# Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus

**Pascale Fung**[1]
Computer Science Department
Columbia University
New York, NY 10027
pascale@cs.columbia.edu

**Abstract**

We propose a novel **context heterogeneity** similarity measure between words and their translations in helping to compile bilingual lexicon entries from a non-parallel English-Chinese corpus. Current algorithms for bilingual lexicon compilation rely on occurrence frequencies, length or positional statistics derived from parallel texts. There is little correlation between such statistics of a word and its translation in non-parallel corpora. On the other hand, we suggest that words with productive context in one language translate to words with productive context in another language, and words with rigid context translate into words with rigid context. Context heterogeneity measures how productive the context of a word is in a given domain, independent of its absolute occurrence frequency in the text. Based on this information, we derive statistics of bilingual word pairs from a non-parallel corpus. These statistics can be used to bootstrap a bilingual dictionary compilation algorithm.

## 1 Introduction

Building a domain-specific bilingual lexicon is one significant component in machine translation and machine-aided translation systems. These terms are often not found in standard dictionaries. Human translators, not being experts in every technical or regional domain, cannot produce their translations effectively. Automatic compilation of such a bilingual lexicon in specific domains is therefore highly desirable.

We present an algorithm in finding word correlation statistics for automatic bilingual lexicon compilation from a non-parallel corpus in Chinese and English. Most previous automatic lexicon compilation techniques require a sentence-aligned clean parallel bilingual corpus (Kupiec 1993; Smadja & McKeown 1994; Kumano & Hirakawa 1994; Dagan *et al.* 1993; Wu & Xia 1994). We have previously shown an algorithm which extracts a bilingual lexicon from *noisy* parallel corpus without sentence alignment (Fung & McKeown 1994; Fung 1995). Although bilingual parallel corpora have been available in recent years, they are still relatively few in comparison to the large amount of monolingual text. Acquiring and processing of parallel corpora are usually labour-intensive and time-consuming. More importantly, the existence of a parallel corpus in a particular domain means *some* translator has translated it, therefore, the bilingual lexicon compiled from such a corpus is at best a reverse engineering of the lexicon this translator used. On the other hand, if we can compile a dictionary of domain-specific words from non-parallel corpora of monolingual texts, the results would be much more meaningful and useful.

[1] This research is partly supported by the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology.

As demonstrated in all the bilingual lexicon compilation algorithms, the foremost task is to identify word features which are similar between a word and its translation, yet different between a word and other words which are not its translations. In parallel corpora, this feature could be the positional co-occurrence of a word and its translation in the other language in the same sentences (Kupiec 1993; Smadja & McKeown 1994; Kumano & Hirakawa 1994; Dagan *et al.* 1993; Wu & Xia 1994) or in the same segments (Fung & Church 1994; Fung 1995). In a non-parallel corpus, there is no corresponding sentence or segment pairs, so the co-occurrence feature is not applicable. In Fung & McKeown (1994); Fung (1995), the word feature used was the positional difference vector. Whereas this is more robust than sentence co-occurrence feature, the matching between two positional difference vectors presumes the two texts are rough translations of one anther. Moreover, whereas the occurrence frequency of a word and that of its translation are relatively similar in a parallel corpus, they have little correlation in non-parallel texts. Our task is, therefore, to identify a word feature correlating a pair of words even if they appear in texts which are not translations of each other. This feature should also be language and character set independent, i.e. it should be applicable to pairs of languages very different from each other. We propose that **context heterogeneity** is such a feature.

## 2 A Non-parallel Corpus of Chinese and English

We use parts of the HKUST English-Chinese Bilingual Corpora for our experiments (Wu 1994), consisting of transcriptions of the Hong Kong Legislative Council debates in both English and Chinese. We use the data from 1988-1992, taking the first 73618 sentences from the English text, and the next 73618 sentences from the Chinese text. There are no overlapping sentences between the texts. The topic of these debates varies though is to some extent confined to the same domain, namely the political and social issues of Hong Kong. Although we select the same number of sentences from each language, there are 22147 unique words from English, and only 7942 unique words from Chinese.

## 3 Some Linguistic Characteristics of Chinese

We have chosen Chinese and English as the two languages from which we will build a bilingual dictionary. Since these languages are significantly different, we need to develop an algorithm which does not rely on any similarity between the languages, and which can be readily extended to other language pairs.

It is useful to point out some significant differences between Chinese and English in order to help explain the output of our experiments:

1 Chinese texts have no word delimiters. It is necessary to perform tokenization on the text by using a Chinese tokenizer. Since the tokenizer is not perfect, the word translation extraction process is affected by this preprocessing.

