Effective Compositional Model for Lexical Alignment

Béatrice Daille, Emmanuel Morin

Effective Compositional Model for Lexical Alignment

Béatrice Daille **Emmanuel Morin**
Université de Nantes, LINA - FRE CNRS 2729
2, rue de la Houssinière, BP 92208
F-44322 Nantes cedex 03
{beatrice.daille,emmanuel.morin}@univ-nantes.fr

Abstract

The automatic compilation of bilingual dictionaries from comparable corpora has been successful for single-word terms (SWTs), but remains disappointed for multi-word terms (MWTs). The increase of coverage of bilingual dictionary thanks to compositional translation improved the results, but still shows some limits for MWTs of different syntactic structures. In this paper, we propose to bridge the gap between syntactic structures through morphological links. The results show a significant improvement in the compositional translation of MWTs that demonstrate the efficiency of the morphologically based-method for lexical alignment.

1 Introduction

Current research in automatic compilation of bilingual dictionaries from corpora makes use of comparable corpora. Comparable corpora gather texts sharing common features (domain, topic, genre, discourse) without having a source text-target text relationship. They are considered by human translators more trustable than parallel corpora (Bowker and Pearson, 2002). Moreover, they are available for any written languages and not only for pair of languages involving English. The compilation of specialized dictionaries should take into account multiword terms (MWTs) that are more precise and specific to a particular scientific domain than single-word terms (SWTs). The standard approach is based on lexical context analysis and relies on the simple observation that a SWT or a MWT and its translation tend to appear in the same lexical contexts. Correct results are obtained for SWTs with an accuracy of about 80% for the top 10-20 proposed candidates using large comparable corpora (Fung, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002) or 60% using small comparable corpora (Déjean and Gaussier, 2002). In comparison, the results obtained for MWTs are disappointed. For instance, (Morin et al., 2007) have achieved 30% and 42% precision for the top 10 and top 20 candidates in a 0.84 million-word French-Japanese corpus. These results could be explained by the low frequency of MWTs compare to SWTs, by the lack of parallelism between the source and the target MWT extraction programs, and by the low performance of the alignment program. For SWTs, it proceeds in two steps: a dictionary look-up, and if no direct translation is available, the contextual analysis. For MWTs, an intermediate step is necessary that will propose several translation candidates to compare with the target MWTs. These candidate translations are obtained thanks to a compositional translation method (Melamed, 1997; Grefenstette, 1999) which increases the coverage of the bilingual dictionary. This method shows some limits when MWTs in the source and the target languages do not share the same syntactic patterns.

In this paper, we propose an extended compositional method that bridge the gap between MWTs of different syntactic structures through morphological links. We experiment this method of French-Japanese lexical alignment, using a multilingual terminology mining chain composed of two term ex-

traction programs, one in each language, and an alignment program. The term extraction programs are publicly available and both extract MWTs. The alignment program makes use of the direct context-vector approach (Fung, 1998; Peters and Picchi, 1998; Rapp, 1999). The results show an improvement of 33% in the translation of MWTs that demonstrate the efficiency of the morphologically based-method for lexical alignment.

2 Multilingual terminology mining chain

Taking as input a comparable corpora, the multilingual terminology mining chain outputs a list of single- and multi-word candidate terms along with their candidate translations (see Figure 1). This chain performs a contextual analysis that adapts the direct context-vector approach (Rapp, 1995; Fung and McKeown, 1997) for SWTs to MWTs. It consists of the following five steps:

1. For each language, the documents are cleaned, tokenized, tagged and lemmatized. For French, Brill's POS tagger[1] and the FLEM lemmatiser[2] are used, and for Japanese, ChaSen[3]. We then extract the MWTs and their variations using the ACABIT terminology extraction program available for French[4] (Daille, 2003), English and Japanese[5] (Takeuchi et al., 2004). (From now on, we will refer to lexical units as words, SWTs or MWTs).

2. 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 languages, 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. In order to identify specific words in the lexical context and to reduce word-frequency effects, we normalize context vectors using an association score such as Mutual Information (Fano, 1961) or Log-likelihood (Dunning, 1993).

3. 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 weight the different translations by their frequency in the target language.

