Bilingual Terminology Mining – Using Brain, not brawn comparable corpora

E. Morin, B. Daille
Université de Nantes
LINA FRE CNRS 2729
2, rue de la Houssinière
BP 92208
F-44322 Nantes Cedex 03
{morin-e,daille-b}@
univ-nantes.fr

K. Takeuchi
Okayama University
3-1-1, Tsushimanaka
Okayama-shi, Okayama,
700-8530, Japan
koichi@
cl.it.okayama-u.ac.jp

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

Abstract

Current research in text mining favours the quantity of texts over their quality. But for bilingual terminology mining, and for many language pairs, large comparable corpora are not available. More importantly, as terms are defined vis-à-vis a specific domain with a restricted register, it is expected that the quality rather than the quantity of the corpus matters more in terminology mining. Our hypothesis, therefore, is that the quality of the corpus is more important than the quantity and ensures the quality of the acquired terminological resources. We show how important the type of discourse is as a characteristic of the comparable corpus.

1 Introduction

Two main approaches exist for compiling corpora: "Big is beautiful" or "Insecurity in large collections". Text mining research commonly adopts the first approach and favors data quantity over quality. This is normally justified on the one hand by the need for large amounts of data in order to make use of statistic or stochastic methods (Manning and Schütze, 1999), and on the other by the lack of operational methods to automatize the building of a corpus answering to selected criteria, such as domain, register, media, style or discourse.

For lexical alignment from comparable corpora, good results on single words can be obtained from large corpora — several millions words — the accuracy of proposed translation is about 80% for the top 10-20 candidates (Fung, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002). (Cao and Li, 2002) have achieved 91% accuracy for the top three candidates using the Web as a comparable corpus. But for specific domains, and many pairs of languages, such huge corpora are not available. More importantly, as terms are defined vis-à-vis a specific domain with a restricted register, it is expected that the quality rather than the quantity of the corpus matters more in terminology mining. For terminology mining, therefore, our hypothesis is that the quality of the corpora is more important than the quantity and that this ensures the quality of the acquired terminological resources.

Comparable corpora are "sets of texts in different languages, that are not translations of each other" (Bowker and Pearson, 2002, p. 93). The term *comparable* is used to indicate that these texts share some characteristics or features: topic, period, media, author, register (Biber, 1994), discourse... This corpus comparability is discussed by lexical alignment researchers but never demonstrated: it is often reduced to a specific domain, such as the medical (Chiao and Zweigenbaum, 2002) or financial domains (Fung, 1998), or to a register, such as newspaper articles (Fung, 1998). For terminology

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 664–671,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

mining, the comparability of the corpus should be based on the domain or the sub-domaine, but also on the type of discourse. Indeed, discourse acts semantically upon the lexical units. For a defined topic, some terms are specific to one discourse or another. For example, for French, within the sub-domain of obesity in the domain of medicine, we find the term *excès de poids* (overweight) only inside texts sharing a popular science discourse, and the synonym *excès pondéral* (overweight) only in scientific discourse. In order to evaluate how important the discourse criterion is for building bilingual terminological lists, we carried out experiments on French-Japanese comparable corpora in the domain of medicine, more precisely on the topic of diabetes and nutrition, using texts collected from the Web and manually selected and classified into two discourse categories: one contains only scientific documents and the other contains both scientific and popular science documents.

We used a state-of-the-art multilingual terminology mining chain composed of two term extraction programs, one in each language, and an alignment program. The term extraction programs are publicly available and both extract multi-word terms that are more precise and specific to a particular scientific domain than single word terms. The alignment program makes use of the direct context-vector approach (Fung, 1998; Peters and Picchi, 1998; Rapp, 1999) slightly modified to handle both single- and multi-word terms. We evaluated the candidate translations of multi-word terms using a reference list compiled from publicly available resources. We found that taking discourse type into account resulted in candidate translations of a better quality even when the corpus size is reduced by half. Thus, even using a state-of-the-art alignment method well-known as data greedy, we reached the conclusion that the quantity of data is not sufficient to obtain a terminological list of high quality and that a real comparability of corpora is required.

2 Multilingual terminology mining chain

Taking as input a comparable corpora, the multilingual terminology chain outputs a list of single- and multi-word candidate terms along with their candidate translations. Its architecture is summarized in Figure 1 and comprises term extraction and alignment programs.

