Anchor points for bilingual lexicon extraction from small comparable corpora

Emmanuel Prochasson, Emmanuel Morin, Kyo Kageura

Anchor Points for Bilingual Lexicon Extraction
from Small Comparable Corpora

Emmanuel Prochasson
Emmanuel Morin
LINA CNRS UMR 6241
Université de Nantes
2 rue de la houssinière
44322 Nantes cedex 3
{prochasson-e, morin-e}
@univ-nantes.fr

Kyo Kageura
Graduate School of Education
University of Tokyo
7–3–1 Hongo, Bunkyo-ku
Tokyo 113–0033, Japan
kyo@p.u-tokyo.ac.jp

Abstract

We examine the contribution of reliable elements in French– and English–Japanese alignment from comparable corpora, using transliterated elements and scientific compounds as anchor points among context-vectors of elements to align. We highlight those elements in context-vector normalisation to give them a higher priority in context-vector comparison. We carry out experiments on small comparable corpora to show that those elements can efficiently be used to improve the quality of the alignment.

1 Introduction

We are currently working on French- and English-Japanese term alignment from comparable corpora. Much work has been carried out on bilingual lexicon extraction, used to automatically update linguistic resources. This is especially interesting for specialised vocabulary and is needed by translators since regular bilingual dictionaries can not catch up with the growth of terminology. More specifically since the 90s, studies have focused on extraction from comparable corpora (Rapp, 1999; Fung, 1998). This is partly because there is a lack of parallel corpora, especially for language pairs not involving English. This holds even for language pairs such as French and Japanese, both of which have substantial number of speakers. In contrast, comparable corpora, defined as "sets of texts in different languages that are not translations of each other" (Bowker and Pearson, 2002), are more readily available for wider range of language pairs. It is therefore natural to explore comparable corpora for bilingual term alignment.

For comparable corpora, the standard approach is the lexical context mapping using dictionaries (Rapp, 1999; Fung, 1998; Peters and Picchi, 1998), but the performance is generally lower than the performance using parallel corpora because there are much fewer clues for alignment. For language pairs not involving English, the situation is aggravated by the fact that fewer dictionary resources are available and the coverage of available dictionaries tend to be lower than bilingual dictionaries involving English. Moreover, alignments using small comparable corpora (about 250,000 words in our case) lead to lower quality results than with bigger corpora (several million words) used in many case. We try to circumvent these issues by relying on a particular vocabulary to improve the discriminative strength of the context of the terms we want to characterise and align.

We study two kinds of vocabulary: transliterated units in Japanese (and their English/French matches) and scientific compounds (words built on Latin/Greek roots, in English and French, and their translations into Japanese). These two are important lexical classes in specialised discourse. Indeed, they are relevant regarding corpora topics, automatically identifiable and stable (no polysemy). Thus, we expect that they would contribute to the improvement of the term alignment process, using the direct approach (Fung, 1998; Rapp, 1999).

The remainder of this paper is organised as follows. In Section 2, we describe the process for lexicon alignment and the improvement we propose. Section 3.2 presents anchor points we have chosen, describing their interesting features and the method for their automatic detection. Finally, we present the experiments and a discussion in section 4.

2 Lexicon alignment from comparable corpora

2.1 Direct approach

Our alignment program makes use of the direct approach (Fung, 1998; Rapp, 1999). Figure 1 synthesises the different steps of this process.

Our implementation consists of the following four steps:

1. **Building Context-Vector** For each lexical unit i, we collect all lexical units in its context and count the number of times these lexical units appear in a window of n word around i. We obtain, for each lexical unit i of the source and the target languages, a context-vector v_i which collects the set of co-occurring units j associated with the number of times that j and i occur together.

2. **Normalisation of Context-Vector** In order to identify specific words in the lexical context and to reduce word frequency artifacts, we normalise context-vectors using an association score. Context-vectors therefore record the association pattern of a word and its neighbours.

3. **Translation of the vector** Using a bilingual dictionary, we translate the lexical units of the source context-vector. If the bilingual dictionary provides several translations for a lexical unit, we consider all of them but we weight the different translations by their frequency in the target language.

4. **Selection of similar Context-Vector** For a lexical unit to be translated, we compute the similarity between the translated context-vector and all target vectors through vector distance measures (Manning and Schütze, 1999).

The candidate translations of a lexical unit are the target lexical units closest to the translated context-vector according to vector distance measure.

