Nb. tokens*	68,5			33,5	
3	Total*	36,6	37,8	9,9	10,2	13,3
4	Correct*	13,0	14,7	8,1	8,5	12,8
5	Rappel (4/2)	0,19	0,21	0,12	0,25	0,38
6	Précision (4/3)	0,36	0,39	0,81	0,83	0,96
7	F-mesure	0,25	0,27	0,21	0,38	0,54

* Nombre par paire de phrases

TAB. 1 – Tableau comparatif des résultats d'alignement

des modèles IBM 1-5 et d'un modèle HMM. Nous avons ensuite effectué une évaluation sur le résultat d'alignement des 39 paires de phrases (soit 2671 mots).

Le tableau 1 montre le résultat de cette évaluation. Les colonnes de (A) à (D) présentent les résultats de GIZA++, (A), (B) et (C) étant ceux obtenus avec le texte original comportant des mots grammaticaux et (D) avec le texte dans lequel les mots grammaticaux ont été supprimés. La colonne (E) présente les résultats obtenus lors de l'évaluation de notre système par consultation du dictionnaire avec calcul de la similarité des chaînes. La mise en correspondance avec le dictionnaire était réalisée avec un texte sans mots grammaticaux. La ligne (2) indique la moyenne du nombre des tokens par paire de phrases et la ligne (1) celle des lemmes par paire de phrases. Les lignes de (3) à (7) correspondent aux valeurs indiquant respectivement le nombre total de mots alignés par paire de phrases, le nombre de mots correctement alignés par paire de phrases, le rappel (5 = 4/2), la précision (6 = 4/3) et la F-mesure calculée à partir du rappel et de la précision.

Résultat de GIZA++ avec mots grammaticaux (MG) Nous avons tout d'abord réalisé un alignement avec comme texte source les phrases françaises et comme texte cible les phrases japonaises. Comme on peut le constater dans les résultats présentés colonne (A), plus de la moitié des mots ont été alignés avec un ou plusieurs mots de l'autre texte mais les paires correctes étaient extrêmement restreintes. La figure 1 (à gauche) montre un exemple d'alignement réalisé par ce système. Plusieurs mots du texte cible (japonais) sont mis en correspondance avec un même mot du texte source (français). Parmi ces correspondances (marquées par une ligne continue dans la figure), il n'existe souvent qu'un lien qui relie les éléments effectivement équivalents (éléments entourés reliés par une ligne discontinue). Un autre alignement a été réalisé dans le sens inverse, c'est-à-dire avec comme texte source les phrases japonaises et comme texte cible les phrases françaises. Le résultat de cet alignement japonais vers français présenté dans la colonne B (cf. figure 1, au milieu) est semblable à celui de l'alignement français vers japonais.

Résultat de GIZA++ : alignement bi-directionnel avec MG Nous avons essayé ensuite un alignement bidirectionnel (cf. colonne C et figure 1, à droite). Il consiste, afin d'élimer le bruit, à prendre en compte uniquement les paires alignées dans les deux alignements de sens inverses – l'alignement présenté dans la colonne A et celui présenté dans la colonne B. La précision a été nettement améliorée, au prix cependant d'une baisse considérable du rappel.

