國立臺灣大學文學院翻譯碩士學位學程

碩士論文

Graduate Program in Translation and Interpretation

College of Liberal Arts

National Taiwan University

Master Thesis

法律翻譯語料庫建置及分析

Corpora for Legal Translation: Compilation and Analysis

邱筱涵

Sheau-Harn Chiou

指導教授：高照明博士

Advisor: Dr. Zhao-Ming Gao

中華民國 105年6月

June 2016

doi:10.6342/NTU201603145

Acknowledgements

It has been my extremely good fortune to have been gifted with all the opportunities, guidance, and support that have brought about the existence of this thesis. I am eternally grateful to Professor Zhao-Ming Gao, my thesis advisor, who inspired my initial interest in corpus linguistics and translation technology, who equipped me with the knowledge and tools and skills needed to undertake this study, and whose ever-patient guidance saw me through all the lapses of times that this thesis had stubbornly refused to be written. I am deeply indebted to Professor Shih-Ping Wang, the chair of my oral defense committee, who has gone out of his way to give me invaluable advice, the kindest support, and much-needed pointers on academic writing from the earliest paper version of this thesis up to its final draft. I would like to thank Professor Shan-Shan Wang, who has generously offered wonderful advice and words of encouragement as both my proposal reviewer and oral defense committee member.

Thank you to all my professors at the NTU Graduate Program in Translation and Interpretation and Department of Foreign Languages and Literatures, who have over the years provided me with the training for taking on translation work and studies, who have encouraged me to pursue an M.A. degree, and who have opened so many doors of opportunities for me into the world of translation. Thank you to my classmates at the T&I program; I will always treasure the thoughts, the discoveries, the experiences, and ups and downs we shared in schoolwork, translation jobs, and research efforts.

My greatest thanks to my dearest family and friends, without whose love and support the completion of this thesis would not have been possible. I am thankful for everything that has brought me to where I am, for all the people I have had the honor to meet and learn from, and for all the things that I have learned and gained in this process.

doi:10.6342/NTU201603145

摘要

語料庫語言學及相關技術在翻譯領域中的應用日趨重要，專門語料庫作為特定專業領域的翻譯參考資源，也極具價值。為解決譯者在專門領域的翻譯工作中，可能面臨參考資源不足的問題，本研究嘗試應用現有之軟體資源輔助，建置中英平行結合英文單語之法律語料庫。建置過程採用軟體工具及半自動化方式，進行大批語料處理、自動斷詞、詞性標記、段落（句）對齊，以及詞組對應擷取之工作。建置完成的語料庫，以語料庫軟體輔助，進行關鍵詞、對應詞組、N 連詞、雙語關鍵詞檢索，以及單語關鍵詞檢索之分析。研究結果顯示，本文中嘗試採用的語料庫分析方式，可有效幫助譯者取得多種翻譯過程中需要的參考範例。研究過程中取得的關鍵詞、詞組翻譯、常用表達方式和翻譯策略，也可經累積後應用於其他形式的翻譯資源建置。本研究採用的語料庫建置與分析方式，還可應用在其他專門領域之翻譯，以支援譯者工作需求；分析過程中觀察得的許多現象，也值得進一步分析探索，期能貢獻於未來的翻譯實務與研究工作。

關鍵詞：平行及單語語料、語料庫工具、翻譯參考資源、法律翻譯、法律英語

doi:10.6342/NTU201603145

Abstract

Corpora, the well-organized bodies of “naturally occurring language data” sampled to represent a variety of language (McEnery, 2003, p. 449), have been making a growing impact in the field of translation (Bernardini, Stewart, & Zanettin, 2003). Scholars have asserted the immense value of corpora as reference tool for translation practice in specialized subject domains (Bowker & Pearson, 2002; Varantola, 2003), where intrinsic features of the language may cause difficulties for the translator. To address the potential lacking in reference tools for specialized translation assignments, this study explores a number of methods and computerized tools in compiling and analyzing a parallel and monolingual corpus of Chinese and English legislation. Incorporating semi-automated tools for text processing, part-of-speech tagging, sentence alignment, and phrasal alignment, this study utilizes keyword analysis, n-gram and n-gram part-of-speech sequence, as well as bilingual and monolingual concordance search to address identification of terminology equivalents, stylistic features, usage patterns, and translation strategies for legal contexts. Findings suggest that with the proposed methods, the corpus compiled in this study could effectively provide a number of information to aid the work of legal translators. The information identified can also be applied to compiling other forms of translation resources. It is hoped that in future research, the corpus tools and approaches employed in this study can be applied to facilitating other specialized fields of translation, and that preliminary findings observed here could be further explored to benefit future work in this discipline.

Keywords: parallel and monolingual corpora, corpus analysis tools, translation reference tool, legal translation, legal English

doi:10.6342/NTU201603145

Table of Contents

doi:10.6342/NTU201603145

doi:10.6342/NTU201603145

List of Figures

doi:10.6342/NTU201603145

List of Tables

doi:10.6342/NTU201603145

Chapter 1 Introduction

Corpora are well-organized bodies of “naturally occurring language data” sampled to represent a particular variety of language (McEnery, 2003, p. 449), and they have for some years been making a growing impact in the field of translation. The influence of corpora is reflected in both academic and practice-oriented aspects of the discipline (Bernardini, Stewart, & Zanettin, 2003).

With an aim to explore the benefits that corpus linguistics methods and tools can create for the practicing translator of legal texts, this study employs a variety of corpus-based approaches to compile and analyze a Chinese and English corpus of legislation and statute translations. This chapter introduces the background and motivation of this study (Section 1.1), its purpose and research questions (Section 1.2), and significance of the study (Section 1.3). Section 1.4 provides definitions for the terminology employed, while Section 1.5 outlines the structure of this thesis.

1.1 Background of Study

Technological developments today have brought drastic changes to the translation profession and industry, in which electronic texts, the Internet, and computerized tools play an essential part (Chen, 2012). Translators now have the choice of utilizing a vast array of computational resources, including electronic reference tools and computer-assisted translation (CAT) tools, which create an advantage for translators or language professionals who can master and integrate such “state-of-the-art” tools and software (Bernardini et al., 2003, p. 3).

Application of the above technologies in translation is further introduced in the following subsection, while Subsection 1.1.2 discusses reference tools for translation in the legal genre.

doi:10.6342/NTU201603145

1.1.1 Technologies in Translation

Reference tools or sources are important "information sources for decision making" to the translation process (Varantola, 2003, p. 57). Translators may wish to consult dictionaries, glossaries, and encyclopedic or other sources of domain-specific knowledge for a variety of information, ranging from lexical information, collocation, idiomatic usage, to stylistic information and encyclopedic knowledge, so as to apply them to the appropriate context.

CAT tools, also referred to as computer-aided translation tools or machine-aided translation tools (Quah, 2006), are computerized tools that help human translators work more efficiently (Bowker, 2002) by, in the more narrowly-defined sense, performing some portion of the translation process (Sager, 1994). The two major types of CAT tools, typically included in an integrated "workbench" or "workstation" system along with other CAT tools and resources, are the translation memory system and terminology management system (Quah, 2006).

Translation memory (TM), which are similar to a parallel corpus (Bernardini et al., 2003), can be seen as databases of text segments that allows for storage of previously completed translations and source texts, which are compiled by the user, to be retrieved for reuse where similar source texts occur (Shuttleworth & Lagoudaki, 2006).

The terminology management system (TMS) is a program used for constructing a termbase, which comprises a collection of data records called "term records" containing information on a single concept. A number of associated information could be included in a term record, in addition to the term itself, such as equivalents in another or other languages, grammatical information, synonyms, definition, subject field, and other usage notes (Bowker, 2002).

Varantola (2002) pointed out the necessity of introducing corpora as a new source

doi:10.6342/NTU201603145

of information for translation. Translators spend up to an estimated 50% of the time during their translation process on obtaining lexical information, which goes beyond equivalents to include substantial stretches of contextual information, reassurance, and other non-dictionary information.

