Looking for Transliterations in a Trilingual English, French and Japanese Specialised Comparable Corpus

Emmanuel Prochasson[∗], Kyo Kageura[†], Emmanuel Morin[∗], Akiko Aizawa[⋆]

[∗]Laboratoire d'Informatique de Nantes-Atlantique, Université de Nantes,
2 rue de la Houssinière, 44322 Nantes, France
{emmanuel.prochasson, emmanuel.morin}@univ-nantes.fr
[†]Graduate School of Education, the University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
kyo@p.u-tokyo.ac.jp
[⋆]Digital Contents and Media Sciences Research Division, National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
aizawa@nii.ac.jp

Abstract

Transliterations and cognates have been shown to be useful in the case of bilingual extraction from parallel corpora. Observation of transliterations in a trilingual English, French and Japanese specialised comparable corpus reveals evidences that they are likely to be used with comparable corpora too, since they are an important and relevant part of the common vocabulary, but they also yield links between Japanese and English/French corpora.

1. Introduction

Bilingual lexicon extraction from comparable corpora has received specific attention in recent years. This attention is motivated by the scarcity of parallel corpora, especially for language pairs not involving the English language. However, since comparable corpora are "sets of texts in different languages, that are not translations of each other" (Bowker and Pearson, 2002, p. 93), methods proposed for parallel corpora — that make use of fixed correlations between bilingual textual units such as word, sentence, paragraph… — are not applicable. For comparable corpora, the standard approach is based on lexical context analysis and relies on the assumption that a word and its translation tend to appear in the same lexical contexts (Rapp, 1995; Fung and McKeown, 1997; Peters and Picchi, 1998).

Although processing methods are distinct, bilingual corpora such as parallel or comparable corpora share, by essence, some transverse features such as words in one language that are orthographically or phonetically similar to a semantically related word in another language (cognates or transliterations). Cognates and transliterations yield anchor points that are useful to find extra clues for alignment of parallel texts (Simard et al., 1993). In the same way, we want to investigate the usefulness of the transliterations for the task of bilingual terminology extraction from specialised comparable corpora. We first introduce the concept of transliteration, especially concerning Japanese language and then present observations about transliterations in a trilingual English/French/Japanese specialised comparable corpus.

Note that this paper is not about automatic transliteration in comparable corpora, all transliterated units were extracted and aligned manually, as we were only concern by their prominence and relevance among specialised comparable corpora.

2. Overview of the transliteration phenomenon

In this study, we call *transliteration* the phenomenon of picking a word in one language to use it in another language, generally using different and non equivalent graphical symbols (to be accurate, a *loan word* is *transliterated* to fit a target language). This phenomenon differs from *cognates*, which are words sharing a common origin but evolved in different ways. For example, the English/Japanese pair volley-ball/バレーボール (ba-re-e-bo-o-ru – note that we will always give the Hepburn romanised version of Japanese terms introduced, each mora separated by a hyphen) is a transliteration, whereas the Spanish/Portuguese pairs estrella/estrela, meaning *star*, is a cognate.

In some cases, transliteration process is direct and the word is not changed at all (for example, *café*, *voilà*, *vis-à-vis* or *raison d'être*, which are used in French and in English, even though English language does not include any diacritical symbols in its alphabet). In other cases, however, the word need to be drastically transformed, which happen in English/French to Japanese transliterations, since Japanese does not share the same alphabet and does not include some very common English or French speech sound, such as cluster of consonants. Therefore, *hovercraft* is transformed to ホバークラフト (ho-ba-a-ku-ra-fu-to). Thus, transliterations can be seen as *the projection of a word, from a source language, into a target language*.

This phenomenon appears with many pairs of language such as western language (English, French, German…) and oriental language (Arabic, Chinese, Japanese...), in both ways. It is frequent in all languages which keep evolving, to allow a dynamic evolution of the vocabulary to fit needs of speakers. This is especially the case with technical vocabulary, which is intended to be shared by a community of experts and, at first, do not go through the regu-