Association measures Computed at the second step in the alignment process, they give, for every element of a context-vector, the importance of its relation to the head elements of the vector. We use the

	j	$\neg j$
i	$a = occ(i, j)$	$b = occ(i, \neg j)$
$\neg i$	$c = occ(\neg i, j)$	$d = occ(\neg i, \neg j)$

Table 1: Contingency table for terms i and j

Log Likelihood, eq. 1 (Dunning, 1993), computed from a contingency table (see table 1).

$$
\begin{aligned}
\lambda(i,j) = \quad & \\
& a\log(a) + b\log(b) + c\log(c) + d\log(d) \\
& +(a+b+c+d)\log(a+b+c+d) \\
& -(a+b)\log(a+b) - (a+c)\log(a+c) \\
& -(b+d)\log(b+d) - (c+d)\log(c+d) \quad (1)
\end{aligned}
$$

2.2 Results of the direct approach

It is not an easy task to compare results obtained by different published studies, due to the differences between the corpora used (especially concerning the way they were built and their size), but also due to the coverage and relevance of bilingual resources used at the second step of the alignment process. As far as we know to date, there is no reference experiment and no reference set of resources (corpus or dictionary) available.

The results of the direct approach are evaluated on the number of correct candidates, found in the x first candidates output by the alignment process (the Top_X). Rapp (1999) obtains 72% correct results for the Top_1 and 89% for the Top_{10}, working on a 135 million word corpus of English and 163 million of German. He used a bilingual dictionary of 16,380 entries (single word terms – SWT). Chiao and Zweigenbaum (2002), using a medical, French/English corpus composed of 600,000 words (for each part) and a specialised dictionary of 18,437 entries, obtained 20% correctness for Top_1 and 60% for Top_{20}. Those results are much lower than Rapp's, but can be easily explained by the different sizes of the comparable corpora. We focus here on single word terms, but other papers present studies about multi-word term alignment (Daille and Morin, 2005).
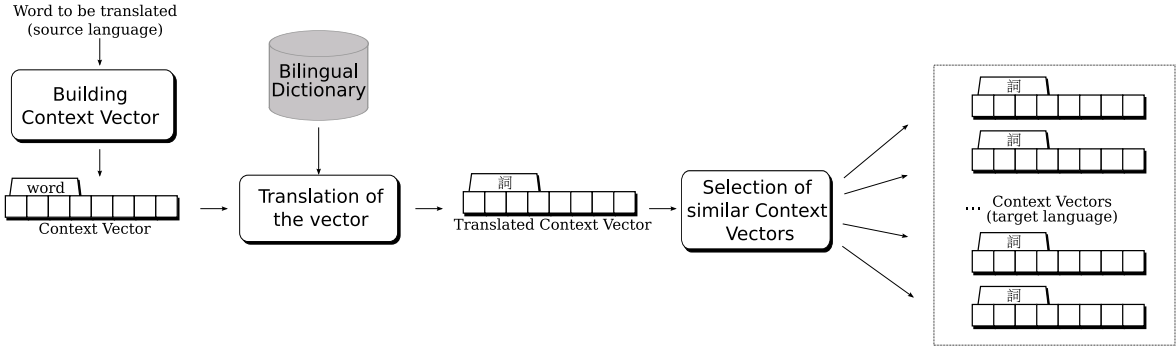
Figure 1: Direct alignment

3 Anchor points in comparable corpora

3.1 Context

We explored the Web in order to compile an English/French/Japanese comparable corpus. Documents selected refer to *diabetes* and *nutrition* and are all of *scientific* discourse ("experts addressing experts"; Pearson (1998), p. 36). Documents were manually extracted, following search engine results or using PubMed[1] for the English part.

We converted those documents into text and cleaned them (manually removing non-informative parts such as *References*, frequent in scientific documents). We obtained 257,000 token words for the French corpus, 235,000 for the Japanese corpus and 250,000 words for the English corpus.

3.2 Specialised vocabulary as anchor points

To be usable in the automatic process of bilingual lexicon extraction, anchor points need to have these three properties :

1. They must be easily identified.

2. They must be relevant, regarding corpora topics.

3. They should not be ambiguous (no polysemy).

We propose the hypothesis that we can rely on those words to improve the discriminative strength of context-vector and therefore improve the quality of results obtained with the direct approach on small corpora. The first property allows us to use them in an automatic process. Second and third properties ensure that those anchor points are relevant, in other word, able to characterise efficiently the specialised terms we are trying to translate. They also ensure that no additional ambiguities are introduced.