Despite the CAT tools now available and the information they may provide, either the TM or TMS mentioned above begin as empty databases. Users must construct a reasonably large amount of new entries, either in their process of translation or from previously completed work, before the tools can contribute to their translation process (Shuttleworth & Lagoudaki, 2006; Bowker, 2002). While shared TMs and termbases may have been created and made available to translators in industries like localization (a process in which business companies customize their products to address foreign language markets), where CAT tools have already become widely employed (Bowker, 2002), the same does not necessarily apply to other specialized fields of translation.

Translators, when working with texts in a specialized subject domain, often fall into the role of non-experts in need of acquiring a “language for special purposes” (LSP), defined in Bowker and Pearson (2002) as the language used for communication on a specialized field of knowledge, typically among experts and semi-experts for facilitating interchanges. Corpora, in electronic format by modern definition, are extremely valuable for learning an LSP, and are an ideal reference tool that complements other sources of reference (Bowker & Pearson, 2002; Varantola, 2003).

Corpus analysis tools, categorized by some scholars as a more broadly-defined type of CAT tools (Bowker, 2002), allow users to manipulate and investigate the contents of LSP corpora through word frequencies (wordlister tool) or search item in contexts (concordancers). Translators can, therefore, utilize LSP corpora as a translation resource for verifying or investigating terminology equivalents, term usage or

doi:10.6342/NTU201603145

collocation, writing style, and even conceptual knowledge. Its advantage in potential extensiveness, ease of update, and convenience for consulting also allows LSP corpora to overcome the constraints of traditional materials in comprehensiveness, physical volume, time-efficiency, and accessibility (Bowker & Pearson, 2002).

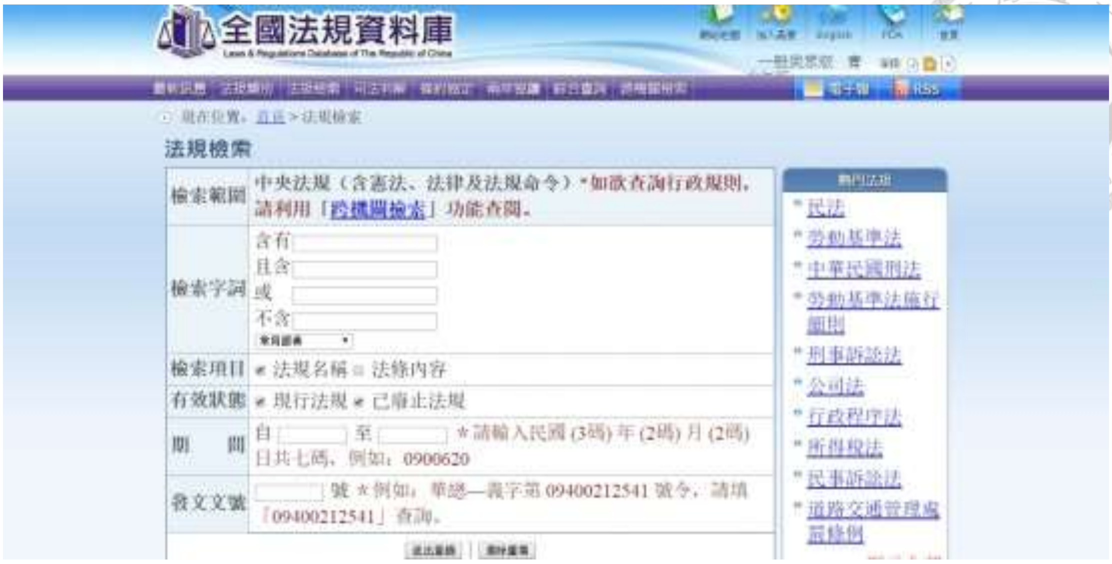
1.1.2 References for Legal Translation

Legal translation, which has long played an essential role in many aspects of international communication and law (Šarčević, 1997), is one such specialized domain that is difficult to master due to the high level of linguistic skills and domain knowledge it requires. With the increasing demand for legal translation as a result of the world-wide globalization trend (Cao, 2007), more translators may find a need to consult additional reference sources in dealing with the singular features intrinsic to the language of law.

One available reference tool for legal translation in Taiwan is the “Laws & Regulations Database of the Republic of China” (http://law.moj.gov.tw/), maintained by the Ministry of Justice. The Database contains texts of the Constitution, laws, regulations, and administrative rules of Taiwan. English translations are available for legislation passed by the parliament (Legislative Yuan), and for regulations and orders that concern foreign nationals, institutions, organizations, or are deemed necessary of translation by relevant authorities (Ministry of Justice [MOJ], 2015). A screenshot of the system interface is shown in Figure 1.1 (on p. 5).

While the system allows for title and content search, however, the texts have not been processed to support corpus analysis; the Chinese statutes and translated texts are simply presented on separate webpages, with a link provided for users to navigate between. For translators, therefore, it would be rather laborious to browse through the potentially relevant statutes, compare between source and target texts, and attempt to

doi:10.6342/NTU201603145



全國法規資料庫
Laws & Regulations Database of The Republic of China

現在位置：首頁 > 法規檢索

法規檢索

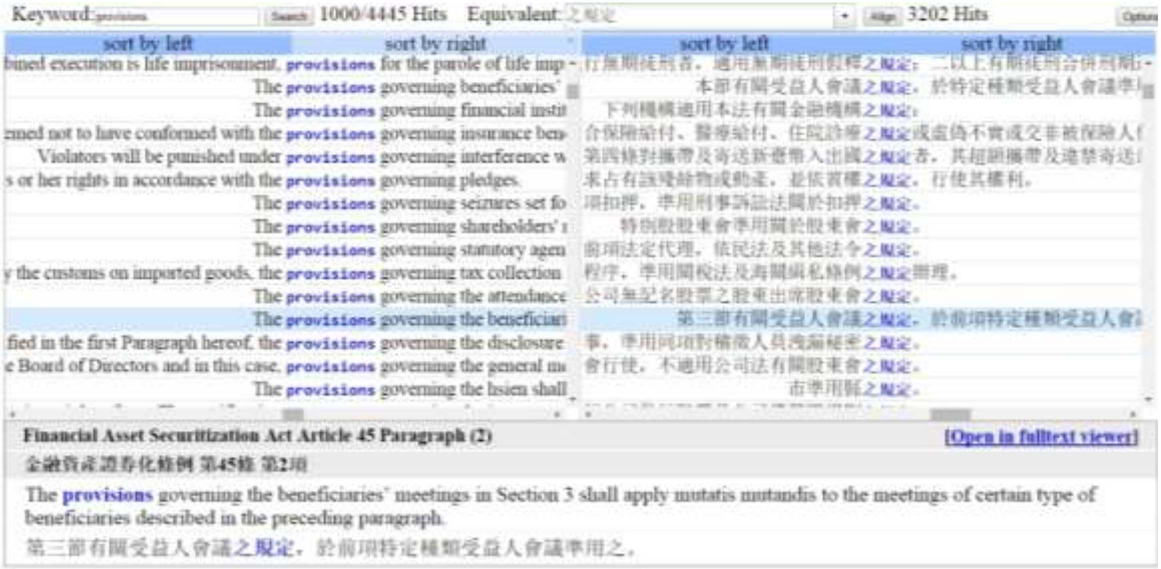
檢索範圍	中央法規（含憲法、法律及法規命令）*如欲查詢行政規則，請利用「跨機關檢索」功能查閱。
檢索字詞	含有 且含 或 不含
檢索項目	☑ 法規名稱 □ 法條內容
有效狀態	☑ 現行法規 ☑ 已廢止法規
期　間	自 至 *請輸入民國(3碼)年(2碼)月(2碼)日共七碼，例如：0900620
發文文號	號 *例如：華總一義字第09400212541號令，請填「09400212541」查詢。

熱門法規
- 民法
- 勞動基準法
- 中華民國刑法
- 勞動基準法施行細則
- 刑事訴訟法
- 公司法
- 行政程序法
- 所得稅法
- 民事訴訟法
- 道路交通管理處罰條例

Figure 1.2. Screenshot of Laws & Regulations Database (http://law.moj.gov.tw/)

identify terminology equivalents or to investigate other linguistic or conceptual information. Also, availability of English translations is largely dependent on many different authorities, which may create constraints for Chinese-English translators.