	1	2	3	4	5	6	7	8
1	あ/ア/a	い/イ/i	う/ウ/u	え/エ/e	お/オ/o			
2	か/カ/ka/	き/キ/ki	く/ク/ku	け/ケ/ke	こ/コ/ko	きゃ/キャ/kya	きゅ/キュ/kyu	きょ/キョ/kyo
3	さ/サ/sa	し/シ/shi	す/ス/su	せ/セ/se	そ/ソ/so	しゃ/シャ/sha	しゅ/シュ/shu	しょ/ショ/sho
4	た/タ/ta	ち/チ/chi	つ/ツ/tsu	て/テ/te	と/ト/to	ちゃ/チャ/cha	ちゅ/チュ/chu	ちょ/チョ/cho
5	な/ナ/na	に/ニ/ni	ぬ/ヌ/nu	ね/ネ/ne	の/ノ/no	にゃ/ニャ/nya	にゅ/ニュ/nyu	にょ/ニョ/nyo
6	は/ハ/ha	ひ/ヒ/hi	ふ/フ/fu	へ/ヘ/he	ほ/ホ/ho	ひゃ/ヒャ/hya	ひゅ/ヒュ/hyu	ひょ/ヒョ/hyo
7	ま/マ/ma	み/ミ/mi	む/ム/mu	め/メ/me	も/モ/mo	みゃ/ミャ/mya	みゅ/ミュ/myu	みょ/ミョ/myo
8	ら/ラ/ra	り/リ/ri	る/ル/ru	れ/レ/re	ろ/ロ/ro	りゃ/リャ/rya	りゅ/リュ/ryu	りょ/リョ/ryo
9	や/ヤ/ya		ゆ/ユ/yu		よ/ヨ/yo			
10	わ/ワ/wa				を/ヲ/wo			
11	が/ガ/ga	ぎ/ギ/gi	ぐ/グ/gu	げ/ゲ/ge	ご/ゴ/go	ぎゃ/ギャ/gya	ぎゅ/ギュ/gyu	ぎょ/ギョ/gyo
12	ざ/ザ/za	じ/ジ/ji	ず/ズ/zu	ぜ/ゼ/ze	ぞ/ゾ/zo	じゃ/ジャ/ja	じゅ/ジュ/ju	じょ/ジョ/jo
13	だ/ダ/da	ぢ/ヂ/ji	づ/ヅ/zu	で/デ/de	ど/ド/do			
14	ば/バ/ba	び/ビ/bi	ぶ/ブ/bu	べ/ベ/be	ぼ/ボ/bo	びゃ/ビャ/bya	びゅ/ビュ/byu	びょ/ビョ/byo
15	ぱ/パ/pa	ぴ/ピ/pi	ぷ/プ/pu	ぺ/ペ/pe	ぽ/ポ/po	ぴゃ/ピャ/pya	ぴゅ/ピュ/pyu	ぴょ/ピョ/pyo
16	ん/ン/n							

Table 1: Standard Japanese mora. Column from 6 to 8 are mora composed with two symbols (note that the second one is smaller). Line from 10 to 15 are voiced sound, transformed with the ゛ and ゜ diacritical symbol (は/ha → ば/ba → ぱ/pa). There is one more mora, to be used inside words, the "small tsu", ッ/っ refers to a silent mora (romanised by repeating the following consonant).

lar process of being appropriated and integrated by regular users of a language. Numerous examples can be found in computer science technical vocabulary, being used "as-is" in French (*shell*, *login*, *OS*, *web*, *cd-rom*, *e-mail*...) even when translation can be easily found (*ligne de commande*, *enregistrement/connexion*, *SE*, *toile*, *disque compact*, *courrier électronique*...). Spotting transliterations is therefore even more interesting since it concerns a vocabulary likely to be missing in regular multilingual dictionaries.

We chose here to focus on Japanese transliterations and introduce some features of the Japanese language in the next part.

3. Characteristics of transliterations in Japanese language

3.1. Japanese writing systems

Japanese language is written using three different sets of symbols (see Kageura (2005), for complete description). Kanjis, namely Chinese symbols, are used for their meanings and can be combined to form plain words, whereas katakana and hiragana are two equivalent phonetic alphabets composed of 46 symbols each (see table 1). Hiragana are used for common words where no kanjis are available or are unknown to the writer (typically for children), for grammatical purpose and at scarce occasions to represent onomatopoeia emitted by human. Katakana is mostly used to represent transliterated terms which give us an easy way to spot them and drastically prune terms comparison process. We should also note that katakana are also frequently used for emphasise (for example, in advertising) and to represent onomatopoeia.

3.2. Origin of Japanese transliterations

Japanese language borrowed word from many languages, especially Asian languages (more often Chinese) and western languages (English, French, German…). Most of Japanese western transliterations have been borrowed to English language (even country names are for the most transliterated using the English pronunciation, for example スペイン/`su-pe-i-n`, standing for Spain). However, some transliterations are issued from other languages:

- from French, for example クロワッサン/`ku-ro-wa-s-sa-n` – *croissant* or エスカルゴ/`e-su-ka-ru-go` – *escargot*, in English *snails*, (cooked one, the name of the animal being カタツムリ/`ka-ta-tsu-mu-ri` – this last example shows that species name are often written using katakana too) ;
- from German, for example レントゲン/`re-n-to-ge-n`, corresponding to *x-rays*, from Wilhelm Röntgen who discovered them
- from other western languages, for example パン/`pa-n` from Portuguese (*bread*).