2 Chinese part-of-speech classes are very ambiguous; many words can be both adjective or noun, noun or verb. Many adjectives can also act as adverbs with no morphological change.

3 Chinese words have little or no morphological information. There are no inflections for nouns, adjectives or verbs to indicate gender, number, case, tense or person (Xi 1985). There is no capitalization to indicate the beginning of a sentence.

4 There are very few function words in Chinese compared to other languages, especially to English. Moreover, function words in Chinese are frequently omitted.

5 A vast number of acronyms are employed in Chinese, which means many single words in Chinese can be translated into compound words in English. Hong Kong Chinese use many terms borrowed from classical Chinese which tend to be more concise. The usage of idioms in Chinese is significantly more frequent than in English.

Points 3,4, and 5 contribute to the fact that the Chinese text of our corpus has fewer unique words than in English.

## 4 Context Heterogeneity of a Word

In a non-parallel corpus, a domain-specific term and its translation are used in different sentences in the two texts. Take the example of the word *air* in the English text. Its concordance is shown partly in Table 4. It occurred 176 times. Its translation 空氣 occurred 37 times in the Chinese text and part of its concordance is shown in Table 4. They are used in totally different sentences. Thus, we cannot hope that their occurrence frequencies would correspond to each other in any significant way.

On the other hand, *air*/空氣 are domain-specific words in the text, meaning something we breathe, as opposed to of some kind of ambiance or attitude. They are used *mostly* in similar *contexts*, as shown in the concordances. If we look at the content word preceding *air* in the concordance, and the content word following it, we notice that *air* is not randomly paired with other words. There are a limited number of word bigrams $(x, W)$ and a limited number of word bigrams $(W, y)$ where $W$ is the word *air*; likewise for 空氣. The number of such unique bigrams indicate a degree of heterogeneity of this word in a text in terms of its neighbors.

We define the context heterogeneity vector of a word $W$ to be an ordered pair $(x, y)$ where:

$$
\begin{aligned}
\text{left heterogeneity } x &= \frac{a}{c}; \\
\text{right heterogeneity } y &= \frac{b}{c}; \\
a &= \text{number of different types of tokens} \\
&\quad \text{immediately preceding W in the text;} \\
b &= \text{number of different types of tokens} \\
&\quad \text{immediately following W in the text;} \\
c &= \text{number of occurrences of W in the text;}
\end{aligned}
$$

The context heterogeneity of any function word, such as *the*, would have $x$ and $y$ values very close to one, since it can be preceded or followed by many different words. On the other hand, the $x$ value of the word *am* is small because it always follows the word *I*.

We postulate that the context heterogeneity of a given domain-specific word is more similar to that of its translation in another language than that of an unrelated word in the other language, and that this is a more salient feature than their occurrence frequencies in the two texts.

For example, the context heterogeneity of *air* is $(119/176, 47/176) = (0.676, 0.267)$ and the context heterogeneity of its translation in Chinese, 空氣 is $(29/37, 17/37) = (0.784, 0.459)$. The context heterogeneity of the word 休會/*adjournment*, on the other hand, is $(37/175, 16/175) = (0.211, 0.091)$. Notice that although *air* and 休會 have similar occurrence frequencies, their context heterogeneities have very different

values, indicating that *air* has much more productive context than 休會. On the other hand, 空氣 has more similar context heterogeneity values as those of *air* even though its occurrence frequency in the Chinese text is much lower.

Table 1: Part of the concordance for *air*

| Word position in text 1 | concordance | |
|---|---|---|
| 8754 | people to enjoy fresh | air , exercise , and a complete change of |
| 14329 | , is it possible for room | air - conditioners to be provided |
| 14431 | houses and institutions . I believe that | air - conditioners |
| 20294 | Chicago Expo told people all about | air - conditioning and the 1 9 3 9 Expo in |
| 31780 | likely to be attracted to visit Expo by | air would only aggravate the problem . |
| 86604 | overnment needs to come out of its old | air - tight armour suit which might serve |
| 102837 | the problems of refuse , sewage , polluted | air , noise and chemical |
| 118017 | ociety marching parallel with decline our | air and water and general |
| 118113 | . It will cover whole spectrum pollution : | air , noise , water and wastes. |
| 119421 | KMB is now experimenting with | air - conditioned double - deckers |