Another reference tool is the “Bilingual KWIC for Taiwan Laws” (http://kwic.law.nagoya-u.ac.jp/taiwan/, screenshot shown in Figure 1.2), an online parallel corpus and bilingual concordancer developed by the Nagoya University



Keyword: provisions 1000/4445 Hits Equivalent: 之規定 3202 Hits

sort by left	sort by right	sort by left / sort by right
bined execution is life imprisonment,	provisions for the parole of life imp	行無期徒刑者，適用無期徒刑假釋之規定；二以上有期徒刑合併刑期
The	provisions governing beneficiaries'	本節有關受益人會議之規定，於特定種類受益人會議準
The	provisions governing financial instit	下列機構適用本法有關金融機構之規定：
rned not to have conformed with the	provisions governing insurance ben	合保險給付、醫療給付、住院診療之規定或虛偽不實或交非被保險人
Violators will be punished under	provisions governing interference w	第四條對攜帶及寄送新臺幣入出國之規定者，其超額攜帶及違禁寄送
s or her rights in accordance with the	provisions governing pledges.	未占有該質物或動產，並依質權之規定，行使其權利。
The	provisions governing seizures set fo	項扣押，準用刑事訴訟法關於扣押之規定。
The	provisions governing shareholders' r	特別股股東會準用關於股東會之規定。
The	provisions governing statutory agen	前項法定代理，依民法及其他法令之規定。
y the customs on imported goods, the	provisions governing tax collection	程序，準用關稅法及海關緝私條例之規定辦理。
The	provisions governing the attendance	公司無記名股票之股東出席股東會之規定。
The	provisions governing the beneficiari	第三節有關受益人會議之規定，於前項特定種類受益人會
fied in the first Paragraph hereof, the	provisions governing the disclosure	事，準用同項對稽徵人員洩漏秘密之規定。
e Board of Directors and in this case,	provisions governing the general m	會行使，不適用公司法有關股東會之規定。
The	provisions governing the hsien shall	市準用縣之規定。

Financial Asset Securitization Act Article 45 Paragraph (2) [Open in fulltext viewer]

金融資產證券化條例 第45條 第2項

The **provisions** governing the beneficiaries' meetings in Section 3 shall apply mutatis mutandis to the meetings of certain type of beneficiaries described in the preceding paragraph.

第三節有關受益人會議之規定，於前項特定種類受益人會議準用之。

Figure 1.1. Screenshot of Bilingual KWIC for Taiwan Laws (http://kwic.law.nagoya-u.ac.jp/taiwan/)

doi:10.6342/NTU201603145

Graduate School of Law. The corpus comprises the texts and English translations of 156 Taiwan laws and 669 interpretations by justices of the Judicial Yuan, and the concordancer accepts search items in either Chinese or English (Toyama, 2011). Chinese source segments are bound (aligned) to their corresponding English translation segments, which are presented together when search items are identified in one of the segments. The system succeeds in identifying word or phrasal correspondence for some search items, enabling it to function as a bilingual dictionary. Search results are also linked to their full segments and statute texts, available if the user wishes to view them.

The Bilingual KWIC is already a powerful tool for translators who wish to identify possible terminology equivalents, observe translation strategies, or obtain explanatory contexts. The only drawback is perhaps that, since raw corpora are not available to the average user, further or other forms of statistical data cannot be compiled from the same set of corpora. As texts of law in the Bilingual KWIC corpora were selected from the MOJ Laws & Regulations Database (Toyama, 2011), the availability of English translations remains a potential constraint; in the case of the Bilingual KWIC, the selection scope and selection criteria of the corpora are also unknown to the user.

Another issue for translators working into English lies in the question whether translated English texts are sufficient as the sole reference source. Parallel corpora, by definition, consist of source texts in one language and their translations in another language, aligned to each other by corresponding segments (Bowker & Pearson, 2002). Since the Bilingual KWIC does not include any texts translated into Chinese, the corpora may not be appropriate for verifying idiomatic English usages; even equipped with this tool, therefore, translators may still be in need of reference sources that will allow them to examine the legal language as used in non-translational English texts.

doi:10.6342/NTU201603145

1.2 Purpose and Research Questions

To address the translator’s need for reference tools, the purpose of this study is to explore a set of practice-oriented methods that enable translators to effectively assemble and make use of their own corpora for a specialized field of translation.

In the case of Chinese to English legal translation in Taiwan, the two known reference sources introduced in the previous section, while extensive and powerful as they are, are nevertheless inadequate in some ways. The Database is difficult to sift through for linguistic information, while the corpus (KWIC) does not allow for use of corpus tools beyond the concordancer. Both reference tools are potentially limited in their availability of translated texts, and neither offers information on idiomatic target language usage in similar types of texts.

While corpora comprising or including non-translational legal English have also been compiled in the past, it is also not easy to find one with contents sufficiently comparable to that of the Laws & Regulations Database or Bilingual KWIC. The Cambridge Corpus of Legal English, compiled by the Cambridge University Press, for example, is a collection of legal books and newspaper articles; the MultiJur Multilingual Corpus of Legal Texts (University of Helsinki) consists of international conventions and treaties; the BOnonia Legal Corpus (BoLC), compiled by the University of Bologna, meanwhile, comprises European Community documents (Biel, 2010; Rossini Favretti, Tamburini, & Martelli, 2007).

Varantola (2003) explored the compilation and application of disposable or *ad hoc* corpora, which are corpora collected to serve the transitory needs of single translation assignments. While it was found that this type of user-compiled and task-oriented corpora provided reassurance for strategic and lexical decisions in translation, participants of the study also noted the difficulty in corpora compilation and

doi:10.6342/NTU201603145

recommended joint efforts for the task, which in professional practice are not always possible.

The present thesis, therefore, aims at user-compiled LSP corpora of a less specific scope, so the corpora may remain reusable in future assignments of texts in the same specialized domain. By employing a number of readily available tools for corpus processing and analysis, this study sets out to compile a legal Chinese-English corpus consisting of both parallel and monolingual corpora. Attempts are then made to demonstrate the ways in which this user-compiled corpus can be utilized to obtain information on terminology equivalents, frequent collocation, idiomatic usage patterns, and translation strategies so that legal translators can be better equipped to tackle their translation assignments.

The research questions to be addressed are:

(1) How can corpus-based approaches and available computerized tools be utilized to facilitate identification of terminology equivalents and translation units for legal translation?

(2) How can user-compiled corpora be utilized to investigate stylistic features and patterns specific to the legal genre?

(3) How can translational and non-translational corpora be utilized in combination to discover additional information for aiding the process of legal translation?

1.3 Significance of Study

In the process of compiling a parallel and monolingual corpus, this study utilizes a variety of computerized tools and semi-automatic approaches, which will hopefully facilitate the work of future translators who apply them to assembling their own specialized corpora, making the task more manageable when such a need arises.

This study also explores several corpus-based methods for identifying useful and reusable units that can likely be applied in Chinese-English translation of legislation or similar types of texts. A number of results was generated in the forms of key terminology, terminology equivalents, frequently used word strings, other patterns of usage, and observations on translation strategies. While immediately applicable results are limited in quantity, being cited as illustrative examples rather than actual translation resource, the identified key terms, phrase-like units, and usage patterns involve some significant features of the legal language which are likely to be encountered in many cases of legal translation. The examples provided could also serve as starting points for much more and further investigation in the future.

Moreover, it is hoped that the corpus analysis methods employed in this study will offer examples of and insights into the ways specialized corpora can be utilized to serve the practice-oriented needs of legal translators; the combined use of parallel and monolingual corpora as translation reference tool, in particular, seems to be little documented beyond applications to identification of terminological information.

It is therefore anticipated that the employed methods and approaches of corpus compilation and analysis in this study can be applied, at least in part, to other specialized fields of translation to benefit future work in translation practice, translator education, and translation research.