Starting from the corpora presented in section 3.1, we observed two classes of vocabulary that satisfy these features. They are Japanese transliterations and English/French scientific compounds.

We call *transliteration* a loan term, from one source languages, that has been adapted to fit the target language speech sound and scripts (by extent, we call *transliteration* the relation between the source term and the target term). (Prochasson et al., 2008) shows that translitteration are proeminent in Japanese language, and that they provide many links between Japanese and other languages, especially concerning English and French. Furthermore, they show that Japanese transliterations reflect specialised vocabulary used in document. Finally, Japanese transliteration are easy to identify, since they are written using a set of symbols mostly dedicated to foreign terms, the *katakanas*. Japanese transliteration are for the most adapted from English, but can be aligned with French term, since French and English share a large common vocabulary. For example, the Japanese term インスリン / i-n-su-ri-n can be aligned with English *insulin* and with French *insuline*.

---

[1] `http://www.ncbi.nlm.nih.gov/PubMed/`, query was "Diabetes Mellitus/diet therapy"[MeSH] OR "Diabetes Mellitus/etiology"[MeSH] OR "Diabetes Mellitus/prevention and control"[MeSH]) **AND** ("nutrition" OR "feeding") with limit to "English language"

We also studied *scientific compounds*. They are word, in French and in English, built with specifics roots (Namer, 2005). (Claveau, 2007), studying automatic translation of medical vocabulary observe that biomedical terms are built on common Greek and Latine roots, and their derivations are consistent. These compounds are characteristic of a specialised vocabulary, especially in medical documents (Lovis et al., 1997; Namer and Zweigenbaum, 2004). Therefore, they seem to be relevant anchor points in the corpora we are using. Moreover, they can easily be identified from their morphology in French and English.

3.3 Improvement

The main idea of this paper is to introduce depth in flat context-vectors, relying on selected terms that are more relevant than others, *i.e.*, anchor points. We highlight those elements in context-vectors, in order to give them more importance when comparing context-vectors (step 3 in section 2.1). Therefore, the context-vector comparison step relies in priority on anchor points, then on other elements.

One way to do so is to dispatch association scores of non-highlighted terms on highlighted terms. That means we lower non-highlighted element scores and give it back to highlighted elements in order to keep a balanced overall score among context-vectors, see equation 2 to 4. In these equations, AP is the set of anchor points used ($|AP|_l$ the number of anchor points found in the context-vector l and $|\neg AP|_l$ the number of other elements), $assoc_j^l$ is the association measure of element j in the context-vector l.

$$assoc\_weighted_j^l := assoc_j^l + \beta, \; if \; j \in AP \quad (2)$$

$$assoc\_weighted_j^l := assoc_j^l - offset_l, \; if \; j \notin AP \quad (3)$$

$$offset_l := \frac{|AP|_l}{|\neg AP|_l} \times \beta \quad (4)$$

The β parameter is used to calibrate the importance given to the highlighted elements. Thus, overall weight (sum of all association scores for all items

of a given vector) is equal before and after balancing. This modification of association measures implies that, if a pair of anchor points (source term and its translation) is to be found between two compared vectors, their similarity score will increase. On the other hand, if an anchor point is to be found in only one of two compared vectors, their similarity score will decrease.

Anchor points must be translations pairs. Indeed, the last step of the direct approach is to compare translated source context-vectors with target context-vectors. If an anchor points is not transfered from source language to target language at the translation step, its discriminative power will be lost at the similarity computation step.

3.4 Anchor points detection

3.4.1 Transliteration detection

We adopted a tool to perform automatic transliteration between English and Japanese language (Tsuji et al., 2005). This tool, based on the Markov chains, gives good results for English/Japanese; it generates a set of potential transliterations for a given katakana or English input. Output is then compared with target vocabulary in the comparable corpora.

Although direct French–Japanese transliterations are quite rare, a lot of English–Japanese transliterations can also be aligned with French vocabulary, due to cognate relations between French and English. We first used a specialised French–Japanese transliteration detection tool but obtained poor results (especially concerning false-positive alignments). We eventually decided to fall back on the tool used for English/Japanese. Before processing, we withdrew every diacritic specific to the French language.

Using this tool, we obtained 589 pairs of transliteration for English/Japanese and 526 for French/Japanese. In order to have valuable anchor points (that is, able to be transfered from one language to another at the translation step of the direct approach), we added detected transliteration to bilingual resources (see Section 4.1).