2.1 Term extraction programs

The terminology extraction programs are available for both French[1] (Daille, 2003) and Japanese[2] (Takeuchi et al., 2004). The terminological units that are extracted are multi-word terms whose syntactic patterns correspond either to a canonical or a variation structure. The patterns are expressed using part-of-speech tags: for French, Brill's POS tagger[3] and the FLEM lemmatiser[4] are utilised, and for Japanese, CHASEN[5]. For French, the main patterns are N N, N Prep N et N Adj and for Japanese, N N, N Suff, Adj N and Pref N. The variants handled are morphological for both languages, syntactical only for French, and compounding only for Japanese. We consider as a morphological variant a morphological modification of one of the components of the base form, as a syntactical variant the insertion of another word into the components of the base form, and as a compounding variant the agglutination of another word to one of the components of the base form. For example, in French, the candidate MWT *sécrétion d'insuline* (insulin secretion) appears in the following forms:

- **base form** of N Prep N pattern: *sécrétion d'insuline* (insulin secretion);
- **inflexional variant**: *sécrétions d'insuline* (insulin secretions);
- **syntactic variant** (insertion inside the base form of a modifier): *sécrétion pancréatique d'insuline* (pancreatic insulin secretion);
- **syntactic variant** (expansion coordination of base form): *secrétion de peptide et d'insuline* (insulin and peptide secretion).

The MWT candidates *secrétion insulinique* (insulin secretion) and *hypersécrétion insulinique* (insulin

[1] http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/ and release LINUX.

[2] http://research.nii.ac.jp/~koichi/study/hotal/

[3] http://www.atilf.fr/winbrill/

[4] http://www.univ-nancy2.fr/pers/namer/

[5] http://chasen.org/$\sim$taku/software/mecab/

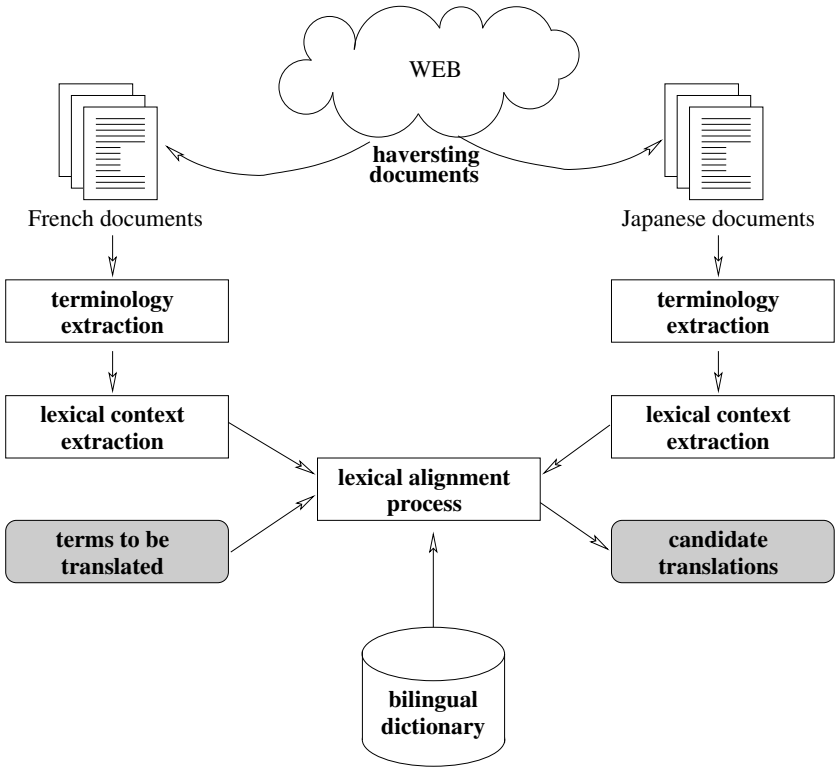


Figure 1: Architecture of the multilingual terminology mining chain

hypersecretion) have also been identified and lead together with *sécrétion d'insuline* (insulin secretion) to a cluster of semantically linked MWTs.

