從平行語料庫到計算辭典編纂學：自動建構中英雙語詞組翻譯資料庫

高照明

摘　要

翻譯記憶系統和術語庫對於翻譯從業人員而言並非新的工具。雖然早已有像 Trados 這類的產品可以半自動方式找到原文和譯文句子的對應關係，但截至目前為止，還沒有任何產品可以自動從平行語料擷取詞組的翻譯。為提升電腦輔助翻譯系統的實用性，本文整合計算語言學工具發展出一套可以從中英平行語料自動對應中英詞組翻譯的雛形系統，本文所提出的系統雛形未來可望擴展成為網路新型的翻譯記憶系統，有效提升譯者的工作效率和品質。

關鍵詞：平行語料庫、電腦輔助翻譯、機器翻譯、雙語句對應、雙語關鍵詞上下文檢索程式、雙語詞組對應、翻譯記憶

高照明，國立臺灣大學外國語文學系副教授

From Parallel Corpora to Computational Lexicography: Automatically Constructing a Translation Database of Chinese-English Phrases

Zhao-Ming Gao

Abstract

Translation memory systems and terminology banks are no novel tools to translation practitioners. While there have already been tools such as Trados which can semi-automatically identify sentence correspondences in the source and target languages, no commercial tools exist which can automatically extract translation equivalents at phrase levels. This paper tries to meet translators' real technological needs by integrating state-of-the-art tools in computational linguistics. Given a collection of parallel Chinese-English texts, our system can automatically create a translation database of bilingual phrases. The proposed system will be used as a basis of a web-based translation memory system.

Keywords: parallel corpora, computer-assisted translation, machine translation, sentence alignment, bilingual concordancer, phrase alignment, translation memory

Zhao-Ming Gao, Associate Professor, Department of Foreign Languages and Literatures, National Taiwan University

壹、翻譯科技與譯者的需求

「工欲善其事，必先利其器」。從 30 年前文書處理的功能到現在的搜尋引擎，資訊科技在翻譯中所扮演的角色越形重要。目前譯者常用的翻譯科技有下列幾項：翻譯記憶系統（translation memory systems），術語庫（term banks），機器翻譯系統（machine translation systems），及雙語關鍵詞上下文檢索程式（bilingual concordancers）。上述幾項工具推出已有一段時間，但尚未出現整合的系統，對譯者而言仍有很大的限制。例如 機器翻譯系統雖然效率很高，但輸出結果部份正確，部份錯誤，系統並沒有顯示哪些部分可能有問題，整體的翻譯品質仍然不夠理想。翻譯記憶系統則是根據之前的翻譯資料，找出最接近的翻譯，實用性很高，但是前提是必須要有翻譯資料庫，且必須利用工具找出句對應然後一一檢查並修正對應。換言之，沒有翻譯資料庫的譯者無法利用翻譯記憶系統來增進效率，而就算有翻譯資料庫，譯者也必須花大量時間檢查及修正句子的對應才能使用。術語庫的情形也是一樣，大都需要譯者自行建立，才能使用。雙語關鍵詞上下文檢索程式對譯者的翻譯有很大的幫助，但前提是雙語平行語料庫需要夠大，且句對應的正確性必須高。對於一個譯者而言，一套理想的電腦輔助翻譯工具必須解決上述的問題，並自動提供詞組和術語的翻譯以及例句供譯者參考。數年前 Trados 推出可以半自動找到原文和譯文句子對應關係的產品，但截至目前為止，還沒有任何產品可以自動從平行語料擷取詞組翻譯。本文整合計算語言學工具發展出一套雛形系統，將機器翻譯，術語庫，翻譯記憶，以及語料庫的有機的整合在一起，將成為未來網路新型的翻譯記憶系統的基礎。

貳、機器翻譯

1950 年代電腦問世後幾年，英美因東西方方冷戰的因素開始研究機器翻譯，希望能透過機器翻譯大量翻譯前蘇聯的報章雜誌，以便即時取得情報。當時採取的方法是請語言學家撰寫翻譯的規則，這種規則式的系統通常採取三大步驟，也就是是分析原文（analysis）的詞彙與語法，將詞彙轉換成譯文（transfer），最後再綜合語法及語義形成譯文（synthesis）。規則式的機器翻譯一直沿用數十年。1990 年代初期年 IBM 的研究員 Peter Brown（Brown et al., 1993）與同事發表全新的機器翻譯方法，此種方法完全以語料庫統計的方式自動對譯任何兩種語言，不需任何的規則，開創統計式機器翻譯系統的先河。拜網路及搜尋引擎發明之賜，得以自動收集網路上大量的平行語料並以統計演算法自動進行句對應及詞對應。統計式機器翻譯系統的原理是透過收集大量的語料（包括原文，譯文，及兩種語言句對應的語料）來建立語言模型，並將翻譯的基本問題變成是選詞及詞排序的機率問題。（Brown el al.,1993）所提出的 IBM Models 因為僅利用詞及句對應的機率模型無法產生真正實用的系統，因此歐美日機器翻譯學界