3.4.2 Identifying scientific compounds

We extracted scientific compounds using a list of 606 medical suffixes and prefixes used in En-

glish[2]. The process is quite simple: we compile regular expressions for every suffix and prefix and have them matched on the bilingual dictionaries used (see section 4.1). Words extracted are kept with their Japanese translation. Such pairs are then used as anchor points in the alignment process. This list, dedicated to the English language can easily be adapted to French (in accordance with the Claveau (2008) observation). We draw our inspiration from this work to write some simple conversion rules. For example, the `-y` suffix in English (as in *psychology*) corresponds to the `-ie` suffix in French (as in *psychologie*). After adapting rules to the French language, we performed the same extraction process than with English on the French dictionary, with the converted list of prefixes/suffixes.

Some suffixes and prefixes are very productive (especially the `a-` prefix) and corresponding extracted terms are not necessarily built from this root. All suffixes and prefixes generating more than 1,000 pairs on bilingual resources were therefore withdrawn. They are however quite rare, only 12 have been discarded for English, and 17 for French.

We obtained 17,210 scientific compounds in English (60,341 pairs of translation, linguistic resources often give more than one translation for a given word) and 8,254 in French (24,240 pairs of translations). The difference comes from the nature of linguistic resources for English and French. When projected onto our corpora vocabulary, we obtained 604 scientific compounds for English (1,197 pairs of translation) and 819 for French (822 pairs of translation).

Unlike transliterations, scientific compounds can not be matched in Japanese using morphological or phonetical clues. That is why they are extracted directly from bilingual resources. That also ensure that extracted scientific compound pairs are translation.

4   Experiments and results

In order to evaluate the influence of anchor points, three kinds of experiments were carried out on English/Japanese and French/Japanese alignments.

(a) direct approach (control experiment)

(b) taking into account automatically detected transliterations

(c) taking into account automatically extracted scientific compounds

All experiments were run on the same set of context-vectors (before normalisation process, which is experiment-dependent), and comparisons were made between results obtained with equivalent parameters (same window size for building context-vectors, same similarity measure and equivalent association measure). We used the Cosine measure (equation 5) for similarity and the Log-Likelihood (equation 1) for the association measure (Dunning, 1993). The *term frequency limit* is set to three for all experiments (it means that a word must appears three times or more in the neighbourhood of a term to be a part of its context-vector). The term lists used for evaluation, introduced in Section 4.1, are the same for all experiments.

4.1   Material

The corpora that we used have already been introduced in Section 3.1. The French-Japanese bilingual dictionary required for the translation phase is composed of four dictionaries freely available from the Web[3], and of a French-Japanese Scientific Dictionary (1989). It contains 173,156 entries, of which 114,461 are single word terms (SWT), with an average of 2.1 translations per entry. We used the JMDict for English/Japanese [4] which is freely available under a Creative Commons (By-SA) licence. We completed it with lists of technical terms from different domains: a list of technical terms compiled by the Japanese Ministry of Education and the National Institute of Informatics (Tokyo)[5] and the *Dictionary of Technical Terms* (Kotani and Kori, 1990). Overall, it contains 589,946 entries (unique words) with an average of 2.3 translations per entry and only 49,208 SWT.

---

[2] http://www.medo.jp/a.htm

[3] http://kanji.free.fr;   http://quebec-japon.com/lexique/index.php?a=index&d=25; http://dico.fj.free.fr/index.php; http://quebec-japon.com/lexique/index.php?a=index&d=3

[4] http://www.csse.monash.edu.au/~jwb/j_jmdict.html

[5] http://sciterm.nii.ac.jp/cgi-bin/reference.cgi

$$Cosine(V_s, V_t) = \frac{\sum_i V_s[i] \times V_t[i]}{\sqrt{\sum_i V_s[i]^2}\sqrt{\sum_j V_t[j]^2}} \tag{5}$$

experiment	a	b	c
En/Jp (Top_1)	17.1%	20.2% [18.2%]	20.2% [18.2%]
En/Jp (Top_{10})	36.3%	39.3% [8.2%]	40.4% [11.2%]
Fr/Jp (Top_1)	20.4%	20.4% [0.0%]	22.4% [10.0%]
Fr/Jp (Top_{10})	36.7%	37.8% [2.8%]	38.8% [5.6%]

Table 2: Alignment results for French/Japanese and English/Japanese ($\beta = 8$)

To evaluate the quality of our method, we built lists of known translations. We selected the most frequent French words ($N_{occ} > 50$) for which a Japanese translation was available. Among those translations, we selected the most frequent Japanese words ($N_{occ} > 50$) in order to obtain a 98 elements test list. We proceeded in the same way with English/Japanese corpora and obtained an 99 elements test list. This protocol for building an evaluation term list is quite similar to the one presented in (Chiao and Zweigenbaum, 2002). They used $N_{occ} > 100$ for the source language, and $N_{occ} > 60$ for the target language, in order to compile a test set of 95 words in an English/French comparable corpus.