4. For a lexical unit 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).

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

In this approach, the translation of the lexical units of the context vectors (step 3 of the previous approach), which depends on the coverage of the bilingual dictionary vis-à-vis the corpus, is the most important step: the greater the number of elements translated in the context vector, the more discriminating the context vector in selecting translations in the target language. Since the lexical units refer to SWTs and MWTs, the dictionary must contain many entries which occur in the corpus. For SWTs, combining a general bilingual dictionary with a specialized bilingual dictionary or a multilingual thesaurus to translate context vectors ensures that much of their elements will be translated (Chiao and Zweigenbaum, 2002; Déjean et al., 2002). For a MWT to be translated, steps 3 to 5 could be avoided thanks to a compositional method that will propose several translation candidates to directly compare with the target MWTs identified in step 1. Moreover, the compositional method is useful in step 3 to compensate the bilingual dictionary when the multiword units of the context vector are not directly translated.

3 Default compositional method

In order to increase the coverage of the dictionary for MWTs, that could not be directly translated, we generated possible translations by using a default compositional method (Melamed, 1997; Grefenstette, 1999).

[1] `http://www.atilf.fr/winbrill/`

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

[3] `http://chasen-legacy.sourceforge.jp/`

[4] `http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/` and release for Mandriva Linux.

[5] `http://cl.cs.okayama-u.ac.jp/rsc/jacabit/`

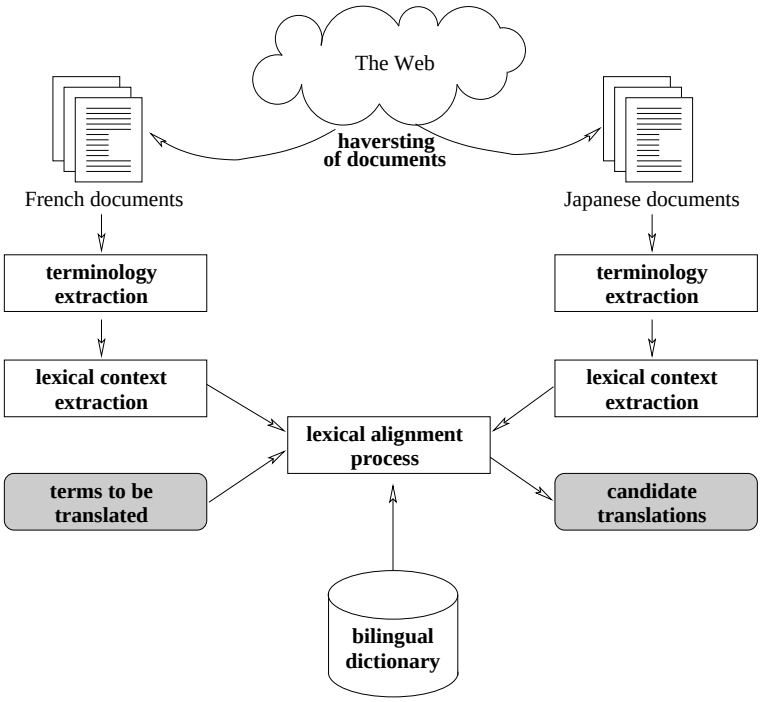

Figure 1: Architecture of the multilingual terminology mining chain

For each element of the MWT found in the bilingual dictionary, we generated all the translated combinations identified by the term extraction program. For example, for the French MWT *fatigue chronique* (*chronic fatigue*), there are four Japanese translations for *fatigue* (*fatigue*) – 疲れ, 疲労, 倦怠, 飽き – and two translations for *chronique* (*chronic*) – 記事番組, 慢性. Next, we generated all possible combinations of the translated elements (see Table 1[6]) and selected those which refer to an existing MWT in the target language. In the above example, only one term for each element was 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 step.

This approach also 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 unit 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 shorter multi-word unit elements of a length less than or equal to n. For example, the French MWT *syndrome de fatigue chronique* (*chronic fatigue disorder*) 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 com-

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)

[6]The French word order is reversed to take into account the different constraints between French and Japanese.

bination of type iv) above since 90% of the French candidate terms provided by the term extraction process after clustering are only composed of two content words.

4 Pattern switching

The compositional translation presents problems which have been reported by (Baldwin and Tanaka, 2004; Brown et al., 1993):

Fertility SWTs and MWTs are not translated by a term of a same length. For instance, the French SWT *hypertension* (*hypertension*) is translated by the Japanese MWT 高血圧 (here the kanji 高 (*taka*) means *high* and the term 血圧 (*ketsuatsu*) means *blood pressure*).