如 Yamada and Knight（2001），Charniak et al.（2003），Melamed（2004）等研究如何將句法結構納入統計式機器翻譯的理論架構。目前譯者可以使用的網路機器翻譯系統以 Google Translate 和微軟的 Bing Translate 為代表，兩者都可以翻譯數十種以上語言。對於固定的翻譯如術語或專有名詞等，通常能達到相當不錯的效果。若不是固定的翻譯，統計式機器翻譯系統則無法得到令人滿意的效果。

參、翻譯記憶系統

翻譯記憶系統（translation memory system）是近年在歐美日廣泛使用的另一項電腦輔助翻譯工具。譯者可以利用此類工具將經常出現的術語或詞組或句子與其翻譯存入資料庫。之後只要文件中出現這些資料，系統即自動將該部分代換成資料庫中所存的翻譯。有別於機器翻譯系統，翻譯記憶系統只是從翻譯資料庫中找出相同或類似的翻譯供使用者參考，而不是電腦自動產生翻譯。相對於機器翻譯系統經常產生一些難以理解也無參考價值的翻譯，電腦輔助翻譯系統的實用性或許更高。只要翻譯資料庫夠大，收集的翻譯品質高，無論是雙語語料檢索程式或翻譯記憶系統都能對譯者產生相當大的助益。然而此兩項工具在國內使用的情形卻不普遍。主要的關鍵在於目前大多數的雙語關鍵詞語境檢索系統與翻譯記憶系統的建立大都靠人工的方式將雙語對應資料存入資料庫。這種方式固然正確率較高但曠日廢時很難在短時間內建立足夠大的資料庫。國外著名的翻譯軟體公司 Trados 利用計算語言學界近年的研究成果發展出能夠將雙語文章段落及句子對應的軟體 WinAlign，人工檢查並修正後即可將雙語對應句存入 Trados 的翻譯記憶系統。當翻譯者使用 Microsoft Word 或 PowerPoint 等軟體翻譯文章時，Trados 翻譯記憶系統會自動將與翻譯記憶系統中相似度百分比超過使用者訂值的句子自動代換成系統裡存的翻譯。譯者再根據情況修改這些翻譯。Trados 句對應的程式 WinAlign 主要依據句長的關連性，對於英漢這兩種不同語系的語言，正確性差很多，只要一個句子對應錯，後面幾乎全錯。因為錯誤率高，人工校對所花的時間頗長，大大影響譯者建立雙語句對應資料庫或翻譯記憶系統的意願。

肆、平行語料庫雙語句對應演算法

語料庫的檢索最常用的是關鍵詞前後文語境檢索（concordancer），也就是輸入一個關鍵詞可以顯示語料庫中包含這個詞的例子。一般的關鍵詞檢索只能提供單一語言的檢索，無法同時列出另一個語言的翻譯。要解決這個問題必須先將每一個句子的翻譯對應句找出來。

雙語句對應的研究開始於 90 年代初期。Gale 與 Church（1991）及 Brown 等（1991）觀察到長句的翻譯對應句一般而言較長，而短句的翻譯句通常較短。他們利用句長的關連性配合動態規劃或 EM 演算法得到 96% 以上的正確率。Gale 與 Church