3.3. Transliteration relations with French language

Even though French to Japanese transliterations are rare, it might still be interesting to try to align them with French vocabulary (Tsuji et al., 2002). Indeed, a lot of French vocabulary is common, or very close to English vocabulary and by extend, to western languages (several terms being cognates or transliterations among those languages), especially concerning specialised technical vocabulary. Therefore, transliteration alignment between French and katakana can give interesting result due to a common *bridge word*. Table 2 shows a set of examples extracted from our corpora. Note that knowing the origin of a transliterated term is not really relevant since bridge terms and French terms are generally cognates, originally from a third common language, mostly Greek and Latin.

However, this can lead to attempt to align transliterations with *faux amis*. As an example, the Japanese term フィルム/`fi-ru-mu` is to be aligned with the English term *film*, which also exists in French although the meaning is slightly different. Whereas in French *film* is generally

Japanese / Romanised	→ Bridge term →	French
インスリン / i-n-su-ri-n	→ insulin →	insuline
ホルモン / ho-ru-mo-n	→ hormone →	hormone
ミネラル / mi-ne-ra-ru	→ mineral →	minéral
ヘモグロビン / he-mo-gu-ro-bi-n	→ hemoglobin →	hémoglobine
ビタミン / bi-ta-mi-n	→ vitamin →	vitamine

Table 2: Example of katakana/French indirect transliterations

used for *movie*, in English it mostly refers to *reel*, which is also the meaning of フィルム/fi-ru-mu. We therefore take cautious to talk about transliteration relation between two term only when both conditions are met: terms are phonetically related and are mutual translations.

On the next part, we will shortly present the comparable corpus and observation concerning transliterations and their importance among corpora.

4. Analysis

4.1. Point of observation

We harvested the Web in order to compile an English-French-Japanese comparable corpus. Documents selected all refer to *diabetes* and *nutrition* and are all of *scientific* discourse ("*expert addressing experts*"; (Pearson, 1998), p. 36). Documents were extracted manually, following search engine results or using PubMed[1] for the English part. Documents were finally converted from HTML or PDF to plain text. We obtained 257,000 words for the French corpus, 235,000 for the Japanese corpus and 1,877,000 words for the English corpus. The Japanese corpus is processed through the Chasen morphological analyser[2], French and English corpora are tokenised to isolate words.

The first observation concern all potential transliterations extracted from the Japanese corpus (see part 4.2.) sorted depending on language alignment possibility criteria. We then try to find corresponding source term in English and French corpora (see 4.3.) and finally take a look at a sample of the vocabulary involved in transliteration found between English and Japanese comparable corpora (see part 4.4.). Our goal here is to show the importance and the relevance of transliteration in specialised French/English and Japanese comparable corpora, in order to use them for bilingual lexicon extraction.

4.2. Starting from Japanese corpora

We extract all potential transliterations from the Japanese corpus, by isolating every sequence of katakana. We only work on Japanese single word and exclude hapax for this part, for they are likely to be unstable. 627 different terms were extracted. Note that, due to issue in PDF to text conversion, some candidates are incorrect and are therefore removed (typically single katakana). We finally obtain 493 potential transliterations (i.e. existing Japanese terms written using katakana), which stand for about 8% of the Japanese part unique vocabulary used in context vectors. We then manually translate them, in French when possible, in English if not. Table 3 summarises statistics and shows some samples concerning every sets. *French only trans.* (resp. *English only trans.*) refers to the amount of transliterations, in the Japanese corpus, that can be aligned with a French term (resp. English term — that is, phonetically related and translation of each other) but not with an English term (resp. French term). On the other hand, *French/English trans.* stands for the amount of transliterations that can be aligned with a French and an English term. Finally, *Adapted* refers to transliterations originally from any language, which can not be aligned with French or English because they have been adapted, generally shorten, such as コンビニ/ko-n-bi-ni referring to *convenient store*.

4.3. Relations with English and French corpora

We found several transliterated term in the Japanese corpus, but can we find relation with other corpora ? To answer this question, starting from the manually translated and sorted list, we seek in French and English corpora if we can find corresponding terms. There are 449 transliterations corresponding to an existing English term in the Japanese corpus (see table 3 – 228 transliterations for English only, 221 for English and French) and 225 transliteration corresponding to an existing French term (221 for English and French, 4 for French only). That means we can, at most, find 449 English terms and 225 French terms in English and French comparable corpora.