Pattern switching MWTs in the source and the target language do not share the same syntactic patterns. For instance, the French MWT *cellule graisseuse* (*fat cell*) of `N ADJ` structure is translated by the Japanese MWT 脂肪細胞 of `N N` structure where the French noun *cellule* is translated by the Japanese noun 細胞 (*saiboo - cellule - cell*) and the French adjective *graisseuse* by the Japanese noun 脂肪 (*shiboo - graisse - fat*).

Foreign name When a proper name is part of the MWT, it is not always translated: within the French MWT *syndrome de Cushing* (*Cushing syndrome*), Cushing is either transliterated クッシング症候群 or remains unchanged *Cushing*症候群. The foreign name is of course not present in the dictionary.

The pattern switching problem involves the Adjective/Noun and the Noun/Verb part-of-speech switches. The Adjective/Noun switch commonly involves a relational adjective (`ADJR`). According to grammatical tradition, there are two main categories among adjectives: epithetic such as *important* (*significant*) and relational adjectives such as *sanguin* (*blood*). The first ones cannot have an agentive interpretation in contrast to the second: the adjective *sanguin* (*blood*) within the MWT *acidité sanguine* (*blood acidity*) is an argument to the predicative noun *acidité* (*acidity*) and this is not the case for the adjective *important* (*significant*) within the noun phrase *acidité importante* (*significant acidity*). Such adjectives hold a naming function (Levi, 1978) and are particularly frequent in scientific fields (Daille, 2001). Relational adjectives are either denominal adjectives, morphologically derived from a noun thanks to suffix, or adjectives having a noun usage such *mathématique* (*mathematical*/*mathematics*). For the former, it exists appropriate adjective-forming suffixes that lead to relational adjectives such as *-ique*, *-aire*, *-al*. For a noun, it is not possible to guess the adjective-forming suffix that will be employed as well as the alternation of the noun stem that could occur. Relational adjectives part of a MWTs are often translated by a noun whatever is the target language. From French to Japanese, examples are numerous: *prescription médicamenteuse* (処方薬 - *medicinal presciption*), *surveillance glycémique* (血糖管理 - *glycemic monitoring*), *fibre alimentaire* (食物繊維 - *dietary fibre*), *produit laitier* (乳製品 - *dairy product*), *fonction rénale* (腎臓機能 - *kidney function*).

The fertility problem could only be solved thanks to a contextual analysis on the contrary of the foreign name problem that could be solved by an heuristic. We decide to concentrate on the MWT pattern switching problem.

5 Morphologically-based compositional method

When it is not possible to directly translated a MWT — i.e. i) before performing the steps 3 to 5 of the contextual analysis for a multi-word term to be translated or ii) during step 3 for translation of multi-word units of the context vector —, we try first to translate the MWT using the default compositional method. If the default compositional method fails, we use a morphologically-based compositional method. For each MWT of `N ADJ` structure, we generate candidate MWTs of `N Prep N` structure thanks to the rewriting rule:

$$\begin{aligned}
&\mathrm{N}_1\ \mathrm{ADJ} \rightarrow \mathrm{N}_1\ \mathrm{Prep}\ \mathrm{Art}^?\ \mathcal{M}(\mathrm{ADJ}, \mathrm{N}_2)\\
&\mathcal{M}(\mathrm{ADJ}, \mathrm{N}_2) = [-\mathrm{ique}, -\mathrm{ie}]\\
&\mathcal{M}(\mathrm{ADJ}, \mathrm{N}_2) = [-\mathrm{ulaire}, -\mathrm{le}]\\
&\mathcal{M}(\mathrm{ADJ}, \mathrm{N}_2) = [-\mathrm{seux},]\\
&\ldots
\end{aligned} \tag{1}$$

$\mathcal{M}(\mathrm{ADJ}, \mathrm{N}_2)$ gathers a relational adjective ADJ

such as *glycém-ique* and the noun N_2 from which the adjective has been derived such as *glycém-ie* thanks to the stripping-recoding rule $[-\text{ique}, -\text{ie}]$. We generate all possible forms of N_2 as matching stripping-recoding rules and keep those that belong to the biligual dictionary such as *glycém-ie*. Thus, we have created a morphological link between the MWT *contrôle glycémique* (*glycemic control*) of `N ADJ` structure and the multi-word units (MWU) of `N Prep N` structure *contrôle de la glycémie* (lit. *control of glycemia*). Since it has not been possible to translate all the parts of the MWT *contrôle glycémique* as *glycémique* was not found in the dictionary, we use the associated MWT *contrôle de la glycémie* of which all the parts are translated. The generated MWU could be seen as an intermediate lexical form in the translation process that possibly does not exist in the source language. For instance, if *index glycémique* (*glycemic index*) is a French MWT, the MWU *index de la glycémie* (lit. *index of the glycemia*) does not exist in French.