（1991）及 Brown 等（1991）兩者最大的差別是前者透過人工先得到先驗機率（prior probability）而後者利用 EM 演算法得到相關的參數。Wu（1994）及 Xu and Tan（1996）以句長為主結合一個包含日期及數字等訊息小的辭典得到 96% 的正確率。以句長為基礎的統計方法的優點是不需要語言知識及辭典就可以運作。缺點是如果語料中含有豐富的多對多的句對應關係，或是翻譯的語料中有增添或刪減的現象發生就會造成正確率大幅下降。另一個不需要辭典的方法是 Kay and Röscheisen（1993）以詞彙的頻率（去除低頻的詞及高頻的詞）及在文章中出現的分佈，建立可能的詞對應表及句對應表並不斷的修正，以 relaxation 方法達到收斂。與 Gale 與 Church（1991）及 Brown 等（1991）方法一樣，Kay and Röscheisen（1993）的方法只有在一對一的情形佔絕大多數時才會有好的效果。此外此種方法過度重視詞頻，文章的長度太短會造成正確率的大幅下降，甚至可能找不到對應的詞和句子。Melamed（1997a）提出 Smooth Injective Map Recognizer（SIMR）利用統計和同源詞（cognate），正確率高於 Gale 與 Church（1991），但我們以光華雜誌做初步實驗發現正確率仍然只有 60% 左右。

以統計為主的方法不管是以長度，詞頻及詞彙內部分佈，或 geometric，在正確率及強健性方面似乎都不理想，因此使用雙語辭典似乎是提昇正確率所必需，但如果只以雙語辭典找句對應效果也不理想，原因是翻譯的基本單位在很多時候並不是詞，而是詞組或結構，因此 Catizone et al.（1989）提出結合辭典與統計訊息。Haruno and Yamazaki（1996）比較純統計式，辭典，與混合式三種方法，發現混合式在精確率 precision 召回率 recall 介於 91.6% 到 97.1% 之間，比採純統計式或只用辭典的方式好。Utsuro 等（1994）也採取辭典與句長為主的混合法，但錯誤率介於 4.6% and 21.6%。顯示即使採用混合法正確率也隨著語料的不同與演算法細節的不同而有相當大的差異。

伍、雙語詞組對應演算法

Kaji 等（1992）利用 CKY 剖析演算法剖析英日語對應句再抽取對應的詞組。Matsumoto 等（1993）也用英語日語兩種剖析器分別剖析英日語對應句，再利用特徵結構得到語法依存關係從而得到詞組結構的對應關係。Grishman（1994）與 Kaji 等（1992）Matsumoto 等（1993）也使用兩種語言的剖析器但著眼於如何利用辭典以 bottom-up 的 Iterative 的方式透過結構對應得到更多詞對應。Wu（1995,1997）提出隨機倒置語法 Inversion Transduction Grammars（ITG）的演算法模型 , 與前面研究不同的是不需要利用任何一種語言的剖析器只需要雙語詞對應關係就可以得到兩個語言在某些節點的線性次序是否一致或相反，但此演算法無法得到內部結構的關係。Yarowsky et al.（2001）利用英文和另一個語言的雙語語料的詞對應，再利用現有的英文的詞性標注程式和名詞組辨識程式來找詞組翻譯。Hwa et al.（2005）先利用 Giza++ 這個統計式的詞對應系統得到英文與其它語言的雙語語料的詞對應，接著用英文的剖析器得到句法依存關係，再透過詞對應的結果得到另一個語言的句法依存關係，從而建立那個語

言的句法樹庫及剖析器。Melamed（2004）及 Melamed et al.（2004）改良 Wu（1995, 1997）的架構提出 Generalized Multitext Grammar。Wu（2000a）與 Lu 等人（2001）提出利用一種語言的剖析器來得到另一種語言的語法結構。基本上他們的方法是利用詞彙對應的機率以動態規劃的方式，求出整體中英文配對權值最高的結果。本文的作法類似 Grishman（1994）、Kaji（1992）、Matsumoto 等（1993）也使用兩種語言的剖析器，但與 Grishman（1994）的方法最接近著眼於如何利用辭典以 bottom-up 的 Iterative 的方式透過結構對應得到更多詞對應。

陸、系統架構與流程

與上述文獻主要不同點是我們綜合了大多數的資源，包括辭典及機器翻譯系統，以正確率為主要考量。系統架構如圖 1 所示。輸入雙語平行語料，輸出其中的雙語句對應和詞組對應。