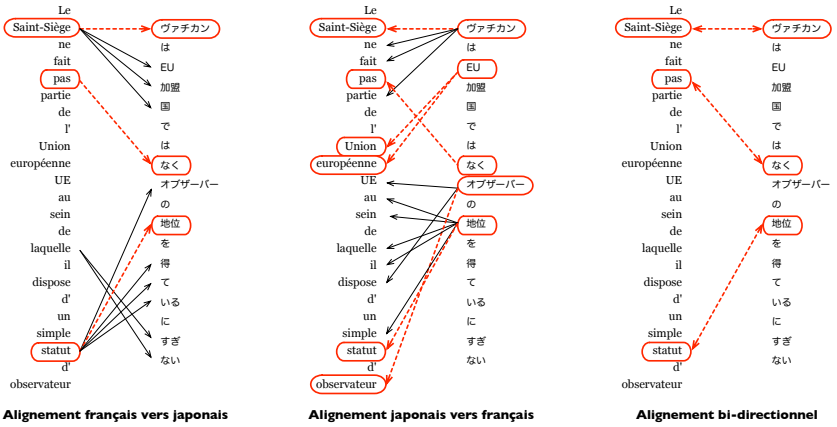

FIG. 1 – Résultats d'alignement des mots par GIZA++

Résultat de GIZA++ : alignement bi-directionnel sans MG Suite à cette observation, nous avons posé comme hypothèse que la présence des mots grammaticaux perturbait le bon fonctionnement de la mise en correspondance. En effet, lors de recherches antérieures pour le développement d'un système d'alignement des phrases (Nakamura-Delloye, 2005), nous avons constaté la difficulté d'alignement des mots grammaticaux du fait de la nature différente des mots grammaticaux français et des mots grammaticaux japonais. Par exemple, en japonais il n'existe pas d'article, ni de pronom relatif introduisant une subordonnée[4]. La présence de ces mots dans le texte français peut donc non seulement provoquer une mauvaise mise en correspondance mais aussi empêcher un bon alignement. Par ailleurs, les pronoms du français posent également des problèmes dans l'alignement avec le japonais. En effet, dans la phrase japonaise, l'omission d'éléments est extrêmement fréquente. Le seul élément obligatoire de la phrase est le mot variable prédicatif et aucun de ses compléments n'est obligatoire s'ils sont déductibles à partir du contexte. Ainsi, dans la phrase japonaise, sont très souvent absents les termes équivalents aux pronoms français – surtout anaphoriques qui servent à reprendre les termes déjà introduits et déictiques qui désignent les éléments présents dans le contexte. Nous avons donc supprimé les mots grammaticaux et les pronoms avant l'alignement. Malgré nos attentes, la suppression des mots grammaticaux n'a entraîné aucune amélioration significative comme le montre la colonne (D). De plus, cet alignement statistique n'a pas pu fournir de meilleur résultat que notre méthode basée sur la consultation du dictionnaire et le calcul de la similarité des chaînes (cf. colonne E).

Conclusion et travaux connexes Nous pourrions obtenir un meilleur résultat avec un modèle adapté au japonais proposé par (Yamada & Knight, 2001), qui a montré une meilleure efficacité par rapport au modèle IBM 5 (avec le score 0,582 contre 0,431). Une amélioration de ce modèle est également proposée par (Yamada *et al.*, 2003). Cette amélioration consiste en l'utilisation des paires de mots communs extraites par deux approches d'apprentissage asymétrique et elle a amélioré l'alignement des mots de 5,7 %. Cependant, le meilleur résultat présenté par la F-mesure dans cet article reste de 35 à 37 % pour le corpus journalistique, et les auteurs signalent dans la conclusion la nécessité et la possibilité d'autres améliorations. L'étude de l'alignabilité entre les mots grammaticaux français et japonais est un sujet incontournable pour le progrès des

[4]La subordination est marquée uniquement par la forme et la place du mot prédicatif de la subordonnée par rapport au mot prédicatif principal.

techniques d'alignement. Mais c'est aussi un sujet assez peu étudié. En effet, les travaux sur la mise en correspondance des mots traitant le japonais, sont souvent ceux des unités lexicales et dans ces travaux les mots grammaticaux sont généralement supprimés préalablement (Haruno & Yamazaki, 1996) (Kitamura & Matsumoto, 1997). Les travaux sur d'autres types d'alignement sous-phrastique qui se basent sur l'alignement des mots tels que (Kaji *et al.*, 1992) (Watanabe *et al.*, 2000) ne cherchent à mettre en correspondance que des mots lexicaux à l'aide d'un dictionnaire bilingue.