The stripping-recoding rules could be manually encoded, mined from a monolingual corpus using a learning method such as (Mikheev, 1997), or supplied by a source terminology extraction program that handle morphological variations. For such program, a MWT is a canonical form which merge several synonymic variations. For instance, the French MWT *excès pondéral* (*overweight*) could be seen as a canonical form of the following variants: *excès pondéral* (*overweight*) of `N ADJ` structure, *excès de poids* (*overweight*) of `N PREP N` structure. If the pattern switching could only been partially solved as MWT variations are not always attested forms in the corpus, the morphological links could be used to generate stripping-recoding rules. It is this last method that we employ for our experiment.

6 Evaluation

In this section, we outline the different linguistic resources used for our experiments. We then evaluate the performance of the default and morphologically-based compositional methods.

6.1 Linguistic resources

In order to obtain comparable corpora, we selected the French and Japanese documents from the Web. The documents 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 *alimentation*, *diabète* and *obésité* (*food*, *diabetes*, and *obesity*); for Japanese, 糖尿病 and 肥満 (*diabetes*, and *overweight*). Then the documents were manually selected by native speakers of each language who are not domain specialists. These documents (248 for French and 538 for Japanese) were converted into plain text from HTML or PDF, yielding 1.5 million-word corpus (0.7 million-word for French and 0.8 million-word for Japanese).

The French-Japanese bilingual dictionary used in the translation phase was composed of four dictionaries freely available on the Web ([dico 1][7], [dico 2][8], [dico 3][9], and [dico 4][10]), and the French-Japanese Scientific Dictionary (1989) (called [dico 5]). Besides [dico 4] which deals with the medical domain, the other resources are general (as [dico 1, 2, and 3]) or technical (as [dico 5]) dictionaries. Merging the dictionaries yields a single resource with 173,156 entries (114,461 single words and 58,695 multi words) and an average of 2.1 translations per entry.

6.2 French N ADJ reference lists

In order to extract French `N ADJ` reference lists, we proceed as follows:

1. We identify the candidate terms corresponding to `N ADJ` structure in the French corpus using ACABIT.
2. We preserve only the candidate terms whose occur more than 2 times in the French corpus. As a result of filtering, 1,999 candidate terms were extracted.
3. We manually select only those corresponding to a correct term. Here, 360 candidate terms were removed, mainly some misspelled terms,

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

English terms, broken terms, or incoherent terms.

4. We take off the terms that are translated by the bilingual dictionary and found in the comparable corpora. We identified 61 terms of which 30 use a relational adjective such as *vaisseau sanguin* (*blood vessel* - 血管), *produit laitier* (*dairy product* - 乳製品) and *insuffisance cardiaque* (*heart failure* - 心不全).

Finally, we created two French reference lists:

- [N ADJE] composed of 749 terms where ADJE is a epithetic adjective;
- [N ADJR] composed of 829 terms where ADJR is a relational adjective.

6.3 Default compositional method

We first evaluate the quality of the default compositional method for the two French reference lists. Table 2 shows the results obtained. The first three columns indicate the number of French and Japanese terms found, and the number of correct French-Japanese translations.

The results of this experiment show that only a small quantity of terms were translated by the default compositional method. Here, the terms belonging to [N ADJE] were more easily translated (10% with a precision of 69%) than the terms belonging to [N ADJR] (1%). We are unable to generate any translations for 56 (12%) and 227 (27%) terms in the [N ADJE] and [N ADJR] lists, respectively, due to there being no word translations for one or several content words in the dictionary. The best translations for the [N ADJE] list are those where the adjective refers to a quantity such as *faible* (*low*), *moyen* (*medium*), or *haut* (*high*). Since our French-Japanese dictionary contained a small quantity of medical terms, the identified translations for the [N ADJR] list refers to the generic relational adjectives such as *poids normal* (*normal weight* - 正常体重), *étude nationale* (*national study* - 全国調査), or *activité physique* (*physical activity* - 身体活動).

6.4 Morphologically-based compositional method

We now turn to the evaluation of the morphologically-based compositional method that are dedicated to the translation of the [N ADJR] list (see Table 4).