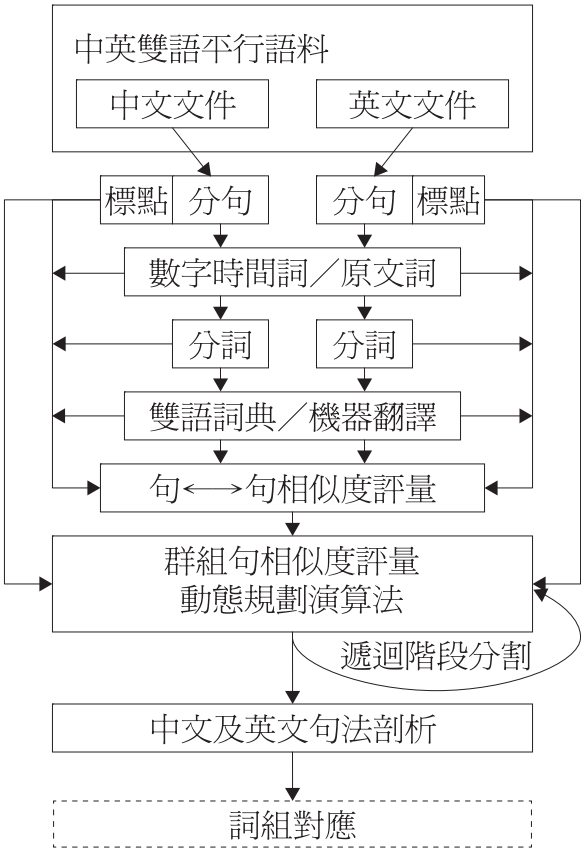

圖 1　系統架構與流程

柒、自動建立中英雙語詞組翻譯資料庫的步驟

1. 以網路代理人程式自動下載中英雙語對應資料。我們希望能取得沒有版權問題且源源不斷的資料來源。符合此兩項條件的中英雙語資料就我們所知只有美國之音（http://www.voa.gov）。我們透過一個網路代理人（agent）程式每天固定時間到美國之音網站下載中英文新聞。
2. 自動找出對應的中英文新聞，這個部份主要是靠辭典查詢和機器翻譯系統，然後找出最接近的中文和英文新聞。
3. 分句。中文以標點符號作為句子為單位。英文分句使用 Shlomo Yona Lingua::EN::Sentence 的 Perl 模組，以解決如 Mr. 等之縮寫與句子句點辨識困難的問題。
4. 分詞。英文方面因為有比較明顯的分詞界線（空白），比較沒有分詞上的困難。中文分詞，我們使用中研院詞庫小組中文斷詞系統。
5. 英文還原成原形。為了利用辭典的查詢，我們透過 WordNet 詞彙知識庫將英文還原成原形。
6. 辭典的查詢。我們以中英文雙向查詢呼叫 pydict 辭典，並將結果存起來。為避免詞彙對應錯誤，中英文的功能詞如 of, the, 的皆略過不查。
7. 找出句對應。我們採用辭典查詢結合 Champollion Tool Kit（CTK）找出中英文對應句。CTK 利用雙語辭典，數字，及簡體中文中的英文詞的對應再加上句長對應的關聯性透過動態規劃找出中英文句對應。
8. 以模糊比對（fuzzy match）找出最多的詞對應，並過濾錯誤的詞對應。模糊比對可以將兩個字串的相似度以百分比表示出來。另外詞的頻率也可以做為重要參考，頻率高的詞，比較容易造成非精確比對的 spurious match 所以權值應該比低頻字的權值低。
9. 英文語法分析。我們使用 Klein and Manning（2003）所發展的 Stanford Parser 來分析英文的句子結構和依存關係（http://nlp.stanford.edu/software/lex-parser.shtml#Download）Stanford Parser 。Stanford Parser 將語法分析的結果以 Penn Treebank Style（http://www.cis.upenn.edu/~treebank/home.html）來表示。在 Penn Treebank Style 中，一個英文句子是用一個樹結構來表示，而樹結構中根節點（root node）、葉節點（leaf node）、內部節點（internal node）則分別表示該英文句子中的各個結構（詞彙或詞組）。
10. 中文的句法分析。我們以支撐向量機（Support Vector Machine）機器學習演算法和中研院句法樹庫（Sinica Treebank）來發展中文的依存關係，並輸出的依存關係。
11. 根據中英文剖析器的結果找出名詞組和不含子句的動詞組。
12. 將中英文的名詞組和動詞組呼叫機器翻譯系統。

13. 利用辭典查詢和機器翻譯的結果，透過 Dice Coefficient 以字串比對的方式計算相似度並自動找出中英文的名詞組和動詞組的對應。

我們的系統實際運作的情形如下。程式每天自動從網路下載中文英對照的新聞。透過程式查詢辭典得到中英詞彙的對應關係，並透過中英句對應程式 CTK 將中英文句子對齊，接著呼叫中英文的剖析器，從剖析的結果找出名詞組和不包含子句的動詞組。再將名詞組和動詞組呼叫微軟機器翻譯系統。透過中英詞彙的對應關係以及機器翻譯的結果，再利用字串比對的功能找出詞組的對應關係。假設下列例 (1) 中文和英文句子已經自動對應好。

