École Doctorale STIM
Sciences et Technologie de l'Information et des Matériaux

Spécialité : Système d'Aide à la Décision
Laboratoire : LINA UMR CNRS 6241
Equipe : Traitement Automatique du Langage Naturel

Observation des translittérations dans un corpus comparable spécialisé trilingue

Emmanuel Prochasson
Mél : Emmanuel.Prochasson@univ-nantes.fr
Directrice de thèse : Béatrice Daille
Co-encadrant : Emmanuel Morin

Résumé : Nous observons la présence des translittérations dans le cas d'un corpus spécialisé japonais, puis nous mettons en évidence les relations que l'on peut trouver entre ce corpus japonais et des corpus français et anglais. Nous soulignons enfin la présence de liens forts, supportés par les translittérations, que l'on pourra utiliser dans le cas de l'extraction lexicale bilingue à partir de corpus comparable.
Mots clés : *corpus comparable, alignement lexical, translittérations*

1 Introduction

Les ressources linguistiques multilingues sont des données précieuses (elles sont par exemple utiles dans le cadre de la traduction automatique, de la recherche d'information multilingues ou plus directement, pour l'assistance à la traduction manuelle). Pour faire face à l'évolution du vocabulaire des langues vivantes, il est intéressant d'être capable de mettre à jour automatiquement ces ressources. C'est un besoin dans certaines disciplines (par exemple en médecine), en raison de la quantité de données publiées chaque jour, mais aussi en raison du vocabulaire très spécialisé utilisé par les experts de la discipline, rendant un traitement manuel fastidieux.

Il existe différentes méthodes d'acquisition automatique de lexique multilingue, ces méthodes se basent sur des corpus parallèles (des ensembles de documents dans des langues différentes, traductions les uns des autres) ou sur des corpus comparables (des ensembles de documents ayant des traits communs, par exemple un même thème). Alors que, dans le cas des corpus parallèles, il est possible de s'appuyer sur des points d'ancrages forts (la position d'un élément dans un document et la position de sa traduction sont liées), il est peu concevable d'utiliser ce type d'indices dans le cas des corpus comparables.

Nous nous sommes penchés sur l'apport des *translittérations*, c'est-à-dire les mots d'emprunts, dans l'optique d'utiliser ces relations entre langues comme un nouveau type de point d'ancrage, valorisable dans le cas de l'extraction lexicale à partir de corpus comparable. Après avoir introduit les méthodes d'alignement à partir de corpus comparables, nous présentons la notion de translittérations en insistant sur le cas du japonais, puis nous mettons en évidence ce phénomène au sein d'un corpus trilingue spécialisé anglais/français/japonais.

2 Extraction automatique de lexique multilingue

2.1 Extraction à partir de corpus parallèles

Un corpus parallèle est un ensemble de documents où chaque document *source* est lié un document *cible* qui en est la traduction (on peut donc séparer le corpus total en deux sous-corpus de langue l_s et l_c). Il est alors possible, à partir de la connaissance d'une langue d'inférer des éléments de traduction dans les autres langues en s'appuyant sur deux propriétés des corpus parallèles : i) chaque élément d'un document dans une langue l_s possède une correspondance dans le document correspondant de langue l_c et ii) la position d'un élément dans un document correspond à la position de sa traduction dans le document associé (au niveau, paragraphe, phrase...). L'extraction d'éléments de traduction à partir de corpus parallèles se fait donc principalement en s'appuyant sur la position et la distribution de ces éléments dans les documents *sources* et *cibles* [1]. Notons toutefois que ce n'est pas aussi direct, en effet il est fréquent qu'un mot dans une langue soit traduit avec plusieurs mots dans une autre langue, ou que la traduction regroupe plusieurs phrases en une seule, ou inversement. Ce phénomène de distorsion peut-être contourné en prenant en compte des points d'ancrages dans les documents. En supposant que

l'on connaisse avec certitude un couple de mots traduction dans deux documents parallèles, il est alors possible de s'appuyer sur ces traductions pour aligner leurs voisins.