1.4 Terminology

This section provides definitions for key terms and concepts that are frequently discussed in this study. The terms are listed as follows in alphabetical order:

(1) Alignment: the mapping and binding of corresponding source and target text units that translate each other; often performed automatically by computer programs and can be carried out on text units at different levels (Véronis, 2000).

(2) CAT tools: computer-aided translation tools; refers to computerized tools for

doi:10.6342/NTU201603145

helping human translators work more efficiently, often by performing some portion of the translation process (Quah, 2006; Bowker, 2002; Sager, 1994).

(3) Collocation: the “characteristic co-occurrence patterns” of words that “appear together with a greater than random probability” (Bowker, 2002, p. 64).

(4) Colligation: the co-occurrence of grammatical categories with one another, or with a word or phrase (Sinclair, 2003).

(5) Concordance: an “index to the places in a text” where the particular search items occur, commonly displayed in a KWIC (Key Word in Context) format, with search items vertically aligned at the center of context lines (Sinclair, 2003).

(6) N-gram: a recurrent string of uninterrupted *n* items, such as words, lemmas, or part-of-speech tags; *n* stands for a specified number. Also referred to as “cluster” or “lexical bundle” (Stubbs, 2007; Lu, 2014).

(7) LSP: language for special purposes; the language used for communication on a specialized field of knowledge for facilitating interchanges; as opposed to LGP, or language for general purpose (Bowker & Pearson, 2002).

(8) SMT: statistical machine translation, a method of translating texts between natural languages by computer which deduces the most probable results using statistical means on a parallel corpus of previously translated texts (Mitkov, 2003).

(9) Tagset: a set of tags (labels) for denoting the POS categories of words (Lu, 2014), typically assigned automatically by a software program (Bowker & Pearson, 2002); the Penn Treebank tags used in this study are listed in the Appendix.

(10) TM: translation memory; system for storage of translation units and retrieval for reuse upon identification of similar source text (Shuttleworth & Lagoudaki, 2006).

(11) Translational corpora: corpora consisting of translated texts, as opposed to original or non-translational texts written in a naturally occurring environment (Laviosa,

1998).

(12) Translation unit: a pair of aligned source segment and its corresponding translation segment (Shuttleworth & Lagoudaki, 2006).

1.5 Outline of Thesis

Having introduced the background of this study (1.1), its purpose and research questions (1.2), significance (1.3), and the terminology employed (1.4), this section goes on to outline the structure of this thesis.

Chapter 2 summarizes the theories and previous studies that found the basis for this thesis, particularly statistical techniques for corpus analysis and the application of corpora in translation. Also reviewed are relevant corpus processing approaches and previously conducted corpus-based studies on legal English or translation.

The corpora and methods adopted will be introduced in Chapter 3, including the major procedures taken and the tools employed in each stage of compilation and analysis. The results generated are presented in Chapter 4, along with discussions on the results and their implications as they relate to the research questions proposed above. Chapter 5 concludes with a summary of findings, limitations of the present study, and suggestions for future research.

doi:10.6342/NTU201603145

Chapter 2 Literature Review

As made apparent in the introduction in Chapter 1, this study relies heavily on a variety of corpus-based theories and approaches to compile and analyze corpora for legal translation. The relevant theories and previous studies will be reviewed in this chapter: Section 2.1 summarizes the statistical techniques of corpus analysis which provide theoretical basis for the methodology of this study. Section 2.2 focuses on applications of corpora in translation practice and as a reference tool in particular. Section 2.3 introduces a selection of computational linguistics approaches that facilitated the important corpus processing tools this study employs. Lastly, Section 2.4 reviews previous corpus-based studies that have been conducted on the subjects of legal English or legal translation.

2.1 Statistical Techniques in Corpus Analysis

McEnery (2003) defined a corpus as a large, well-organized, and typically machine-readable body of "naturally occurring language data" (p. 449) sampled to represent a particular variety of language. As corpus linguistics is an empirical approach to language analysis (Mitkov, 2003) that typically involves large bodies of data, statistical techniques are naturally entailed in the process of corpus analysis. Introduced below are the major approaches incorporated into the methodology for the current thesis: keywords (2.1.1), n-grams (2.1.2), and concordances (2.1.3).

2.1.1 Frequency Data, Keywords and Keyness

Frequency data has been a common starting point for corpus analysis (Flowerdew, 2012). In the case of a specialized corpus, interesting points for further exploration can be detected by simply comparing the high-ranking items on its word frequency list to those in a reference corpus of a similar size. Composition of a corpus can be studied

doi:10.6342/NTU201603145

through type/token ratio, which is the percentage that occurrences of an individual word account for among all word occurrences in the corpus.

Another way of discovering objects for analysis is by applying frequency data to identifying items of significance that distinguish one corpus from another (Rayson & Garside, 2000). Scott and Tribble (2006) introduced the concept of “keyness,” a quality suggesting the importance of a given word in a text or set of texts. A “keyword” occurs with significantly higher frequencies in a certain text or corpus as compared against their occurrences in a general reference corpus while reaching a frequency threshold at 2 or 3 occurrences.

Whether frequency contrasts are “significant” is determined by statistical tests of probability, which make numeric comparisons between the given frequencies and the expected frequencies estimated with statistics from the reference corpus. As summarized in Ji (2012), a null hypothesis is adopted in statistical analyses, the assumption being that no significant relationships exist between the two sets of data examined. The computed significance value is measured against a predetermined threshold value, normally set at 5%; the null hypothesis can be rejected only when the significance value is lower than the threshold value. In this case, the alternative hypothesis can be accepted, confirming that a significant difference exists between the two sets of data, rather than just random or chance variation.

A significance test commonly employed in corpus analysis is the G^2 test or log-likelihood ratio (LLR), preferred for its applicability without depending on the assumption of normally distributed data (Dunning, 1993). Rayson and Garside (2000) provided the following formulas for calculating the expected frequencies (E) and log-likelihood (LL) values of given words: (Table 2.1 on p. 14 shows the frequency variables used for the formulas.)

doi:10.6342/NTU201603145

Table 2.1

Contingency Table of Word Frequency Variables

	Corpus 1	Corpus 2
Word frequency	a	b
Corpus size	c	d

Note. Adapted from “Comparing Corpora Using Frequency Profiling,” by P. Rayson and R. Garside, 2000, *Proceedings of the Workshop on Comparing Corpora, 9*, p. 3. Copyright 2000 by the Association for Computational Linguistics.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

$$-2 \ln \lambda = 2 \sum_i O_i \ln \frac{O_i}{E_i}$$

In calculating the expected values (*E*) of a word’s frequency in the two corpora, the values “a” and “b” corresponds to the observed values (*O*) of frequency O_1 and O_2, while the values “c” and “d” are the total number of words (*N*) in Corpus 1 (N_1) and Corpus 2 (N_2). For a given word, therefore:

$$E_1 = \frac{\mathrm{c(a + b)}}{\mathrm{c + d}};\ E_2 = \frac{\mathrm{d(a + b)}}{\mathrm{c + d}}$$

$$\mathrm{LL} = 2\left[(\mathrm{a} \times \log \frac{\mathrm{a}}{E_1}) + (\mathrm{b} \times \log \frac{\mathrm{b}}{E_2})\right]$$

The LL scores thus calculated represent the significance of frequency differences between the two corpora; the higher the score, the greater the significance (Rayson & Garside, 2000). When the scores exceed certain “critical values,” the significance value (or “p-value”) is deemed to be lower than the threshold value (Y.-C. Cheng, 2013). Critical values for this calculation method are set at 3.84 (significant at the 5% level or 95th percentile level) for the 95% critical value, 6.63 (1% level or 99th percentile level), 10.83 (0.1% or 99.9th percentile level), and 15.13 for the 0.01% or 99.99th percentile level (Rayson, Berridge, & Francis, 2004).

doi:10.6342/NTU201603145

In Rayson et al. (2004), experiments comparing the reliability of two significant tests found that when applied to different-sized corpora, the log-likelihood test generates fewer inaccurate results than the chi-squared test. For both of the statistical tests, expected values need to reach certain numbers, depending on the threshold value. But at the 0.01% level, the LL test remains accurate even under conditions of highly unbalanced-sized corpora, and applicability can even be expanded to expected values of 1 or more, instead of having to reach expected values of 5 like with the chi-squared test.