Table 2: Part of the concordance for *air* in Chinese

| Word position in text 2 | concordance | |
|---|---|---|
| 32978 | 上 沒有 免費 東西 , 即使 我 們 呼 吸 | 空氣 , 由於 需要 解決 污染 問題 , 也 絕 非 免費 的 |
| 65488 | 減低 了 燃 油 含 硫 量 , 從而 大大 提高 | 空氣 質素 。 這項 措施 旨在 解決 影響 民 居 最 |
| 153687 | 下列 各項 新 措施 : ( a ) 推 出 兩 條 新 | 空氣 調節 特 快 巴士 線 , 來 往 九龍 |
| 202338 | 及 公布 有關 本港 使用 無 鉛 汽 油 後 | 空氣 含 苯 量 資料 , 以及 會否 採取 管制 措施 , |
| 202594 | 環境保護署 目 前 正 進行 測 量 " 周 圍 | 空氣 每月 含 苯 量 " , 作為 空氣污染 監察 程序 |
| 240355 | 一些 令人 鼓舞 成績 : — 大大 減輕 了 | 空氣 污染 程度 ; — 在 實施 新 廢物 |
| 261651 | 電 工程 師 設計 輸 電 管 , 排 水 , 通 風 及 | 空氣 調節 等 系統 。 完成 此等 工程 |
| 284517 | 服 務 建議 。 我 提出 建議 包括 改善 | 空氣 調節 系統 , 以及 與 小 輪 公司 加強 合作 , 推 |
| 284547 | 鼓勵 乘客 使用 渡輪 和 輕鐵 服 務 。 ( 1 ) | 空氣 調節 不足 , 尤 其 是 在 夏 天 |
| 293127 | 國際間 所 採用 規定 來 立例 規定 | 空氣 中 危險 化 學 品 含 量 標準 ? |

## 5 Distance Measure between two Context Heterogeneity Vectors

To measure the similarity between two context heterogeneity vectors, we use simple Euclidean distance $\mathcal{E}$ where :

$$\mathcal{E} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

.

The Euclidean distance between *air* and 空氣 is $0.2205$ whereas the distance between *air* and 休會 is $0.497$. We use the ordered pair based on the assumption that the word order for nouns in English and Chinese are similar most of the times. For example, *air pollution* is translated into 空氣污染.

## 6 Filtering out Function Words in English

There are many function words in English which do not translate into Chinese. This is because in most Asian languages, there are very few function words compared to Indo-European languages. Function words in Chinese or Japanese are frequently omitted. This partly contributes to the fact that there are far fewer Chinese words than English words in two texts of similar lengths.

Since these functions words such as *the, a, of* will affect the context heterogeneity of most nouns in English while giving very little information, we filter them out from the English text. This heuristic greatly increased the context heterogeneity values of many nouns. The list of function words filtered out are *the*, *a*, *an*, *this*, *that*, *of*, *by*, *for*, *in*, *to*. This is by no means a complete list of English function words. More vigorous statistical training methods could probably be developed to find out which function words in English have no Chinese correspondences. However, if one uses context heterogeneity in languages having more function words such as French, it is advisable that filtering be carried out on both texts.

## 7 Experiment 1: Finding Word Translation Candidates

Given the simplicity of our current context heterogeneity measures and the complexity of finding translations from a non-parallel text in which many words will not find their translations, we propose to use context heterogeneity only as a bootstrapping feature in finding a candidate list of translations for a word.

In our first experiment, we hand-compiled a list of 58 word pairs as in Tables 3 and 4 in English and Chinese, and then used 58 by 58 context heterogeneity measures to match them against each other. Note that this list consists of many single character words which have ambiguities in Chinese, English words which should have been part of a compound word, multiple translations of a single word in English, etc. The initial results are revealing as shown by the histograms in Figure 1.