2.2 Extraction à partir de corpus comparables

Un corpus comparable est un ensemble de documents dans des langues différentes partageant des caractéristiques communes mais n'étant pas des traductions mutuelles. Un groupe d'articles de journaux traitant d'un sujet commun à une même période, mais dans des langues différentes, est un exemple de corpus comparable. Les corpus comparables permettent de trouver des éléments de traduction d'une partie à l'autre (un vocabulaire commun), mais il n'est plus possible de s'appuyer sur la position des parties de document pour trouver ces correspondances puisqu'ils ne sont pas traductions.

L'approche standard s'appuie sur l'analyse des contextes lexicaux et l'hypothèse qu'un mot et sa traduction tendent à apparaître dans des contextes identiques[2]. Nous enregistrons alors sous forme de vecteur, pour chaque terme, l'ensemble des termes avec lesquel il coocurre (c'est-à-dire, apparaissant dans une fenêtre textuelle donnée). Ces vecteurs de contexte reflètent l'environnement des termes au sein des différents documents du corpus. Un processus d'alignement (appelé *approche directe*) consiste, à partir d'un terme en langue l_s, à traduire les éléments de son vecteur de contexte puis à chercher dans les vecteurs de contextes de la langue l_c ceux qui sont les plus semblables en utilisant des mesures de similarités (telles que le *Cosinus*). Cette méthode dépend toutefois largement de la qualité des dictionnaires bilingues utilisées pour la traduction des vecteurs de contexte, mais présente des résultats prometteurs [3].

3 À propos des translittérations

3.1 Cadre général

Dans cette étude, nous appelons *translittération* le phénomène qui consiste à prendre un mot d'une langue pour l'utiliser dans une autre langue, utilisant potentiellement des symboles d'écritures et/ou des phonèmes différents. Ce phénomène est différent des *cognats*, qui sont des mots partageant une origine commune et ayant évolué différement. Par exemple, le couple anglais/japonais `volley-ball`/バレーボール (`ba-re-e-bo-o-ru`) est une translittération, alors que le couple espagnol/portugais `estrella/estrela` (signifiant *étoile*) est un couple de cognats. Les translittérations peuvent être vues comme *la projection d'un mot, d'un langage source vers un langage cible*. Ce phénomène est fréquent dans toutes les langues vivantes pour permettre une évolution rapide du vocabulaire, de manière à s'adapter aux besoins des locuteurs. Il est donc particulièrement présent dans le cas du langage technique, qui est utilisé par une communauté restreinte d'utilisateur et ne passe donc pas par un processus d'assimilation global. De nombreux exemples peuvent être trouvés dans le vocabulaire informatique, utilisé tel quel en français alors même que des traductions simples existent (*shell*, *login*, *web*...).

3.2 Le cas du japonais

La langue japonaise est écrite avec trois ensembles de symboles différents [4]. Les kanjis (à l'origine, des idéogrammes chinois – quelques exemples : 人木子玉金), sont utilisés pour leur sens et peuvent être combinés pour former des mot pleins, alors que les hiraganas et les katakanas sont deux ensembles de symboles phonétiques équivalents, composé de 46 symboles chacun. Le syllabaire hiragana (quelques exemples : あえいおうかけき) est utilisé pour les mots communs pour lesquels il n'y a pas de kanjis disponibles ou lorsqu'ils sont inconnus du scripteur (typiquement pour les apprenants ne maitrisant pas les kanjis), mais aussi dans un but grammatical, pour marquer les partie du discours ou décliner les verbes. Le syllabaire katakana enfin (quelques exemples : アエイオウカケキ), est lui utilisé pour les translittérations (ce qui permet de les repérer facilement) mais aussi pour l'emphase ou pour représenter les onomatopées (sauf celles émises par des humains, alors notées en hiragana). Il y a de rares exemples de mots empruntés à des langues occidentales qui ne sont pas représentés avec le syllabaire katakana, tels que たばこ/`ta-ba-ko` (*tabac*), ici écrit en hiragana. Cette exception est en fait un emprunt au chinois (on peut aussi l'écrire avec les kanjis 煙草, signifiant *fumée* et *herbe*), qui l'avait emprunté à l'anglais.