4 Constat sur l'alignabilité des mots lexicaux

Cette expérience de comparaison de techniques de mise en correspondance des mots nous a fait ressentir la nécessité de l'amélioration des techniques d'alignement des mots, mais surtout elle nous a finalement conduit à revenir à notre première analyse sur notre échec de la mise en correspondance des mots et à remettre en cause l'équivalence même des propositions entre le français et le japonais. Tous les mots lexicaux français ont-ils leur correspondant dans le texte japonais ? Les propositions françaises et japonaises ont-elles une similarité lexicale suffisante pour être alignées et pour être réutilisées ailleurs ? Afin d'élucider ces questions, nous avons examiné le corpus sans mots grammaticaux pour évaluer l'alignabilité des mots lexicaux entre les phrases françaises et japonaises.

Comme nous l'avons déjà abordé, les mots grammaticaux sont souvent impossibles à aligner du fait de l'absence même de notion équivalente dans l'autre langue. Nous avons donc décidé d'étudier l'équivalence en termes de mots lexicaux. Nous avons alors réalisé un alignement des 89 paires de phrases à la main (soient 1142 mots lexicaux français et 1084 mots lexicaux japonais). Dans ces phrases, 526 mots lexicaux (soit 24%) n'avaient pas leur correspondant dans l'autre texte. Il existe différentes explications à cette absence de mot correspondant : les erreurs des pré-traitements, l'inadéquation de la définition des mots grammaticaux, la présence d'insertion et d'omission par les traducteurs et de structures elliptiques.

Certaines absences de mots correspondants sont dues aux erreurs de l'analyse morphologique. L'attribution d'une mauvaise étiquette a entraîné la suppression de certains mots qui sont considérés comme mots grammaticaux sans l'être. Par ailleurs, il existe des pronoms français que nous supprimons en même temps que les mots grammaticaux, qui ont parfois un correspondant japonais. C'est notamment le cas des pronoms sujets. Toute omission de complément déductible à partir du contexte étant possible en japonais, le correspondant du pronom sujet français est souvent absent dans le texte japonais au niveau de surface car généralement il est déjà introduit dans les phrases précédentes. Toutefois il arrive parfois que le sujet soit marqué explicitement, sans doute pour le faire revenir au premier plan quand il est introduit depuis longtemps.

Il existe également des cas où l'absence de correspondant est due à l'omission de traduction, que ce soit consciemment ou inconsciemment, par le traducteur. Inversement, elle est parfois due à l'insertion d'éléments non existants dans le texte source par le traducteur qui l'a considérée comme nécessaire pour la compréhension. Parmi ces insertions, nous avons constaté, dans notre corpus, beaucoup de cas liés à l'ellipse due à la structure de coordination de la phrase française. Par exemple, dans la traduction japonaise de la phrase française : « *C'est un lieu d'échanges et de rencontres pour* ***les responsables politiques et catholiques*** *européens* », on constate la répétition du substantif déterminé comme « *les responsables politiques et les responsables catholiques* ».

5 À la recherche d'une équivalence au niveau des propositions

Ainsi, en termes d'unités lexicales de surface, il est difficile de parler d'équivalence des paires de propositions alignées du français et du japonais, du fait de la présence d'éléments dépendant du cotexte externe (e.g. anaphores, éléments elliptiques), et de la différence des moyens syntaxiques utilisés entre ces deux langues. Le mauvais résultat de la similarité lexicale des propositions qui a causé le résultat final non satisfaisant de l'alignement devait être dû, du moins partiellement, à cette non-équivalence lexicale. Pour compenser ce défaut, sont sans doute bénéfiques pour enrichir le corpus aligné quelques opérations permettant d'améliorer (ou de restaurer) l'équivalence des unités alignées et de réduire leur dépendance au contexte externe, comme par exemple : la résolution des anaphores, la restitution des éléments elliptiques, ou encore le traitement de l'élément de jonction.

Nous présentons donc dans cette section quelques possibilités pour rétablir l'équivalence entre les propositions du français et du japonais. Nous allons tout d'abord aborder le traitement des connecteurs, puis nous examinerons les opérations permettant d'annuler la dépendance des propositions japonaises au contexte externe. Dans les textes japonais, deux opérations supplémentaires sont envisageables : la restitution des compléments omis et la résolution des fonctions cumulatives du syntagme thématisé.