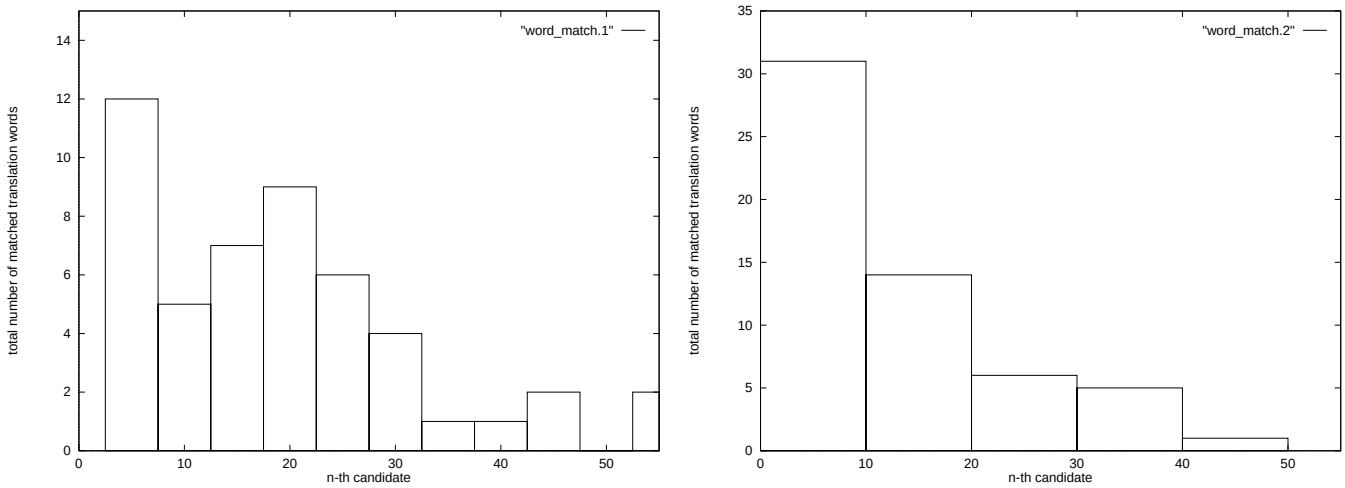


Figure 1: Results of word matching using context heterogeneity

In the left figure, we show that 12 words have their translations in the top 5 candidates. In the right figure, we show the result of filtering out the Chinese genitive 的 from the Chinese texts. In this case, we can see that over 50% of the words found their translation in the top 10 candidates, although it gives fewer words with translations in top 5.

In Sections 7.1 to 7.4, we will discuss the effects of various factors on our results.

Table 3: Test set words - part one

| English word | Chinese word | possible Chinese POS |
|---|---|---|
| Basic | 基本法 | noun |
| British | 英國 | noun-adj |
| CHIM | 詹 | ambiguous |
| CHOW | 周 | ambiguous |
| CHOW | 淑 | ambiguous |
| China | 中國 | noun-adj |
| Committee | 委員會 | noun |
| Council | 局 | ambiguous |
| Declaration | 聲明 | noun-verb |
| Financial | 財政 | noun-adj |
| Government | 政府 | noun-adj |
| Governor | 總督 | noun |
| Hong | 香港 | proper noun |
| Kong | 香港 | proper noun |
| LAM | 林 | ambiguous |
| LAU | 劉 | proper noun |
| Law | 基本法 | noun |
| Ltd | 有限公司 | noun |
| McGREGOR | 覺 | ambiguous |
| Mr | 議員 | noun |
| October | 十月 | noun |
| SECURITY | 保安 | noun-verb |
| Second | 二讀 | noun |
| TAM | 譚 | proper noun |
| TU | 杜 | ambiguous |
| WONG | 黃 | ambiguous |
| YIU | 耀 | ambiguous |

## 7.1 Effect of Chinese Tokenization

We used a statistically augmented Chinese tokenizer for finding word boundaries in the Chinese text (Fung & Wu 1994; Wu & Fung 1994). Chinese tokenization is a difficult problem and tokenizers always have errors. Most single Chinese characters can be joined with other character(s) to form different words. So the translation of a single Chinese character is ill-defined. Moreover, in some cases, our Chinese tokenizer groups frequently co-occurring characters into a single word that does not have independent semantic meanings. For example, 條第/*-th item, number*. In the above cases, the context heterogeneity values of the Chinese

Table 4: Test set words - part two

| English word | Chinese word | possible Chinese POS |
|---|---|---|
| address | 施政報告 | noun |
| air | 空氣 | noun |
| colleagues | 同事 | noun |
| debate | 辯論 | noun-verb |
| decisions | 領導 | noun-verb |
| development | 發展 | noun-verb |
| employers | 僱主 | noun |
| employment | 僱主 | noun |
| expenditure | 開支 | noun-verb |
| figures | 數字 | noun |
| growth | 增長 | noun-verb |
| incidents | 事件 | noun |
| land | 公頃 | quantifier |
| land | 土地 | noun |
| laws | 法例 | noun |
| majority | 大多數 | noun-adj |
| proposals | 建議 | noun-verb |
| prosperity | 繁榮 | noun-adj |
| quality | 素 | ambiguous |
| rate | 率 | ambiguous |
| relationship | 關係 | noun |
| rights | 人權(human rights) | noun |
| risk | 險 | ambiguous |
| safety | 安全 | noun-adj |
| services | 服務 | noun-verb |
| simple | 簡單 | adj |
| step | 步 | ambiguous |
| targets | 目標 | noun |
| tunnels | 隧道 | noun |
| vessels | 船隻 | noun |
| welfare | 社會福利 | noun |
| yesterday | 昨天 | noun |

translation is not reliable. However, translators would recognize this error readily and would not consider it as a translation candidate.