In Japanese, the MWT インスリン.分泌[6] (insulin secretion) appears in the following forms:

- **base form** of NN pattern: インスリン/N_1.分泌/N_2 (insulin secretion);
- **compounding variant** (agglutination of a word at the end of the base form): インスリン/N_1.分泌/N_2.能力/N_3 (insulin secretion ability)

At present, the Japanese term extraction program does not cluster terms.

2.2 Term alignment

The lexical alignment program adapts the direct context-vector approach proposed by (Fung, 1998) for single-word terms (SWTs) to multi-word terms (MWTs). It aligns source MWTs with target single words, SWTs or MWTs. From now on, we will refer to lexical units as words, SWTs or MWTs.

2.2.1 Implementation of the direct context-vector method

Our implementation of the direct context-vector method consists of the following 4 steps:

1. We collect all the lexical units in the context of each lexical unit i and count their occurrence frequency in a window of n words around i. For each lexical unit i of the source and the target language, we obtain a context vector v_i which gathers the set of co-occurrence units j associated with the number of times that j and i occur together occ_j^i. We normalise context vectors using an association score such as Mutual Information or Log-likelihood. In order to reduce the arity of context vectors, we keep only the co-occurrences with the highest association scores.

2. Using a bilingual dictionary, we translate the lexical units of the source context vector.

[6]For all Japanese examples, we explicitly segment the compound into its component parts through the use of the "." symbol.

3. For a word to be translated, we compute the similarity between the translated context vector and all target vectors through vector distance measures such as Cosine (Salton and Lesk, 1968) or Jaccard (Tanimoto, 1958).

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

2.2.2 Translation of lexical units

The translation of the lexical units of the context vectors, which depends on the coverage of the bilingual dictionary vis-à-vis the corpus, is an important step of the direct approach: more elements of the context vector are translated more the context vector will be discrimating for selecting translations in the target language. If the bilingual dictionary provides several translations for a lexical unit, we consider all of them but weight the different translations by their frequency in the target language. If an MWT cannot be directly translated, we generate possible translations by using a compositional method (Grefenstette, 1999). For each element of the MWT found in the bilingual dictionary, we generate all the translated combinations identified by the term extraction program. For example, in the case of the MWT *fatigue chronique* (chronic fatigue), we have the following four translations for *fatigue*: 疲れ, 疲労, 倦怠, 飽き and the following two translations for *chronique*: 記事番組, 慢性. Next, we generate all combinations of translated elements (See Table 1[7]) and select those which refer to an existing MWT in the target language. Here, only one term has been identified by the Japanese terminology extraction program: 慢性.疲労. In this approach, when it is not possible to translate all parts of an MWT, or when the translated combinations are not identified by the term extraction program, the MWT is not taken into account in the translation process.

This approach differs from that used by (Robitaille et al., 2006) for French/Japanese translation. They first decompose the French MWT into combinations of shorter multi-word units (MWU) elements. This approach makes the direct translation of a subpart of the MWT possible if it is present in the bilingual dictionary. For an MWT of length n, (Robitaille et al., 2006) produce all the combinations of MWU elements of a length less than or equal to n. For example, the French term *syndrome de fatigue chronique* (chronic fatigue disease) yields the following four combinations: i) [*syndrome de fatigue chronique*], ii) [*syndrome de fatigue*] [*chronique*], iii) [*syndrome*] [*fatigue chronique*] and iv) [*syndrome*] [*fatigue*] [*chronique*]. We limit ourselves to the combination of type iv) above since 90% of the candidate terms provided by the term extraction process, after clustering, are only composed of two content words.

chronique	*fatigue*
記事番組	疲れ
慢性	疲れ
記事番組	疲労
慢性	疲労
記事番組	倦怠
慢性	倦怠
記事番組	飽き
慢性	飽き

Table 1: Illustration of the compositional method. The underlined Japanese MWT actually exists.

3 Linguistic resources

In this section we outline the different textual resources used for our experiments: the comparable corpora, bilingual dictionary and reference lexicon.