Keywords identified with the significance test approach typically include three types: proper nouns; indications of the theme or "aboutness" in a text or corpus; and indicators of style, which may appear to be function words with unusually high frequencies (Scott, 2000; 2006). The keyword approach, therefore, has also been applied by many studies to different types of texts and specialized corpora, including engineering, political speeches, history, marketing (Bondi & Scott, 2010), and to the compilation of a new Academic Vocabulary List from the Corpus of Contemporary American English (Garner & Davies, 2013).

2.1.2 Phraseology and N-grams

Corpora have also facilitated studies on the important subject of phraseology, described in Stubbs (2001) as "the pervasive occurrence of phrase-like units of idiomatic language use" (p. 59). These recurring, multi-word phrasal units reflect a strong tendency for frequently occurring words to be characterized by fairly restricted sets of collocation, and are components of natural-sounding language use. Collocation can be described as the "characteristic co-occurrence patterns" of words that "appear together with a greater than random probability" (Bowker, 2002, p. 64). Another notable type of co-occurrence relations is colligation, the co-occurrence of grammatical categories with one another, or with a word or phrase (Sinclair, 2003).

doi:10.6342/NTU201603145

In a study discussing quantitative research of phraseology, Stubbs (2007) defined the term “n-gram” to mean “a recurrent string of uninterrupted word-forms” (p. 90). This type of multi-word sequences is also been referred to as “clusters” (Scott & Tribble, 2006), “lexical bundles” (Biber & Conrad, 1999), and other such terms. N-grams, therefore, can also include other continuous sequences of a number (n) of items, such as lemmas[1] and part-of-speech tags (Lu, 2014). This concept applied to part of speech was termed “PoS-gram” in Stubbs (2007), which demonstrated how frequent PoS-grams can be used to narrow down the object of study to certain subsets at a time.

Biber and Conrad (1999) regarded lexical bundles as “extended collocations” comprising three or more sequential words, occurring at least 20 times per million words, and recurring across multiple texts in a register, noting also that lexical bundles are usually not complete structures or fixed expressions. The frequency and distribution criteria were further narrowed to 40 occurrences per million words and occurrence in at least five different texts in Biber, Conrad, and Cortes (2004), which investigated four-word lexical bundles in university classroom teaching and textbook prose.

2.1.3 Pattern Grammar and Concordances

Corpus-based phraseology has also given rise to the development of pattern grammar (Stubbs, 2007), an approach to formulating grammatical descriptions of a lexical item by means of patterns, which are phraseologies frequently associated with the meaning of a word or one sense of its meaning (Hunston & Francis, 2000).

The method of pattern investigation depends on concordances, defined in Sinclair (2003) as “an index to the places in a text where particular phrases occur” (p. 173), commonly presented in the format of “Key Word in Context” (KWIC). The KWIC layout places the queried word(s) or phrase(s) at the center (referred to as the “node”)

[1] A “lemma” refers to a set of derived forms from the same word stem and in the same part-of-speech category (Flowerdew, 2012).

doi:10.6342/NTU201603145

among a line of context, and aligns the generated lines vertically to the center.

A methodology and guide were provided in Sinclair (2003) for sampling and interpreting concordance lines based on small samples, 25 a batch in this guide. Sampled concordances are automatically selected and distributed evenly over all concordances, starting at the first instance and with a fixed interval between samples. The gap between selections is calculated by dividing the number of all generated concordance lines by the required number of samples. Conspicuous patterns on either side of the node are studied to formulate hypotheses, beginning with the strongest pattern and revised with an aim of being as comprehensive and predictive as possible. The hypotheses can then be tested repeatedly and refined as needed.

The pattern grammar approach makes use of a number of concordance lines selected at random, for example 50 lines, as presented in Hunston and Francis (2000). When investigating the patterns of a word, concordance samples are sorted into alphabetical order by the word to the left or right of the examined word, depending on the part of speech of the word examined and the side interesting patterns are likely to emerge on for that category of words. Investigations can also initiate from a pattern to explore a part-of-speech category of words associated with the pattern.

A pattern is identified when the following three criteria are met: a combination of words and structures is found to occur regularly; the combination is dependent on a word choice; and a clear meaning can be associated with the combination. As was also pointed out, patterns contribute to a word’s meaning, because a word often occurs in a typical pattern when used in a particular sense among its different meanings; secondly, words tend to share an aspect of the same meaning if they also share a given pattern. Finally, patterns identified in this way can then be represented in a schematic form, and it was suggested that a simple coding system be adopted.

doi:10.6342/NTU201603145

2.2 Corpora as Translation Reference Tool

Recent years have seen the application of corpora expand beyond the well-established use as a basis for compiling dictionaries and grammar books to include application in several new fields (Flowerdew, 2012). In the face of drastic technological developments and industry changes in the world today, corpora and corpus linguistics have become increasingly important in fields of translation, including corpus-based translation studies, corpora as reference sources in translation practice, computer-aided translation technology, and corpora as teaching or learning aids in translator education (Bernardini et al., 2003; Chen, 2012). The following subsections will focus on the categorization of corpora and their usage in translation practice (2.2.1) as well as the designing of specialized corpora for aiding the translation process (2.2.2).

2.2.1 Corpus Typology and Usage

There are a vast number of ways to categorize corpora according to their content, such as written or spoken, subjects of the corpora (general or specialized), the time periods they cover, the languages included, and whether and how the corpora have been processed in certain ways (Lee, 2010).

The types of corpora most commonly referred to with regard to usage in translation likely include the following three categories: monolingual corpora, comparable bilingual corpora, and parallel corpora (Bernardini et al., 2003). Monolingual corpora are usually mentioned with regard to the target language and are useful for providing information on “native-like” means of expression to the translator. Comparable corpora refer to corpora comprised of two or more subsets of non-translational corpora in different languages and selected according to analogous design criteria. They provide linguistic as well as cultural information, typically of the same subject domain, on both the source and target languages for reference and

doi:10.6342/NTU201603145

comparison. Parallel corpora, meanwhile, consist of original, or non-translational, texts in a source language and their translations into a target language. They allow users to observe what strategies translators have used to overcome constraints imposed by the source texts in the process of translation.

The application of parallel corpora to investigating translation strategies was demonstrated in Pearson (2003). By examining a small set of culture-specific references in popular science articles, participants of this study were able to observe strategies used by translators in dealing with situationally-constrained expressions. It was therefore confirmed that there is an important role for parallel corpora in the translator training environment, in which it can serve a fairly different and supplementary function to comparable corpora.

Bowker and Pearson (2002) summarized a number of ways to investigate a corpus using computerized tools to obtain some of the above-mentioned information frequently sought after in translation. Monolingual corpora in the target language are useful reference tools for verifying information, such as ascertaining whether possible terminology equivalents are correct, if a certain collocation is appropriate, and if a usage or pattern is idiomatic. They can provide information on writing style and conceptual explanations, or even be used to identify translation equivalents. For example, users might conduct a context search, narrowing search results to those with one pattern occurring in the vicinity of another, or generate a list of word clusters containing a certain pattern to acquire equivalents that were previously unknown to them.

Parallel corpora, aside from enabling investigation of translation strategies, can serve to provide information on term usage, collocation, and writing style in translated texts, and even be employed for identifying terminological equivalents (Bowker & Pearson, 2002). In fact, it was stated in Toyama (2011) that parallel corpora can be

doi:10.6342/NTU201603145

viewed as bilingual dictionaries in this respect. If incorporated with CAT tools and technology, parallel corpora can provide additional and readily usable material for constructing TM segments (Quah, 2006).