### 7.2 Effect of English Compound Words

As we have mentioned, our Chinese text has many acronyms and idioms which were identified by our tokenizer and grouped into a single word. However, the English text did not under go a collocation extraction process. We can use the following heuristic to overcome the problem:

For a given word $W_i$ in a trigram of $(W_{i-1}, W_i, W_{i+1})$ with context heterogeneity $(x, y)$:

```
1  if W_i(x) = 1
2     W_i(x) ← W_{i-1}(x);
3  if W_i(y) = 1
4     W_i(y) ← W_{i+1}(y);
5  return (W_i(x), W_i(y));
```

Using this method, we have improved the context heterogeneity scores of 人權/*human rights*, 基本法/*Basic Law*, ,二讀/*Second Reading* and 香港/*Hong Kong*.

### 7.3 Effect of Words with Multiple Functions

As mentioned earlier, many Chinese words have multiple part-of-speech tags such as the Chinese for *declaration/declare, development/developing, adjourned/adjournment*, or *expenditure/spend*. Therefore these words have one-to-many mappings with English words.

We could use part-of-speech taggers to label these words with different classes, effectively treating them as different words.

Another way to reduce one-to-many mapping between Chinese and English words could be to use a morphological analyzer in English to map all English words of the same roots with different case, gender, tense, number, capitalization to a single word type.

### 7.4 Effect of Word Order

We had assumed that the trigram word order in Chinese and English are similar. Yet in a non-parallel text, nouns can appear either before a verb or after, as a subject or an object and thus, it is conceivable that we should relax the distance measure to be:

$$\mathcal{E} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (x_1 - y_2)^2 + (y_1 - x_2)^2}$$

.

We applied this measure and indeed improved on the scores for nouns such as *vessels, Government, employers, debate, prosperity*. In some other languages such as French and English, word order for trigrams containing nouns could be reversed most of the time. For example, *air pollution* would be translated into *pollution d'air*. For adjective-noun pairs, Chinese, English and even Japanese share similar orders, whereas French has adjective-noun pairs in the reverse order most of the time. So when we apply context heterogeneity measures to word pairs in English and French, we might map the left heterogeneity in English to the right heterogeneity in French, and vice versa.

## 8 Experiment 2: Finding the Word Translation Among a Cluster of Words

The above experiment showed to some extent the clustering ability of context heterogeneity. To test the discriminative ability of this feature, we choose two clusters of known English and Chinese word pairs *debate*/辯論. We obtained a cluster of Chinese words centered around 辯論 by applying the Kvec segment co-occurrence score (Fung & Church 1994) on the Chinese text with itself. The Kvec algorithm was previously used to find co-occurring bilingual word pairs with many candidates. In our experiment, the co-occurrence happens within the same text, and therefore we got a candidate list for 辯論 that is a cluster of words similar

to it in terms of occurrence measure. This cluster was proposed as a candidate translation list for *debate*. We applied context heterogeneity measures between *debate* and the Chinese word list, with the result shown in Table 5 with the best translation at the top.

Table 5: Sorted candidate list for *debate*

| | | |
|---|---|---|
| 0.117371 | debate | 觸/* |
| 0.149207 | debate | 月十/* |
| 0.155897 | debate | 辯論/debate |
| 0.158305 | debate | 恢復/resumption |
| 0.185699 | debate | 休會/adjournment |
| 0.200486 | debate | 委員會審議階段/Amendment stage of the Council |
| 0.233063 | debate | 月二十/* |
| 0.246826 | debate | 條第/* |
| 0.255721 | debate | 於一/* |
| 0.268771 | debate | 二讀/Second Reading |
| 0.284134 | debate | 條例草案二讀/Second Reading of the Bill |
| 0.312637 | debate | 九九/* |
| 0.315210 | debate | 條例草案二讀動議/moved to Second Reading of the Bill |
| 0.349608 | debate | 委員會審議/Council Amendment |
| 0.367539 | debate | 今午/this afternoon |
| 0.376238 | debate | 這次/this time |
| 0.389296 | debate | 全局/Council |
| 0.389693 | debate | 照會議常規第/* |
| 0.403140 | debate | 獲按/* |
| 0.404000 | debate | 條例草案經過二讀/Second Reading of the Bill passed |

The asterisks in Table 5 indicate tokenizer error. The correct translation is the third candidate. Although we cannot say at this point that this result is significant, it is to some extent encouraging.