3.1 Comparable corpora

The French and Japanese documents were harvested from the Web by native speakers of each language who are not domain specialists. The texts are from the medical domain, within the sub-domain of diabetes and nutrition. Document harvesting was carried out by a domain-based search, then by manual selection. The search for documents sharing the same domain can be achieved using keywords reflecting the specialized domain: for French, *diabète and obésité* (diabetes and obesity); for Japanese, 糖尿病 and 肥満. Then the documents were classified according to the type of discourse: scientific or popularized science. At present, the selection and classification phases are carried out manually although

[7] the French word order is inverted to take into account the different constraints between French and Japanese.

research into how to automatize these two steps is ongoing. Table 2 shows the main features of the harvested comparable corpora: the number of documents, and the number of words for each language and each type of discourse.

	French		Japanese	
	doc.	words	doc.	words
Scientific	65	425,781	119	234,857
Popular science	183	267,885	419	572,430
Total	248	693,666	538	807,287

Table 2: Comparable corpora statistics

From these documents, we created two comparable corpora:

- [scientific corpora], composed only of scientific documents;
- [mixed corpora], composed of both popular and scientific documents.

3.2 Bilingual dictionary

The French-Japanese bilingual dictionary required for the translation phase is composed of four dictionaries freely available from the Web[8], and of the French-Japanese Scientific Dictionary (1989). It contains about 173,156 entries (114,461 single words and 58,695 multi words) with an average of 2.1 translations per entry.

3.3 Terminology reference lists

To evaluate the quality of the terminology mining chain, we built two bilingual terminology reference lists which include either SWTs or SMTs and MWTs:

- [lexicon 1] 100 French SWTs of which the translation are Japanese SWTs.
- [lexicon 2] 60 French SWTs and MWTs of which the translation could be Japanese SWTs or MWTs.

These lexicons contains terms that occur at least twice in the scientific corpus, have been identified monolingually by both the French and the Japanese term extraction programs, and are found in either the UMLS[9] thesaurus or in the French part of the *Grand dictionnaire terminologique*[10] in the domain of medicine. These constraints prevented us from obtaining 100 French SWTs and MWTs for lexicon 2. The main reasons for this are the small number of UMLS terms dealing with the sub-domain of diabetes and the great difference between the linguistic structures of French and Japanese terms: French pattern definitions tend to cover more phrasal units while Japanese pattern definitions focus more narrowly on compounds. So, even if monolingually the same percentage of terms are detected in both languages, this does not guarantee a good result in bilingual terminology extraction. For example, the French term *diabète de type 1* (Diabetes mellitus type I) extracted by the French term extraction program and found in UMLS was not extracted by the Japanese term extraction program although it appears frequently in the Japanese corpus (一 型 糖 尿 病).

In bilingual terminology mining from specialized comparable corpora, the terminology reference lists are often composed of a hundred words (180 SWTs in (Déjean and Gaussier, 2002) and 97 SWTs in (Chiao and Zweigenbaum, 2002)).

4 Experiments

In order to evaluate the influence of discourse type on the quality of bilingual terminology extraction, two experiments were carried out. Since the main studies relating to bilingual lexicon extraction from comparable corpora concentrate on finding translation candidates for SWTs, we first perform an experiment using [lexicon 1], which is composed of SWTs. In order to evaluate the hypothesis of this study, we then conducted a second experiment using [lexicon 2], which is composed of MWTs.

4.1 Alignment results for [lexicon 1]

Table 3 shows the results obtained. The first three columns indicate the number of translations found

[8] 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

[9] http://www.nlm.nih.gov/research/umls

[10] http://www.granddictionnaire.com/

	NB_{trans}	AVG_{pos}	$STDDEV_{pos}$	TOP_{10}	TOP_{20}
[scientific corpora]	64	11.6	20.2	49	52
[mixed corpora]	76	11.5	16.3	51	60

Table 3: Bilingual terminology extraction results for [lexicon 1]

	NB_{trans}	AVG_{pos}	$STDDEV_{pos}$	TOP_{10}	TOP_{20}
[scientific corpora]	32	16.1	21.9	18	25
[mixed corpora]	32	23.9	27.6	17	20

Table 4: Bilingual terminology extraction results for [lexicon 2]

(NB_{trans}), and the average (AVG_{pos}) and standard deviation ($STDDEV_{pos}$) positions for the translations in the ranked list of candidate translations. The other two columns indicate the percentage of French terms for which the correct translation was obtained among the top ten and top twenty candidates (TOP_{10}, TOP_{20}).