Comparable corpora, which comprise two or more sets of monolingual corpora on subjects in the same domain, can provide the same functions as monolingual corpora when used as a reference tool for translation (Pearson, 2003). Scholars have also maintained that compared to parallel corpora, comparable corpora have the advantage of being easier to compile with higher quality because more monolingual texts are available, though establishing categories and sampling procedures may cause potential difficulties (Maia, 2003).

The term comparable corpora as defined above differs from that in Baker (1995), which referred to the same type of corpora as multilingual corpora in the context of descriptive translation studies. The term comparable corpora was instead reserved for corpora consisting of original texts written in a certain language and translated texts into the same language. Such corpora would effectively include a monolingual corpus and translational corpus of similar design. It was proposed for this type of comparable corpora to be used in identifying patterns specific to translated texts.

This definition of comparable corpora and line of research was adopted in Laviosa (1998), which compiled an English Comparable Corpus comprising a monolingual (non-translational) subset and a translational component of newspaper articles and narrative prose. The study found four major patterns in lexical use of translational English texts, including lower lexical density (percentage of content words against functional words), higher proportion of high-frequency words, more repetition among the most frequent words, and fewer lemmas, as compared against non-translational texts in the corpus.

doi:10.6342/NTU201603145

2.2.2 Designing of Specialized Corpora

Utilization of corpora as a reference tool, as in the ways described above, was recommended in Bowker and Pearson (2002) to translators, who are often required to learn the language for communicating on the specialized subject field they are working with. A language variety of this type is termed a “language for special purposes” (LSP), and can be more effectively acquired through consulting a special purpose corpus, which presents a particular aspect of a language, such as an LSP of a particular subject field, a specific text type, or a particular language variety. Specialized corpora can be a valuable complement to other reference sources, especially since dictionaries, printed texts, or other conventional LSP-learning materials may be constrained due to incompleteness, physical volume, time requirement, or unavailability.

In addition to the purpose intended, as well as languages and subject domains to include, there are a number of issues to consider when designing a corpus, including size, full or excerpt texts, authorship, text format, and even copyright. Corpora can also consist of written or spoken language; they can be synchronic, meaning they are representative of the language use within a limited time frame, or diachronic, which facilitate studies on how the language evolved over time; they can also be constantly expanded and changed (open) or of a finite size (closed). McEnery and Wilson (2001) pointed out that a corpus must be representative of a language variety; other generally recommended criteria include a reasonably large size, full texts rather than excerpts, texts by a variety of authors, and electronic format, but corpora size ranging from thousands to hundred thousands of words have all been effective for LSP studies (Bowker & Pearson, 2002).

A very specialized type of corpora was explored in Varantola (2003), namely the disposable (*ad hoc*) corpora collected for the needs of single translation assignments.

doi:10.6342/NTU201603145

From the Finnish-English/English-Finnish translation assignments completed by workshop participants, it was observed that corpora benefited the translation projects by providing reassurance for strategic and lexical decisions, especially in cases of radical decisions to break from the source material. Some participants also found corpus evidence to support choices of register, because target audience for their particular assignment was taken into account in the stage of corpus compilation. However, participants of the study also questioned the cost-efficiency of corpus compilation, as the undertaking had proved difficult for reasons including accessibility and reliability of many materials.

An example to the other side of the spectrum is perhaps the bi-directional Portuguese-English parallel corpus *Compara* (http://www.portugues.mct.pt/Compara/), the design of which does not address issues of corpora balance and representativeness (Frankenberg-Garcia & Santos, 2003). The corpus is open-ended, with no pre-determined rules as to what variety of texts could be included. The texts initially included were fiction, because texts of other genres were either not common, lacking in either language direction, questionable in quality, or often relayed (translated into Portuguese or English from a third language). However, users are given the options of narrowing down the varieties of language, subject, publication date, author, or translator when conducting searches, effectively allowing corpus users to work with tailored sub-corpora to serve the specific purposes of their tasks at hand.

2.3 Computational Linguistics and Corpus Processing

Computational linguistics was defined in Mitkov (2003) as the field of studies "concerned with the processing of language by computers" (p. x). Many technological applications today function on the basis of computational linguistics techniques; machine translation, information retrieval, speech recognition, and text data mining are

doi:10.6342/NTU201603145

just a few of the numerous examples.

Corpora data have played an essential role in the development and evaluation of many natural language processing applications (McEnery, 2003). At the same time, corpus linguistics has also benefitted from incorporation of these increasingly sophisticated language processing programs. The subsections below will introduce the technologies supporting the corpus processing programs this study makes use of: part-of-speech tagging (2.3.1), statistical machine translation (2.3.2), sentence alignment (2.3.3), and phrasal alignment (2.3.4).

2.3.1 Part-of-Speech Tagging

Part-of-speech (POS) tagging refers to the automatic assignment of grammatic tags, which are attached by computer programs to indicate the POS category of input words (Voutilainen, 2003). POS tagging is perhaps the most common type of annotation, namely the “addition of explicit linguistic information” to corpus texts (Bowker & Pearson, 2002, p. 229).

McEnery (2003) summarized four key advantages to annotating a corpus. Firstly, annotation increases the ease of corpus exploitation by making the results of corpus analyses available to human users unfamiliar with the language as well as machines. Users capable of performing the analyses can also save time by obtaining the information directly from the annotations. Secondly, annotation allows the results of analyses to be recorded for reuse without unnecessary repeat of analyses. Thirdly, annotation enables corpus analyses to serve multiple functions, including purposes for which the analyses were not originally intended. Finally, annotation makes explicit the interpretation performed, and by opening them for scrutiny enables them to stand more objectively than interpretations unrecorded.

POS tagging provides information for addressing a number of linguistic issues.

doi:10.6342/NTU201603145

As pointed out in Reppen (2010), many words have multiple meanings which can belong to different word categories and could not be distinguished from spelling. Working with a POS-tagged corpus allows users to disambiguate among such polysemous words in frequency lists and other empirical results. Users can therefore focus on a specific word class, or filter out irrelevant search results. For example, if a researcher wishes to study the high-frequency verbs in a specialized field, or narrow down search results to the modal “can” instead of including its other POS forms, they can do so relatively easily by exploiting POS tags (Bowker & Pearson, 2002; Reppen, 2010). With term extraction applications, POS tags can improve identification of terminology candidates for automatic retrieval, based on the knowledge that nouns and adjectives provide more likely indicators of terms than words of other categories (Voutilainen, 2003).

The architecture of most taggers includes functions for tokenization, ambiguity look-up, and disambiguation. Word boundaries must first be identified to divide the input text into units that allow analysis, a process that is also referred to as word segmentation. Taggers then begin assigning possible POS solutions to input words by use of a lexicon, which is essentially a collection of word forms and their corresponding parts of speech. The same information may also be provided in the more economic form of generalized morphological rules. Tokens not included in the lexicon are then assigned with possible POS solutions by use of a guesser, which proposes reasonable analyses by eliminating unlikely alternatives based on information about the lexicon; the lexicon could be known to include all pronouns and articles, for example, and thus allow the guesser to eliminate these two word classes as possibilities. Finally, remaining ambiguities are resolved based on word information and contextual information encoded in the tagger. Word information includes knowledge such as the likelihood or

doi:10.6342/NTU201603145

frequency of a word being used as a particular word category over another. Contextual information refers to probabilities of POS sequences that enable deduction of the appropriate analysis (Voutilainen, 2003).

A point worth noting in tokenization is that while word-level segmentation presents relatively less challenges with languages in which words are delimited by a white space, the same process is significantly more complicated for Chinese and other languages in which tokens directly precede and succeed each other (Mikheev, 2003). Word boundaries must then be identified by turning to statistical methods such as maximum sequence matching, n-gram methods, and other probabilistic models.

Word segmentation, like other such text processing applications as text alignment, also requires sentence segmentation and is affected by the quality of its results. Sentence segmentation is usually performed in earlier text processing stages with regular expressions, introduced in Lu (2014) as special characters that can be used for specifying patterns. Sentence boundaries are most commonly identified with a sequence of sentence terminal, blank space, and capital letter. The error rate produced by such an algorithm can be reduced by supplementing information such as abbreviations that are never located at sentence endings, or words that always begin a new sentence when capitalized and succeeding a period (Mikheev, 2003).