(1) Greece and its international lenders continue tense talks on reducing the Greek budget deficit.
希臘及其國際債權人仍然就降低希臘的預算赤字進行著緊張談判

上述英文經過 Stanford Parser 剖析的結果如例 (2)。

(2)

```
(ROOT
(S
 (NP
  (NP (NNP Greece))
  (CC and)
  (NP (PRP$ its) (JJ international) (NNS lenders)))
 (VP (VBP continue)
  (NP (JJ tense) (NNS talks))
  (PP (IN on)
   (S
    (VP(VBG reducing)
     (NP(DT the) (JJ Greek) (NN budget) (NN deficit))))))
 (. .)))
```

我們再根據剖析的結果擷取出名詞組 Greece and its international lenders, its international lenders, tense talks, the Greek budget deficit 和不含子句的動詞組 continue tense talks, reducing the Greek budget deficit。我們的中文剖析器可以判斷句子裡面的語法依存關係。從標示的依存關係可以擷取名詞組和動詞組。例如修飾語（modifier）和名詞（noun）組成名詞組。動詞（V）和受詞（O）組成動詞組。我們根據下面例 (3) 依存關係剖析的結果及詞的位置，從上面的剖析結果可以擷取出所有從小到大的名詞組和動詞組：其債權人、國際債權人、其國際債權人、預算赤字、緊張談判、進行緊張談

判。

(3)

左詞		右詞	關係
希臘	→	及	DUMMY + CAA
其	→	債權人	MODIFIER N
國際	→	債權人	MODIFIER N
仍然	→	降低	ADV V
就	→	降低	ADV V
的	→	赤字	MODIFIER N
預算	→	赤字	MODIFIER N
緊張	→	談判	MODIFIER N
進行	←	緊張談判	V O

上面步驟所擷取出來的中文及英文名詞組和動詞組再呼叫辭典查詢以及微軟或 Google 的機器翻譯系統，利用字串比對，可找出中英名詞組和動詞組的對應。

以名詞組 budget deficit 和 its international lenders 為例呼叫微軟機器翻譯系統分別得到「預算赤字」和「國際貸款機構」。其中「預算赤字」與原來的翻譯完全一致。雖然原來的翻譯與「國際貸款機構」有所出入，但利用模糊比對可以得到最接近的名詞組是「其國際債權人」。字串模糊比對可以用 Dice Cofficient 來計算。

(4) Dice Cofficient ＝ 2 c /（a ＋ b）（Trujillo（1999, 62 頁）

其中 c 是兩句共有的字串長度 ,a 和 b 分別表示兩個字串長度。

以「國際貸款機構」和「其國際債權人」兩個字串為例，以字為單位，兩者的 Dice Cofficient 為（2 ＊ 2）/（6 ＋ 6）＝ 0.333。只要設定門檻值大於 0.3 透過上述的方法仍然可以得到詞組的對應關係。這些對應的詞組自動存到術語庫和詞組翻譯資料庫中。

捌、從平行語料庫到計算辭典學

在翻譯的過程當中不可避免的會遇到一些不熟悉的詞或不知道該如何用外語表達的一些概念，譯者在這種情形下通常求助於辭典。然而一般的辭典只能查詢某一個詞的用法，限於個別詞彙的檢索，很少有能檢索搭配語、詞組甚或句子者。此外辭典受限於篇幅的限制無法收錄大量的例句，凡此皆對譯者造成不便。利用計算語言學技術結合雙語語料庫可以解決譯者翻譯時的問題。在我們的架構下，翻譯記憶系統比對的單位，除了有對應的句子之外，還有對應的詞組。由於程式每天從網路擷取中英雙語新聞並建立句

對應和詞組對應的翻譯資料庫，使用者即使沒有自行建立的翻譯記憶資料仍然有現成的資料庫可以使用。

本文整合計算語言學工具發展出一套比雙語關鍵詞語境檢索更先進的詞組翻譯庫雛形系統，將機器翻譯，術語庫，翻譯記憶，以及平行語料庫有機的整合在一起，讓譯者可以很快速的檢索詞組的翻譯。這樣的工具理論上可以大大提升譯者的工作效率和產量。限於時間因素，本研究僅開發中英詞組對應的雛形系統，尚未評估正確率和實用性。未來希望將此雛形系統擴展成為網路的大型翻譯記憶系統，並提供實證的研究，證明我們的系統可以有效提升翻譯的效率和品質。