3.3 Translittérations entre le français et le japonais

Bien que les emprunts entre la langue japonaise et le français soient rares, il reste interessant d'essayer d'aligner le vocabulaire des translittérations japonaises avec le vocabulaire français. En effet, de nombreux mots français sont proches, voire identiques à une partie du vocabulaire anglais (et par extension, à d'autres langues romanes ou germaniques), en raison des relations de cognats et de translittérations entre ces langues. Ainsi, l'alignement entre le japonais et le français peut se faire en passant par une langue *pivot*. Par exemple, ビタミン/`bi-ta-mi-n` peut s'aligner aisément avec le terme français *vitamine* ou le terme anglais *vitamin* car construit à partir du latin.

4 Distribution des translittérations dans un corpus comparable spécialisé

4.1 Ressources linguistiques

Nous avons collecté sur le web un corpus comparable anglais-français-japonais, composé de documents *scientifiques* (cela signifie que les textes sont écrits par des experts s'adressant à d'autres experts [5]) traitant des problématiques du diabète et de l'alimentation. Les documents ont été selectionnés manuellement, à partir de requêtes précises sur des moteurs de recherches, mais aussi extraits automatiquement à l'aide de PubMed (`http://www.ncbi.nlm.nih.gov/PubMed/`) pour la partie en anglais. Les documents ont enfin été convertis, à partir de documents HTML ou PDF en texte brut. Nous obtenons 257 000 mots pour la partie française, 235 000 pour la partie japonaise et 1 877 000 pour la partie anglaise. Notre première observation porte sur la place des translittérations dans le corpus japonais puis sur les liens de translittérations avec les corpus français et anglais pour mettre en évidence l'importance de ce phénomène dans les corpus spécialisés.

4.2 Analyse du corpus japonais

Nous avons extrait toutes les translittérations potentielles du corpus japonais, en isolant les séquences de katanakas. Nous ne travaillons pour le moment que sur les mots simples. Nous obtenons alors 493 translittérations potentielles, c'est-à-dire des termes qui existent effectivement dans la langue japonaise, en katakanas dans le corpus. Ces candidats représentent environ 8 % du vocabulaire total unique du corpus. Nous traduisons manuellement ces candidats, en français et en anglais lorsque c'est possible. La table 1 synthétise quelques statistiques concernant le classement des candidats en différentes catégories.

	#occ	%	Exemples
franç. seul.	4	0.8%	レバー/`re-ba-a`/*levure*, リール/`ri-i-ru`/*Lille*
angl. seul.	228	46%	ヘルス/`he-ru-su`/*health*, ダイエット/`da-i-e-tto`/*diet*
franç/angl.	221	45%	マネジャー/`ma-ne-ja-a`/*magnesium*, ヒスタミン/`hi-su-ta-mi-n`/*histamine*
adapté	12	2%	ビル/`bi-ru`/*building*, テレビ/`te-re-bi`/*télévision*
non fra/ang.	5	1%	カリウム/`ka-ri-wa-mu`/*potassium*, タイ/`ta-i`/*Thailand*
non trans.	23	5%	ムカデ/`mu-ka-de`/*mille-pattes*, カキ/`ka-ki`/*huître*

TAB. 1 – Statistiques concernant les séquences de katakana dans le corpus japonais

Franç. seul. (resp. *angl. seul.*) indique que les translittérations ne correspondent qu'à un mot français (resp. anglais) uniquement. Notons que la correspondance entre un mot et sa translittération répond à deux critères : i) les mots doivent être sémantiquement proches (traduction) et ii) les mots doivent être phonétiquement proches. D'un autre côté, *franç/angl* correspond aux translittérations pouvant s'aligner identiquement avec un mot français ou anglais. *Non fra/ang* correspond aux mots ne pouvant pas s'aligner avec un mot anglais ni avec un mot français. *Adapté* indique les mots issues d'une langue quelconque (français ou anglais) mais modifiés, tel que コンビニ/`ko-n-bi-ni`, correspondant à *convenient store*, si bien qu'il ne sont plus phonétiquement proche. Enfin, certains mots japonais peuvent être écrits en katakana sans être des translittérations (typiquement, les noms d'espèces), ils sont référencés dans la dernière ligne du tableau.