2.3.2 Statistical Machine Translation

Corpora can be said to have founded the basis for a new paradigm in machine translation that emerged in the 1990s, since which time corpus-based methodologies have been explored by researchers in addition to the ongoing and more traditional, linguistic rule-based approaches (Somers, 2003).

Machine translation (MT) was defined by the European Association for Machine Translation (EAMT) as “the application of computers to the task of translating texts from one natural language to another” (http://www.eamt.org/mt.php). Statistical machine translation (SMT), in particular, is an MT method which uses statistical means and a parallel corpus of previously translated texts to deduce the most probable translation for input texts (Mitkov, 2003; Somers, 2003).

The SMT approach differs significantly from traditional MT methods in that it is highly non-linguistic (Somers, 2003). Appropriate translations determined by an SMT system are based on two sets of statistical probabilities: firstly, the likelihood that a particular set of words in the source text will give rise to particular combinations of target text words; secondly, the possibility that the generated words are arranged in correct sequences in the target language. These two sets of data manifest as a “translation model” and “(target) language model,” respectively (p. 516), which are typically a parallel corpus, most likely aligned at the sentence level, and a monolingual corpus of the target language(s).

Once provided with a source language text to translate, an SMT system divides the input text into units of word groups or phrases. The source text units are then compared against a parallel corpus, from which the translation model identifies a number of target language units that likely translate the source units. The possible equivalent units are then passed on to the language model, which determines the most probable word sequence in terms of linguistic validness in the target language based on n-gram probabilities derived from the monolingual corpus. The SMT system then outputs the results with the highest probability of being an accurate translation of the source text and linguistically valid word-sequence combinations in the target language (Somers, 2003; Quah, 2006).

Machine translation has also initiated much of the modern interest in parallel texts and in turn alignment (Gale & Church, 1991a), which is introduced below.

doi:10.6342/NTU201603145

2.3.3 Sentence Alignment

An important process of compiling parallel corpora is alignment, which refers to the mapping and binding of corresponding source and target text units that translate each other. This process, often performed automatically by computer programs, can be carried out on text units at different levels, including paragraphs, sentences, phrases, and words. The technique is required in a wide variety of applications; in addition to compiling parallel corpora, it is used for compiling translation memories, dictionaries, and bilingual glossaries, while also applied in cross-language information retrieval (Véronis, 2000; Bowker & Pearson, 2002; Quah, 2006).

Véronis (2000) pointed out that most alignment methods at sentence level are based on one or both of two major principles: lexical anchoring and sentence length correlation. Lexical anchoring methods make use of corresponding lexical elements, which are established as “anchor points” and a basis for identifying likely sentence alignments. These lexical anchors can be word pairs, either word-level alignments derived from texts to be aligned, or word translations obtained from an external bilingual dictionary; or, they may be “cognates,” which are graphically similar or identical elements such as names, dates, figures, symbols, special punctuation marks, or words with similar spelling in the source and target languages.

An early example of lexical anchoring was given in Kay and Röscheisen (1993); the study proposed a sentence alignment method supported by partial word-level alignments derived from word distributions in the texts on which sentence alignment was to be performed. The theoretical basis for this method arised from the observation that sentence pairs containing an aligned word pair will certainly be appropriate sentence alignments as well. Using an initial set of possible sentence alignments based on their location within the texts, a most likely set of aligned words is identified

doi:10.6342/NTU201603145

according to the tendency of their appearance in corresponding sentences. The aligned word set is then used to calculate new results of aligned sentence pairs. The resultant information contributes to a new estimate of possibly aligned words, and the induction process is repeated until no new sets of sentence alignments are found.

Sentence length correlation methods, on the other hand, were derived from the knowledge that the lengths of translated sentences have a tendency to correlate highly with that of the source sentences from which they originated (Véronis, 2000). The statistical model proposed by Gale and Church (1991b) based its calculation of sentence length on the number of characters per sentence. According to empirical data, the researchers determined the mean and variance of the ratio of target text characters per source character, in other words, the number of target text characters that each source text character gives rise to.

Sentence alignments were categorized into four types, for each of which their probabilities of occurrence were calculated. The four types were one source text to one target text sentence alignments, one source or target text sentence with no corresponding counterparts, one source or target text sentence to two matching sentences, and two source text to two target text sentence alignments. The above information and lengths of the proposed sentence pair being considered are incorporated to compute a probabilistic score, with which the maximum likelihood for sentence alignment is derived.

A hybrid model making use of both lexical anchoring and sentence length correlation methods was proposed in Brown, Lai, and Mercer (1993). Working with records of Canadian Parliament proceedings, i.e., Hansard, the study used existing comments such as speakers or time as anchor points. After aligning subsections of the French and English records as divided by the anchors, a probabilistic model computed sentence alignments within subsections based on sentence length by word count.

doi:10.6342/NTU201603145

Despite the differences across sentence alignment methods that have been proposed, these alignment models generally operate on a number of common assumptions about the source and target texts to be aligned. It is often assumed that the source and target text will largely correspond sentence by sentence, in approximately if not exactly the same order, with very few one-sentence-to-two, two-to-one, or two-to-two correspondences, very few omissions, and additions (Véronis, 2000). However, as pointed out in Frankenberg-Garcia and Santos (2003), source text sentences are quite often split, combined, inserted with additional elements, or reordered during the translation process. Such alterations create considerable problems for automatic alignment programs. In fact, evaluations of the alignment model in Gale and Church (1991b) showed that in the case of sentence pairs involving addition or deletion, the alignment program had never achieved correct results. The possibility of three or more sentences in either the source or target text of an aligned segment was not considered in the statistical model, yet such occurrences do still exist. It is therefore quite likely that manual adjustments and correction of misaligned results would often be required to obtain more satisfactory sentence alignments.

2.3.4 Word and Phrasal Alignment

Accuracy in sentence alignment becomes an important issue when the results are used as starting point for word-level alignment, in which case partially correct sentence alignment is no longer sufficient (Véronis, 2000). Processes of lexical alignment or extraction typically consist of two phases: detection of words or expressions in the source and target texts, followed by the mapping of those expressions onto each other.

To overcome the costs and language specificity constraints of linguistic approaches, researchers have continued to develop statistical-based methodologies for lexical alignment. In the automatic translation approach they proposed, Brown et al.

doi:10.6342/NTU201603145

(1990) introduced statistical techniques to facilitate automatic glossary compilation based on the belief that in a large corpus, the correct translation for a given source-language word will occur significantly more frequently than other candidates in their corresponding target language sentences. To account for differences in lengths between source- and target-text sentence pairs, the algorithms were further refined by accounting for source text words that produce “null words” (words for which a correspondence does not appear in the target text) or secondary words.

Addressing constraints in case of source and target languages with different ordering arrangements, Wu (1995a; 1995b) introduced another automatic approach for identifying phrasal translation units. This method makes use of an inversion transduction grammar (ITG), a probabilistic formalism for bilingual language modeling and parsing. The input sentence pairs undergo syntactic analysis in order for supposedly correct grammatical structures to be extracted. The ITG algorithms generate separate output streams for both the source and target language and match the corresponding constituents from the two streams, allowing for constituents to be paired up in either a left-to-right or inversed order. ITG, therefore, provides a language-independent and sequentially flexible approach to extracting several types of linguistic information from parallel corpora, including aligned phrasal or word units.