Traitement des éléments de liaison Comme le dit Harris, « dans une phrase S_i, il peut être possible d'identifier une phrase S_j accompagnée de matériel supplémentaire X » (Harris, 1976). Une proposition peut donc comporter un élément qui assure la liaison avec une autre proposition. Cet élément de jonction X peut entraîner des différences entre les propositions françaises et leurs correspondants japonais. C'est le cas par exemple de la marque de coordination. Elle apparaît souvent dans les propositions inversées, car l'ordre d'apparition des propositions coordonnées est souvent conservé dans la traduction, mais l'ordre syntaxique est à l'inverse. De même, comme le montre l'exemple de la figure 1, la jonction peut être assurée par des moyens complètement différents. Dans cet exemple, alors que dans la phrase française elle est réalisée par un connecteur relatif composé (« au sein de laquelle »), la première proposition japonaise qui correspond à la proposition racine (principale) française est reliée à la proposition racine à l'aide de la marque de coordination. Il pourrait donc être intéressant d'extraire les connecteurs des subordonnées (ou des coordonnées) et de mettre ces dernières sous forme de propositions indépendantes.

Restitution des compléments omis dans la phrase japonaise Comme nous l'avons déjà mentionné, dans la phrase japonaise, l'omission d'éléments est extrêmement fréquente tant qu'ils sont déductibles à partir du contexte. Autrement dit, le japonais est une langue dépendant fortement non seulement du cotexte mais aussi du contexte extra-linguistique. Cette particularité de la phrase japonaise qui est assez incomplète intrinsèquement, entraîne bien évidemment le caractère non-équivalent des unités japonaises par rapport aux unités du français, langue dans laquelle l'omission des éléments est beaucoup plus limitée et surtout strictement conditionnée. Il serait donc intéressant de compléter les éléments omis des propositions japonaises pour instaurer une équivalence avec les propositions françaises alignées.

Y a-t-il une véritable équivalence entre les propositions du français et du japonais ?

Résolution des fonctions cumulatives du syntagme thématisé Suivant la théorie de Mikami (Mikami, 1953), nous défendons la thèse que la structure fondamentale de la phrase japonaise est basée sur l'opposition thème-rhème et nous divisons tout d'abord la phrase japonaise en deux parties : thème et propositions syntaxiques. Le syntagme thématisé se situe à un niveau différent de celui des compléments régis par le mot variable prédicatif. En revanche, il peut assurer une fonction[5] que nous appelons cumulative, au sein de différentes propositions, même celles des autres phrases. Certains des compléments implicites de la phrase japonaise sont dus à ce mécanisme de fonction cumulative du thème. Il est donc possible de restituer ces compléments, sans aller chercher les candidats dans un cotexte plus large, voire le contexte extra-linguistique, par la résolution des fonctions cumulatives du syntagme thématisé. Pour mieux illustrer notre propos, référons-nous aux phrases parallèles suivantes :

> **FR :** `[F1` Les pièces de un et de cinq yen,`]` `[F2` **je les** mets dans ma poche revolver,`]` `[F3` mais en principe **je** ne **m'en** sers pas dans les calculs.`]`
> **JP :** *ichien dama - to - goen dama - wa*
> (pièce de un yen - [coordination] - pièce de cinq yens - [marqueur de thème])
> « les pièces de un et de cinq yen [thème] »
> *hippu poketto - ni - ireru - ga - gensoku toshite - keisan - ni wa - tsukawa nai*
> (poche revolver - dans - mettre - [opposition] - en principe - calcul - dans - utiliser [négation])
> « mettre dans la poche revolver , mais en principe n'utiliser pas dans les calculs »

Dans cet exemple, le thème japonais a trois éléments correspondants dans la phrase française : le thème en prolepse F1 « Les pièces de un et de cinq yen », le pronom clitique objet « les » dans la première proposition F2, enfin le pronom clitique objet « en » dans la seconde proposition F3. Cette détermination de la fonction cumulative du thème est une opération capitale, non seulement pour l'alignement ou l'enrichissement de corpus mais pour toute analyse automatique du japonais. L'automatisation de cette tâche est, malgré son importance, un sujet peu étudié, et il reste encore beaucoup de questions linguistiques à résoudre concernant ce sujet, avant d'arriver à le modéliser formellement.