4.3 Correspondance avec les corpus anglais et français

Nous avons dégagé de nombreuses translittérations à partir du corpus japonais, il est maintenant intéressant de regarder si nous pouvons trouver les mots correspondants dans les corpus français et anglais. La table 2 indique le nombre de relation trouvés entre les corpus anglais et japonais, et français et japonais, par rapport au nombre total de translittérations que l'on peut potentiellement aligner à partir du corpus japonais, c'est-à-dire possédant une traduction phonétiquement proche dans la langue anglaise ou dans la langue française.

	Maximum	Présentes (dont hapax)	Rapport
anglais-japonais	449	**314** (26)	70 %
français-japonais	225	**140** (16)	62 %

TAB. 2 – Relations de translittérations entre les corpus

Ces résultats montrent que, non seulement les translittérations apparaissent dans des corpus isolés, mais qu'elles peuvent aussi couvrir une partie du vocabulaire commun entre les corpus qui est précisément celui que l'on cherche à extraire. Elles fournissent également un ensemble conséquent de lien entre des corpus multilingues, sur lesquels nous chercherons à nous appuyer par la suite pour améliorer le processus d'extraction.

5 Conclusion

Nous avons ici mis en évidence différentes propriétés des translittérations dans la langue japonaise et leur importance dans le cadre de corpus spécialisés. En effet, nous soulignons le fait que ce phénomène n'est pas marginal (il existe beaucoup de relations de translittérations entre les différentes langues du corpus). Ces observations nous laissent penser que les translittérations peuvent être utilisées efficacement dans le cadre de l'extraction lexicale bilingue à partir de corpus comparables spécialisés. Toutefois, plusieurs problèmes nécessitent d'être résolus, le premier étant la capacité de détecter automatiquement les couples de translittérations. En effet, les premières expériences utilisant des outils de détection se sont révélées infructueuses [6], le processus dégageant de nombreuses paires erronées difficiles à intégrer dans un processus d'alignement plus large. D'un autre côté, l'usage des translittérations connues n'est pas direct, plusieurs chemins doivent être explorés : les translittérations peuvent venir compléter les ressources multilingues utilisées pour la traduction des vecteurs de contextes, dans le cas des termes simples ou des termes complexes ; ou bien encore être mises en valeur parallèlement aux mesures de similarités, pour discriminer plus efficacement les vecteurs de contextes candidats à la traduction.

Références

[1] William A. Gale and Kenneth W. Church. Identifying word correspondence in parallel texts. In *HLT '91 : Proceedings of the workshop on Speech and Natural Language*, pages 152–157, Morristown, NJ, USA, 1991. Association for Computational Linguistics.

[2] Reinhard Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322, Morristown, NJ, USA, 1995.

[3] Yun-Chuang Chiao and Pierre Zweigenbaum. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan, 2002.

[4] Kyo Kageura. Character system, orthography and types of origin in japanese writing. In Reinhard Köhler, Gabriel Atmann, and Rajmund Piotrowski, editors, *Quantitative Linguistics : An International Handbook*, pages 935–946. Walter de Gruyter, 2005.

[5] Jennifer Pearson. *Terms in Context*. John Benjamins publishing company, 1998.

[6] Keita Tsuji, Béatrice Daille, and Kyo Kageura. Extracting french-japanese word pairs from bilingual corpora based on transliteration rules. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 499–502, 2002.