4.2 Results

Results shown here are the best that we obtained with the control experiments, compared with other experiments with the same set of parameters. Top_X indicates the number of correct translations found in the X first candidates output by the alignment process. Table 2 shows Top_1 and Top_{10} results for (a), (b) and (c) experiments (improvement between brackets).

Results for control experiment (exp. *a*) are quite similar to those obtained by Chiao and Zweigenbaum (2002), see section 2.2. In the case of English, the improvement when using anchor points is important: it reaches 18.2% when using transliteration (exp. $b - Top_1$) and scientific compounds (exp. $c - Top_1$). The improvement is not that important in the case of French/Japanese alignment. It is null for Top_1 when using transliterations and reaches 10% when using scientific compounds. This can be easily explained by the lowest quality of automatically ex-

tracted anchor points, especially concerning transliterations between Japanese and French.

We think that it is not relevant to combine information brought by transliterations with information brought by scientific compounds. Indeed, those classes are barely related and are taken into account for specific, independent reasons. However, we still ran experiment using both classes as anchor points and observed that the improvement is slightly the same than when using scientific compounds alone.

4.3 Influence of anchor points

We showed that using anchor points can lead to improvement of the direct approach method, for Top 1 and 10. The figure 2 displays the evolution of results, between the control experiment and the experiment using scientific compounds in French/Japanese alignment. This figure shows all correct translations found in both experiments functions of their rank (from Top_1 to Top_{100} – ordinate) and their similarity score (on the abscissa).

In figure 2, hollow triangle indicate translations that were found in control experiment and can not be found with anchor points. On the contrary, black triangle indicate translations found with anchor points that were not found in control experiment. Each thin arrow displays the evolution of a translation found in both experiments. The beginning is the position of the translation in control experiment, the end indicate its position when using anchor points. Finally, thick arrows display the sum of all evolution, for each band indicated by horizontal dot lines (Top 1 to 10, 10 to 20, 20 to 50 and 50 to 100).

Results shown here are interesting : they show that there are pretty much as many missing translations than new translations introduced between both
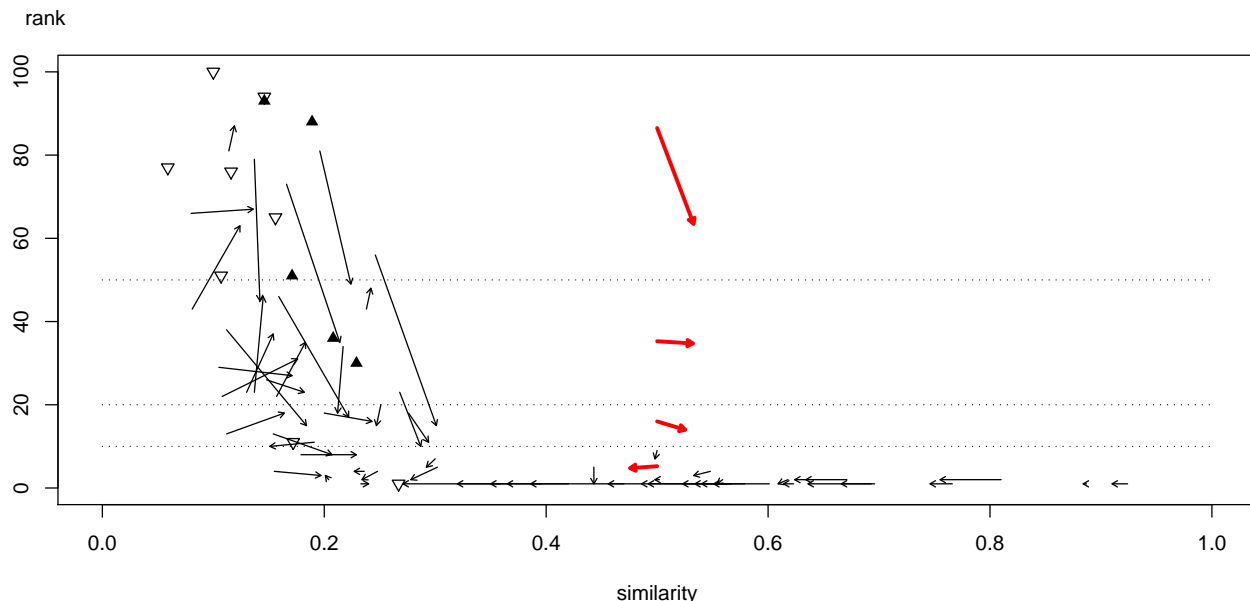
Figure 2: Rank and similarity score of correct translation for French/Japanese alignment, with and without anchor points (scientific compounds $- \beta = 8$.