It is interesting to note that if we applied the same Kvec algorithm to the English part of the text, we would get a cluster of English words which contain individual translations to some of the words in the Chinese cluster. This shows that co-occurrence measure can give similar clusters of words in different languages from non-parallel texts.

## 9 Non-parallel Corpora Need to be Larger than Parallel Corpora

Among the 58 words we selected, there is one word *service* which occurred 926 times in the English text, but failed to appear even once in the Chinese text (presumably the Legco debate focused more on the issue of various public and legal *services* in Hong Kong during the 1988-90 time frame than later during 1991-92. And in English they frequently accuse each other of paying lip *service* to various issues). We expect there would be a great number of words which simply do not have their translations in the other text. Words which occur very few times also have unreliable context heterogeneity. A logical way to cope with such sparse data problem is to use *larger* non-parallel corpora. Our texts each have about 3 million words, which is much smaller than the parallel Canadian Hansard used for the same purposes. Because it was divided into two parts to form a non-parallel corpus, it is also half in size to the parallel corpus used for word alignment (Wu

& Xia 1994). With a larger corpus, there will be more *source* words in the vocabulary for us to translate, and more *target* candidates to choose from.

## 10 Future Work

We have explained that there are various immediate ways to improve context heterogeneity measures by including more linguistic information about Chinese and English such as word class correspondence and word order correspondence, as well as by using a larger context window. Meanwhile, much larger non-parallel corpora are needed for compilation of bilingual lexicons. We are currently experimenting on using some other similarity measures between word pairs from non-parallel corpora. We plan eventually to incorporate context heterogeneity measures and other word pair similarity measures into bilingual lexicon learning paradigms.

## 11 Conclusion

We have shown the existence of statistical correlations between words and their translations even in a non-parallel corpus. Context heterogeneity is such a correlation feature. We have shown initial results of matching words with their translations in a English-Chinese non-parallel corpus by using context heterogeneity measures. Context heterogeneity can be used both as a clustering measure and a discrimination measure. Given two corresponding clusters of words from the corpus, context heterogeneity could be used to further divide and refine the clusters into few candidate translation words for a given word. Its results can be used to bootstrap or refine a bilingual lexicon compilation algorithm.

## 12 Acknowledgment

I wish to thank Kathleen McKeown and Ken Church for their advice and support, and AT&T Bell Laboratories for use of software and equipments.

## References

DAGAN, IDO, KENNETH W. CHURCH, & WILLIAM A. GALE. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1–8, Columbus, Ohio.

FUNG, PASCALE. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, Boston, Massachusettes. To appear.

FUNG, PASCALE & KENNETH CHURCH. 1994. Kvec: A new approach for aligning parallel texts. In *Proceedings of COLING 94*, 1096–1102, Kyoto, Japan.

FUNG, PASCALE & KATHLEEN MCKEOWN. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 81–88, Columbia, Maryland.

FUNG, PASCALE & DEKAI WU. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the 2nd Annual Workshop on Very Large Corpora*, Kyoto, Japan.

KUMANO, AKIRA & HIDEKI HIRAKAWA. 1994. Building an mt dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th International Conference on Computational Linguistics COLING 94*, 76–81, Kyoto, Japan.

KUPIEC, JULIAN. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, 17–22, Columbus, Ohio.

SMADJA, FRANK & KATHLEEN MCKEOWN. 1994. Translating collocations for use in bilingual lexicons. In *Proceedings of the ARPA Human Language Technology Workshop 94*, Plainsboro, New Jersey.

WU, DEKAI. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 80–87, Las Cruces, New Mexico.

WU, DEKAI & PASCALE FUNG. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 180–181, Stuttgart, Germany.

WU, DEKAI & XUANYIN XIA. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 206–213, Columbia, Maryland.

XI, ZHU DE. 1985. *Yu fa da weng - discussions on linguistics*. Hanyu Zhi Shi Cong Shu. Beijing, China: Shang Wu Yin Shu Guan. In Chinese.