The results of this experiment (see Table 3) show that the terms belonging to [lexicon 1] were more easily identified in the corpus of scientific and popular documents (51% and 60% respectively for TOP_{10} and TOP_{20}) than in the corpus of scientific documents (49% and 52%). Since [lexicon 1] is composed of SWTs, these terms are not more characteristic of popular discourse than scientific discourse.

The frequency of the terms to be translated is an important factor in the vectorial approach. In fact, the higher the frequency of the term to be translated, the more the associated context vector will be discriminant. Table 5 confirms this hypothesis since the most frequent terms, such as *insuline* (#occ. 364 - *insulin*: インスリン), *obésité* (#occ. 333 - *obesity*: 肥満), and *prévention* (#occ. 120 - *prevention*: 予防), were the best translated.

	[2,10]	**[11,50]**	**[51,100]**	**[101,...]**
fr	3/17	12/29	17/23	28/31
jp	4/26	32/41	14/20	10/13

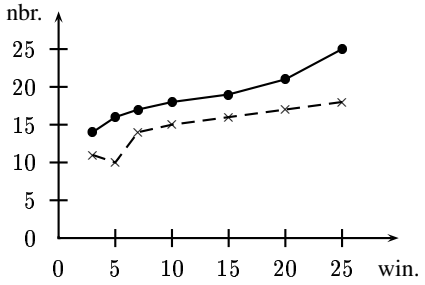
Table 5: Frequency in [corpus 2] of the terms translated belonging to [lexicon 1] (for TOP_{20})

As a baseline, (Déjean et al., 2002) obtain 43% and 51% for the first 10 and 20 candidates respectively in a 100,000-word medical corpus, and 79% and 84% in a multi-domain 8 million-word corpus. For single-item French-English words applied on a medical corpus of 0.66 million words, (Chiao and Zweigenbaum, 2002) obtained 61% and 94% precision on the top-10 and top-20 candidates. In our case, we obtained 51% and 60% precision for the top 10 and 20 candidates in a 1.5 million-word French/Japanese corpus.

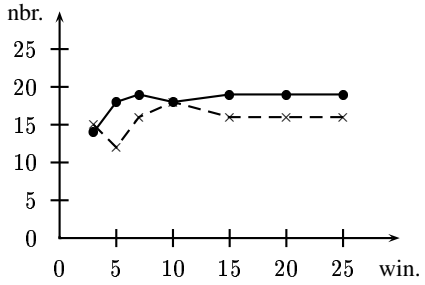
4.2 Alignment results for [lexicon 2]

The analysis results in table 4 indicate only a small number of the terms in [lexicon 2] were found. Since we work with small-size corpora, this result is not surprising. Because multi-word terms are more specific than single-word terms, they tend to occur less frequently in a corpus and are more difficult to translate. Here, the terms belonging [lexicon 2] were more accurately identified from the corpus which consists of scientific documents than the corpus which consists of scientific and popular documents. In this instance, we obtained 30% and 42% precision for the top 10 and top 20 candidates in a 0.84 million-word scientific corpus. Moreover, if we count the number of terms which are correctly translated between [scientific corpora] and [mixed corpora], we find the majority of the translated terms with [mixed corpora] in those obtained with [scientific corpora][11] By combining parameters

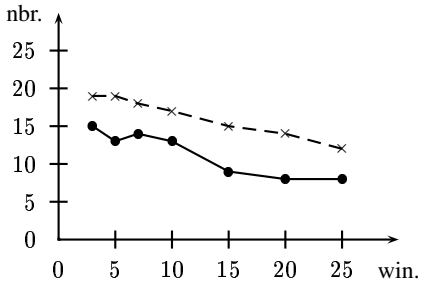
[11]Here, $TOP_{10}18 \cap 17 = 15$, $TOP_{20}25 \cap 20 = 18$ and $NB_{trad}32 \cap 32 = 28$.