experiments. Moreover, arrows shows that there is an average improvement of correct translation ranking. This is especially the case for initially badly ranked translations (Top_{50} to Top_{100}). Their rank is highly improved, as indicate the sum of evolution for this band. This observation is available for other band, although it is less impressive. However, initially well-ranked translation are less likely to be improved, but they are not penalised (even though their similarity score decrease).

These observations complete results shown in section 4.2: they show that correct translation candidates are re-ordered to better ranks when using anchor points, even though Top_1 and Top_{10} improvement are not that impressive. We ran a *t-test* (Harris, 1998) on those results. We settle as a null hypothesis that using anchor points does not lead to improvement in ranking correct translation candidates. Results of the *t-test* ($t = 1,8694$; $p = 0,0333$) allows us to reject the null hypothesis with a 95 % confidence (*Wilcoxon* test return a *p-value* of 0.032). Those statistical tests also allow us to reject the null hypothesis in the cases of English/Japanese alignment, using transliterations and scientific compound, but not in the case of French/Japanese alignment using transliterations. This is once again prob-

ably due to the bad quality of automatically detected transliterations.

5 Conclusion

We proceed the issue of bilingual lexicon extraction from comparable corpora, working on small, specialised corpora. We have put forward a new hypothesis. Due to particular features, it was expected that giving particular importance to trusted vocabulary would lead to improvement in the direct alignment process. This hypothesis has been confirmed by experiments: we improved the quality of the lexicon extraction results and showed the influence of anchor points highlighting on those results.

Some work still need to be done. On the one hand, we would like to improve the anchor points detection and characterisation: indeed, the transliteration detection process can be highly improved. On the other hand, the exploitation of anchor points can also be reconsidered: the method we propose here is consistent with the hypothesis, but other methods for taking anchor points into account in the process should be explored, especially concerning the use of different and consistent association measures. Finally, it would be worth to examine the influence of anchor points in context-vectors from a qualitative

point of view, in order to identify new kinds of reliable anchor points.

References

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

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

Vincent Claveau. 2007. Inférence de règles de réécriture pour la traduction de termes biomédicaux. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'07)*, pages 111–120.

Vincent Claveau. 2008. Automatic Translation of Biomedical Terms by Supervised Machine Learning. In *Proceedings of the 6th edition of the Language Resources and Evaluation (LREC'08)*, pages 684–691.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber, and Eduard H. Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–17.

Mary B. Harris. 1998. *Basic Statistics for Behavioral Science Research*. Allyn & Bacon, 2nd edition.

Takuya Kotani and Atsuhiko Kori. 1990. *Dictionary of Technical Terms*. Kenkyusha.

C. Lovis, R. Baud, P. A. Michel, J. R. Scherrer, and A. M. Rassinoux. 1997. Building medical dictionaries for patient encoding systems: A methodology. *Lecture Notes in Computer Science*, 1211:373–380.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, United States of America.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for french medical terminology: contribution of morphosemantics. In Marius Fieschi, Enrico Coiera, and Yu-Chuan Jack Li, editors, *Studies in Health Technology and Informatics*, volume 107, pages 535–539.

Fiammetta Namer. 2005. Morphosémantique pour l'appariement de termes dans le vocabulaire médical: approche multilingue. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, pages 63–72.

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*, pages 81–90. Kluwer Academic Publishers.

Emmanuel Prochasson, Kyo Kageura, Emmanuel Morin, and Akiko Aizawa. 2008. Looking for transliterations in a trilingual english, french and japanese specialised comparable corpus. In *Proceedings of the 1st Workshop on Building and Using Comparable Corpora, Language Resources and Evaluation Conference (LREC'08)*, pages 83–86.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526.

Keita Tsuji, Satoshi Sato, and Kyo Kageura. 2005. Evaluating the effectiveness of transliteration and search engines in bilingual proper name identifications. In *The 11th Annual Meeting of the Association for Natural Language Processing*, pages 352–355.