玖、致　謝

本研究受到國科會 NSC 95-2411-H-002-045-MY2「中英平行句法樹庫的建立與英漢結構對應演算法」、NSC100-2410-H-002-155「從中英平行語料庫到計算辭典編纂學——語料庫自動擷取英文片語及其中文翻譯與中英例句」、NSC101-2410-H-002-163「結合支持向量機、語料庫統計、與語言規律的混合式中文剖析器」的研究經費補助，中英平行語料的檢索程式由台大資工所碩士沈定先生撰寫，詞組結構對應的雛形程式由台大資工所博士班黃子桓先生及台大資工系姜俊宇、蔡宗翰、邵飛、程至賢四位先生撰寫特此致謝。分詞程式使用中研院詞知識庫發展而成，一併在此致謝。

拾、參考文獻

軟體

Champollion Tool Kit (http://sourceforge.net/projects/champollion/)

Hownet (http://www.keenage.com)

Language::Prolog::Yaswi

http://search.cpan.org/~salva/Language-Prolog-Yaswi-0.14/Yaswi.pm

Lucene Search Engine (http://www.lucene.apache.org)

WordNet http://WordNet.princeton.edu/

WordNet::Similarity http://www.d.umn.edu/~tpederse/similarity.html

WordNet ::QueryData http://people.csail.mit.edu/jrennie/WordNet/

Dagan, I., Church, W, and Gale, W. (1993) “Robust Bilingual Word Alignment for Machine Aided Translation.” In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp. 1-8, Ohio.

Fung, P. and Church, K. (1994) “K-vec: A New Approach for Aligning Parallel Texts.” Proceedings of the International Conference of Computational Linguistics, pp.1096-1102, Kyoto.

Fung, P. and KcKeown, K. (1997) "A Technical Word-and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups." Machine Translation, Vol. 12, Nos. 1-2., pp. 53-87.

Gale, W. and Church, K. (1993) "A Program for Aligning Sentences in Bilingual Corpora." Computational Linguistics, Vol. 19, No 1, pp 75-102.

Gao, Z.-M. (1998) Automatic Extraction of Translation Equivalents from a Parallel Chinese-English Corpus. Ph.d. Thesis. Department of Language Engineering University of Manchester Institute of Science and Technology.

Haruno, M. and Yamazaki, T. (1996) "High-Precision Bilingual Text Alignment Using Statistical and Dictionary Information." Proceedings of Annual Conference of the Association for Computational Linguistics, pp. 131 -138.

Jian, J.-Y., Chang, Y.-C., Chang, J.-S. (2004) Tango: Bilingual Collocational Concordancer. Poster presented at the Annual Conference of the Association for Computational Linguistics.

Johns, T, 1990, 'From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning', CALL Austria 10, pp. 14-34. Revised version in Johns and King (eds.) 1991, pp. 27-45, and with additions and further revisions in

Kay, M. and Roscheisen, M. (1993) "Text-Translation Alignment." Computational Linguistics, Vol. 19, No 1, pp 121-142.

Kilgarriff, Adam et al. (2004) The Sketch Engine. In Proceedings of EURALEX, Lorient, France.

Kumano, A. and Hirakawa, H. (1994) "Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistic Information." in Proceedings of International Conference on Computational Linguistics, pp. 76-81, Kyoto.

Resnik, P., Olsen, M. and Diab, M. (1999) "The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'", Computers and the Humanities, 33(1-2), pp. 129-153.

Smadja, F. and Kathleen, M. and Hatzivassiloglou, V. (1996) "Translating Collocation for Bilingual Lexicons: A Statistical Approach." Computational Linguistics, Vol. 22, No 1, pp. 1-38.

Trujillo, A. (1999) Translation Engines: Techniques for Machine Translation. London: Springer.

Utsuro, T. et al. (1994) "Bilingual Text Matching Using Bilingual Dictionary and Statistics." in Proceedings of International Conference on Computational Linguistics, pp. 1076-1082, Kyoto.

Wu, D. and Xia, X. (1995) "Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon." Machine Translation, Vol. 9, pp. 285-313.