Among English corpus, **314** terms can be found (which means, they are actually 314 transliteration relations between the Japanese and English corpora on a maximum of 449 – 26 concerning hapax, 288 concerning words appearing twice or more) whereas, among French corpus, from a set of 225, **140** relations can be found (of which 16 hapax). Those results shows that, not only transliterations appears among isolated corpora, but they also cover a part of the common vocabulary we are trying to extract and provide several links between comparable corpora. Although effectiveness of transliteration in bilingual extraction is yet to be observed, these first observations reveal a good potential of incorporating transliterated elements into bilingual term extraction methods. We now have to check if those links can be useful as anchor points by observing the vocabulary involved in transliteration relations.

4.4. Transliteration vocabulary

This last observation is hard to claim without concrete experiments, however we think it is worth to introduce it. Indeed, numerous Japanese transliterations extracted refers precisely to corpora topics (*diabetes and nutrition*) or domain (medical), or related theme such as *physical activities*, *diet and recipe*, *screening and treatment*, *doctor/patient*

[1] http://www.ncbi.nlm.nih.gov/PubMed/

[2] http://chasen.naist.jp/hiki/ChaSen/

	#occ	%	Examples
French only trans.	4	0.8%	レバー/re-ba-a/*levure*, リール/ri-i-ru/*Lille*
English only trans.	228	46%	ヘルス/he-ru-su/*health*, ダイエット/da-i-e-tto/*diet*
French/English trans.	221	45%	マネジャー/ma-ne-ja-a/*magnesium*, ヒスタミン/hi-su-ta-mi-n/*histamine*
Adapted	12	2%	ビル/bi-ru/*building*, テレビ/te-re-bi/*television*
Not English, not French trans.	5	1%	カリウム/ka-ri-wa-mu/*potassium*
Not transliteration	23	5%	ムカデ/mu-ka-de/*centipede*, カキ/ka-ki/*oyster*

Table 3: Statistics concerning katakana sequences from the Japanese corpora

conversation... Here is a 50 words sample randomly extracted from the 314 transliteration pairs found between English and Japanese corpora.

> fair / **advice** / library / schedule / mini / **case** / **keywords** / **insulin** / follow-up / **peak** / clear / **candy** / **interferon** / score / shopping / **signal** / copy / **isotope** / map / **nano** / curriculum / **science** / hit / venture / speed / **ion** / prior / **alcohol** / **guide** / blend / **symposium** / segment / **virus** / label / **salad** / **cheese** / **energy** / **jogging** / floor / core / **beta** / later / **sausage** / wide / end / member / file / **guidance** / **fiber** / model

We emphasise all word related to the scientific discourse in reviewed papers (such as *keywords*, *signal*, *symposium*...), to the medical discourse (such as *advice*, *case*, *virus*...) or concerning *diabetes and nutrition* as previously detailed. It would be clumsy to draw a conclusion from these fuzzy data, although this is an encouraging clue to support our proposition, and we will have to check this observation through experiment.

5. Conclusion

We highlighted here different features of Japanese transliterations and their importance in specialised corpora. Indeed, we showed that it was a frequent phenomena (numerous transliteration relations between different language corpora) and that the vocabulary concerned by transliteration relation is likely to be relevant. Those observations make us think that transliteration can be efficiently used in the case of bilingual lexicon extraction from specialised comparable corpora. However, several issues need to be circumvent, the first one being the capacity to automatically extract and align transliterations pairs in corpora. Indeed, our first experiments using tools for transliterations detection (Tsuji et al., 2002) raised a lot of noisy results which are hard to integrate in the larger bilingual lexicon extraction process. On the other hand, using known transliteration relations is not straightforward. Several ways are to be explored: transliterations can be used to increase coverage of bilingual resources used in alignment, for SWT, or for compositional translation, which is particularly interesting since many MWT involve transliterations (Daille and Morin, 2008). Transliteration relations can also be used as an independent information to assist alignment of context vectors.

6. References

Lynne Bowker and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routeledge, London/New York.

Béatrice Daille and Emmanuel Morin. 2008. Compositionality and lexical alignment of multi-word terms. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP'08)*, volume 1, pages 95–102.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202.

Kyo Kageura. 2005. 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.

Jennifer Pearson. 1998. *Terms in Context*. John Benjamins publishing company.

Carol Peters and Eugenio Picchi. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Gregory Grefenstette, editor, *Cross-language information retrieval*, chapter 7, pages 81–90. Kluwer Academic Publishers.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL'95)*, pages 320–322, Morristown, NJ, USA.

Michel Simard, George F. Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *CASCON*, pages 1071–1082.

Keita Tsuji, Béatrice Daille, and Kyo Kageura. 2002. 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.