6 Conclusion et perspectives

D'après les constats que nous avons présentés, nous apportons comme réponse à la question posée en début d'article qu'il est difficile de parler d'équivalence entre les propositions du français et du japonais, non seulement du fait de l'existence de structures fondamentalement différentes, mais aussi en termes d'unités lexicales de surface entre lesquelles on ne peut pas trouver de similarité suffisante. Toutefois, cette non équivalence lexicale peut être améliorée par certaines opérations afin d'augmenter la qualité des corpus alignés. La restitution des compléments non réalisés au niveau de surface est une opération utile voire indispensable pour beaucoup d'autres applications. Certains de ces compléments implicites de la phrase japonaise peuvent être restitués par la résolution des fonctions cumulatives du syntagme thématisé. Cette opération est incontournable pour toute analyse automatique, car elle revient à la neutralisation de la structure thématique de la phrase japonaise et la mise en relief du plan fonctionnel, qui correspond à la structure syntaxique fondamentale de certaines langues, notamment le français.

En ce qui concerne l'alignement des unités sous-phrastiques, du fait de la non-univocité des propositions française et japonaise telles que nous les avons définies, et de la difficulté même

[5] On parle parfois, dans le cadre de la grammaire générative, de « *theta topic* » (Mihara, 1994).

de leur définition et leur détection automatique, nous sommes finalement tentés par un autre type d'alignement sous-phrastique. Il s'agit d'un alignement hiérarchique (Kaji *et al.*, 1992) (Imamura, 2000), où les unités à aligner ne sont pas fixées et qui consiste, plutôt qu'à aligner certaines unités préalablement déterminées, à mettre en relation différentes structures de tout niveau afin d'obtenir un maximum de patrons parallèles. La construction d'une liste de patrons parallèles permettrait d'améliorer la traduction automatique de deux langues à schéma syntaxique fortement différent.

Références

HARRIS Z. S. (1976). *Notes du cours de syntaxe*. Paris : Seuil.

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

IMAMURA K. (2000). A hierarchical phrase alignment from english and japanese bilingual text. In *Proceedings of CICLing 2001*.

KAJI H., KIDA Y. & MORIMOTO Y. (1992). Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, p. 672–678.

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

MIHARA K. (1994). *Nihongo-no tôgo kôzô : seisei bunpô riron to sono ôyô [Structure syntaxique du japonais : la théorie de la grammaire générative et ses applications]*. Tokyo : Shohakusha.

MIKAMI A. (1953). *Gendaigohô josetsu [Introduction à la grammaire contemporaine]*. Tokyo : Toko shoin. Nouvelle édition publiée en 1972 par Kuroshio Shuppan.

NAKAMURA-DELLOYE Y. (2005). Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais. In M. JARDINO, Ed., *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, p. 585–594, Dourdan : ATALA LIMSI.

NAKAMURA-DELLOYE Y. (2007). Méthodes d'alignement des propositions : un défi aux traductions croisées. In F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds., *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, p. 223–232, Toulouse : ATALA IRIT.

NAKAMURA-DELLOYE Y. (à paraître). Typologie des subordonnés et des connecteurs en vue de la détection automatique des propositions syntaxiques du français. In *Description linguistique pour le traitement automatique du français*, Cahiers du Cental. Presses universitaires de Louvain.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

WATANABE H., KUROHASHI S. & ARAMAKI E. (2000). Finding structural correspondences from bilingual parsel corpus for corpus-based traslation. In *Proceedings of COLING 2000*, p. 906–912.

YAMADA K. & KNIGHT K. (2001). A syntax-based statistical translation model. In *Meeting of the Association for Computational Linguistics*, p. 523–530.

YAMADA S., NAGATA M. & YAMADA K. (2003). Improving Translation Models by Applying Asymmetric Learning. *Proceedings of MT Summit IX*.