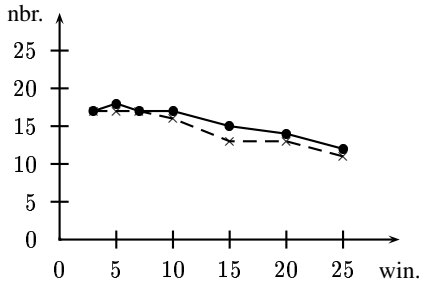
(a) parameter : Log-likelihood & cosinus



(b) parameter : Log-likelihood & jaccard



(c) parameter : MI & cosinus



(d) parameter : MI & jaccard

Figure 2: Evolution of the number of translations found in TOP_{20} according to the size of the contextual window for several combinations of parameters with [lexicon 2] ([scientific corpora] —–; [mixed corpora] - - -, the points indicated are the computed values)

such as the window size of the context vector, association score, and vector distance measure, the terms were often identified with more precision from the corpus consisting of scientific documents than the corpus consisting of scientific and popular documents (see Figure 2).

Here again, the most frequent terms (see Table 6), such as *diabète* (#occ. 899 - *diabetes*: 糖尿.病), *facteur de risque* (#occ. 267 - *risk factor*: 危険.因子), *hyperglycémie* (#occ. 127 - *hyperglycaemia*: 高.血糖), *tissu adipeux* (#occ. 62 - *adipose tissue*: 脂肪.組織) were the best translated. On the other hand, some terms with low frequency, such as *édulcorant* (#occ. 13 - *sweetener*: 甘味.料) and *choix alimentaire* (#occ. 11 - *feeding preferences*: 食品.選択), or very low frequency, such as *obésité massive* (#occ. 6 - *massive obesity*: 高度.肥満), were also identified with this approach.

	[2,10]	**[11,50]**	**[51,100]**	**[101,...]**
fr	1/11	11/25	6/14	7/10
jp	5/21	13/25	5/9	2/5

Table 6: Frequency in [scientific corpora] of translated terms belonging to [lexicon 2] (for TOP_{20})

5 Conclusion

This article describes a first attempt at compiling French-Japanese terminology from comparable corpora taking into account both single- and multi-word terms. Our claim was that a real comparability of the corpora is required to obtain relevant terms of the domain. This comparability should be based not only on the domain and the sub-domain but also on the type of discourse, which acts semantically upon the lexical units. The discourse categorization of documents allows lexical acquisition to increase pre-

cision despite the data sparsity problem that is often encountered for terminology mining and for language pairs not involving the English language, such as French-Japanese. We carried out experiments using two corpora of the specialised domain concerning diabetes and nutrition: one gathering documents from both scientific and popular science discourses, the other limited to scientific discourse. Our alignment results are close to previous works involving the English language, and are of better quality for the scientific corpus despite a corpus size that was reduced by half. The results demonstrate that the more frequent a term and its translation, the better the quality of the alignment will be, but also that the data sparsity problem could be partially solved by using comparable corpora of high quality.

References

Douglas Biber. 1994. Representativeness in corpus design. In A. Zampolli, N. Calzolari, and M. Palmer, editors, *Current Issues in Computational Linguistics: in Honour of Don Walker*, pages 377–407. Pisa: Giardini/Dordrecht: Kluwer.

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

Yunbo Cao and Hang Li. 2002. Base Noun Phrase Translation Using Web Data and the EM Algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 127–133, Tapei, Taiwan.

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, Tapei, Taiwan.

Béatrice Daille. 2003. Terminology Mining. In Maria Teresa Pazienza, editor, *Information Extraction in the Web Era*, pages 29–44. Springer.

Hervé Déjean and Éric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

French-Japanese Scientific Dictionary. 1989. Hakusuisha. 4th edition.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Nonparallel Corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA. Springer.

Gregory Grefenstette. 1999. The Word Wide Web as a Resource for Example-Based Machine Translation Tasks. In *ASLIB'99 Translating and the Computer 21*, London, UK.

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

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.

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, College Park, Maryland, USA.

Xavier Robitaille, Xavier Sasaki, Masatsugu Tonoike, Satoshi Sato, and Satoshi Utsuro. 2006. Compiling French-Japanese Terminologies from the Web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 225–232, Trento, Italy.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.

Koichi Takeuchi, Kyo Kageura, Béatrice Daille, and Laurent Romary. 2004. Construction of grammar based term extraction model for japanese. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *Proceeding of the COLING 2004, 3rd International Workshop on Computational Terminology (COMPUTERM'04)*, pages 91–94, Geneva, Switzerland.

T. T. Tanimoto. 1958. An elementary mathematical theory of classification. Technical report, IBM Research.