	# French terms	# Japanese terms	# correct translations
[N ADJE]	76	98	68
[N ADJR]	8	8	5

Table 2: Production of the default compositional method

By comparison with the previous method, the results of this experiment show that a significant quantity of terms are now translated. Since compositional method can yield several Japanese translations for one French term, we associate 170 Japanese terms to 128 French terms with a high level of precision: 88.2%. Here, we are unable to generate any translations for 136 (16%) terms by comparison with the 227 terms (27%) for the default compositional method.

	# French terms	# Japanese terms	# correct translations
[N ADJR]	128	170	150

Table 4: Production of the morphologically-based compositional method

In Table 3, each French suffix is associated with the number of identify translations. The most productive suffix are *-ique* as *glycémie*/*glycémique* (*glycemia*/*glycemic*), *-al* as *rein*/*rénal* (*kidney*/*renal*), *-el* as *corps*/*corporel* (*body*/*bodily*), and *-aire* as *aliment*/*alimentaire* (*food*/*dietary*).

Finally from 859 terms relative to N ADJR structure, we translate 30 terms (5.1%) by the dictionary, 5 terms (0.6%) by the default compositional method, and 150 terms (17.5%) by the morphologically-based compositional method. It is difficult to find more translations for several reasons: i) some specialized adjectives or nouns are not included in our resources, ii) some terms are not considered by the Japanese extraction program, and iii) some terms are not encountered in the Japanese corpus.

Suffix	# occ.	French term	Japanese term	(English)
-ique	94	*patient diabétique*	糖尿病患者	(*diabetes patient*)
-al	27	*traitement hormonal*	ホルモン療法	(*hormonal therapy*)
-el	18	*trouble nutritionnel*	栄養障害	(*nutritional disorder*)
-aire	15	*cellule musculaire*	筋肉細胞	(*muscular cell*)
-if	5	*apport nutritif*	栄養摂取	(*nutrition intake*)
-euse	4	*cellule graisseuse*	脂肪細胞	(*fat cell*)
-ier	4	*centre hospitalier*	センター病院	(*hospital complex*)
-ien	2	*hormone thyroïdien*	甲状腺ホルモン	(*thyroid hormone*)
-in	1	*lipide sanguin*	血液脂質	(*blood lipid*)

Table 3: Production of relational adjective according to suffix

7 Conclusion and future work

This study investigated the compilation of bilingual terminologies from comparable corpora and shows how to push back the limits of the methods used in alignment program to translate both single and multi- word terms. We proposed an extended compositional method that bridge the gap between MWTs of different syntactic structures through morphological links. We experiment the method on MWTs of `N ADJ` structure involving a relational adjective. By the use of a list of stripping-recoding rules conjugated with a term extraction program, the method is more efficient than the default compositional method. The evaluation proposed at the end of the paper shows that 170 French-Japanese MWTs are extracted with a high precision (88.2%). This increases the coverage of the French-Japanese terminology of MWTs that can be obtained by the bilingual dictionary or the default compositional method.

In this study, we have observed that MWTs are of a different nature in each language: French patterns cover nominal phrases while Japanese patterns focus on morphologically-built compounds. A Japanese nominal phrase is not considered as a term: thus, the Japanese extraction program does not identify カロリーの摂取 (*caloric intake*) as a candidate MWT but カロリー摂取, unlike the French extraction program which does the contrary (*apport calorique - caloric intake*). Since our morphologically-based compositional method associated カロリー摂取 to *apport calorique*, we could get yield the nominal phrase カロリーの摂取 and improve further more lexical alignment.

References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by Machine of Complex Nominals: Getting it Right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.

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

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

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. 2001. Qualitative terminology extraction. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Recent Advances in Computational Terminology*, volume 2 of *Natural Language Processing*, pages 149–166. John Benjamins.

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.

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

Robert M. Fano. 1961. *Transmission of Information: A statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

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

Pascale Fung and Kathleen McKeown. 1997. Finding Terminology Translations from Non-parallel Corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97)*, pages 192–202, Hong Kong, China.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel 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.

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.

Judith Levi. 1978. *The syntax and the semantics of complex nominals*. Academic Press, London.

I. Dan Melamed. 1997. A Word-to-Word Model of Translational Equivalence. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497, Madrid, Spain.

Andrei Mikheev. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

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. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

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, MD, 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, *Proceedings 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.