Several studies have later made use of or developed from the foundation of ITG. One of those studies is Neubig, Watanabe, Sumita, Mori, and Kawahara (2011), in which an unsupervised probabilistic model for extracting phrasal alignments at multiple syntactic levels. Instead of building up from minimal phrase alignments, this ITG-based model generates phrase pairs at every branch of the syntax tree. The end result is a phrase table for SMT translation models that includes phrases at levels ranging from words to full sentences.

doi:10.6342/NTU201603145

While the majority of bilingual concordance programs are based on sentence alignments, a word-based program will be highly advantageous to the user as it can identify the correspondence to the input word without requiring the user to supply the possible corresponding words in a second language (Gale & Church, 1991a). In Dagan, Church, and Gale (1993), which presented a word alignment method developed from a later model of Brown et al. (1990), the researchers also pointed out that word alignment programs can help translators save considerable time by providing them with results of terminology questions already solved by other translators. In fact, a word-based bilingual concordance program has doubled or even more than tripled the speed with which translators produced bilingual terminology lexicons at the partner institution of this study. Even without comprehensive alignment results for all the input words, word or phrasal alignment can be helpful to translators (and lexicographers) in addressing issues of difficult terminology (Dagan et al., 1993).

2.4 Corpus-based Studies on Legal Language and Translation

Biel (2010) pointed out that while application of corpora has become increasingly popular in many fields of linguistics and translation, relatively seldom have corpora been applied in studies of specialized translation and legal translation in particular. In the case of corpus-based studies on legal language, Biel (2010) summarized four major types of objectives that studies in the past have so far addressed:

Studies on external variation examine the differences of legal language from language of general purposes or from other LSPs. Some studies focus on internal variation, which refers to differences among legal genres. Studies can also address temporal variation, and observe how the current and historical legal languages differ from each other. Finally, cross-linguistic variation refers to differences across different languages. Some of the topics have already been addressed in previous corpus-based

doi:10.6342/NTU201603145

studies on legal language and translation relevant to the current thesis, as will be summarized below.

2.4.1 Legal Language and External Variation

In a study of forensic linguistics, the study of language used in the justice system and related to law (Biel, 2010), Coulthard and Johnson (2007) provided examples of corpus-based analysis to illustrate certain features of legal English. Comparison was made between the lists of most frequent words in the COMET corpus of legal contracts and British National Corpus (BNC).

While pages dedicated to the corpus-based approach were relatively few, with results used mainly as supplementary examples, the study nevertheless identified several features pertaining to grammatical words and lexical words in legal contracts. The function words “or,” “any,” “shall,” “be,” and “by,” for example, were found with markedly higher distribution in the contract language than in general language.

Lexical density was also found to be higher in contract language, indicated by the higher number of content words against function words among the most frequent items. The most frequent content words in legal contracts were mainly nouns referring to the parties involved or to the contract itself, the frequent occurrences perhaps partly attributable to the preference of repetition over using pronouns. This use of frequency data in identifying features for detailed observation is also adopted in the present thesis, which uses keywords and keyness derived from word frequencies as the basis for observing distinctive features of statutory language.

2.4.2 Internal Variation of Legal English

Legal English can also be further categorized into different legal genres, which serve different purposes and can be seen as different levels of specialist communication.

doi:10.6342/NTU201603145

The resulting differences in linguistic characteristics, or “linguistic variation,” were examined in Gozdz-Roszkowski (2011), using corpora compiled into the American Law Corpus (ALC). The corpus comprises approximately 5.5 million words in a total of 687 texts from seven major genres: academic journals, briefs, contracts, legislation, opinions, professional articles, and textbooks, the size of each component determined by its relative availability and importance.

The different legal genres were compared against one another in a number of analyses, which were used to describe the linguistic characteristics of different genres, including vocabulary distribution and use, diversity of lexical choice, extended lexical expressions, and lexico-syntactic co-occurrence patterns. Methods of analyses include keywords, lexical bundle, and multi-dimensional analysis (MD), a methodological approach introduced in Biber (1998, 2004) for studying linguistic variation across registers by use of multivariate statistical techniques. Co-occurrence patterns among a variety of linguistic features (“dimensions”) are identified by empirical means, and interpreted by microscopic analyses as to the underlying communicative functions they reflect, such as “information-focused vs. interactive,” “stance vs. context-focused,” and “narrative-focused” (Biber, 2004).

Findings of the ALC study include descriptions of the seven commonly encountered legal language genres in terms of their linguistic characteristics, similarities, and differences (Gozdz-Roszkowski, 2011). Legislations, for example, were found to show a strong presence of informational and normative features, a highly explicit, highly impersonal, and non-narrative style. The above features were also discovered in contracts, which also displayed high-frequency uses of many specialist terms and denser terminology. Textbooks, on the other hand, showed strong narrative and stance-oriented concerns, which were also relatively strong features in academic journals and

doi:10.6342/NTU201603145

professional articles. Legislations also showed a focus on legal reference in lexical bundles, a large proportion of them being procedure or time related.

Having demonstrated that legal genres differ from one another in many ways, the ALC study concluded that “legal English” is in fact “a system of related domain-specific genres” (p. 228), the individual genres of which varying greatly in terms of their linguistic features. The findings above provide theoretical support to the present thesis in focusing on only one of the genres defined in the ALC study; otherwise, studies would likely have to be conducted separately for each genre, making the scope of research unmanageable for a study aiming to investigate linguistic patterns in detail.

2.4.3 Legal Language in Chinese and English Contracts

A corpus-based study on Chinese and English legal translation is seen in Chen (2012), in which a bilingual comparable corpus was compiled for extracting translation equivalents in legal contracts. The corpus consisted of approximately 660 thousand words from 167 English contracts and 910 thousand characters from 229 Chinese contracts. Corpora were non-translational texts selected from the Internet.

With the use of corpus analysis tools, English and Chinese corpora were respectively compared against the Corpus of Contemporary American English and the Academia Sinica Balanced Corpus to identify keywords and key keywords based on the LLR test. Key keywords are “words that are key in many texts” in a corpus (Scott, 1997). The keywords and key keywords were then filtered according to keyness, document frequency (number of texts they occur in), and word frequency to identify the words most distinctive to legal contracts (Chen, 2012).

N-gram lists were next compiled according to the most significant (key) keywords. The n-grams obtained were translated (the English n-grams into Chinese, and vice versa) by using the SMT system Google Translator Toolkit. The SMT-generated

doi:10.6342/NTU201603145

translations and their corresponding source segments were then used as “TM” in the SDLX Translation Suite workbench. The SMT translations were compared to the n-grams directly extracted from contracts of the same language; when similarity exceeded the pre-assigned value, the CAT tool automatically applied the source segment as corresponding “translation” to the n-gram. In this way, translation equivalents were established between the original Chinese and English n-grams.

Evaluation by experienced contract translators determined that the translation equivalents with a similarity of 95% or above achieved an accuracy of 82%. The translation equivalents could then be used as TM to facilitate contract translation. The n-gram pairs could also be used as keywords for investigating relevant linguistic patterns or conceptual knowledge in a specialized corpus of contracts.

While the new method for obtaining translation equivalents or TMs in this study was theoretically simple and effective, the researcher has also pointed out that a great deal of manual work was nevertheless required in the process of correcting or filtering automatically generated results. The workbench and some of the corpus analysis tools (key keywords tool, reference corpora) employed are also not available to the average translator. The efforts and monetary costs involved, therefore, may make this method less appealing to individual translators. Meanwhile, as bilingual texts already exist in the case of corpora selected for the present thesis, it may also be worthwhile to explore alternatives methods for obtaining similar types of resources.

The literature reviewed in this chapter is summarized in Table 2.2 beginning on p. 36. These theories are adopted to form the methods for this study, as will be introduced in the next chapter.

doi:10.6342/NTU201603145

Table 2.2

Summary of Literatures Reviewed

Subject	References	Summary and/or Findings
Statistical Techniques in Corpus Analysis		
Corpus definition	McEnery (2003)	A corpus is a large, well-organized, and typically machine-readable body of “naturally occurring language data” sampled to represent a particular variety of language
Frequency data	Flowerdew (2012)	● Frequency data is a common starting point for analysis ● Type/token ratio: percentage of individual word frequency over total word occurrences in a corpus
Keywords & keyness	Scott & Tribble (2006)	● Keywords: words with significantly higher frequencies in a given corpus as compared to their occurrences the reference corpus ● Keyness: significance of frequency difference as determined by statistical tests of probability
	Scott (2000)	Keywords identify proper nouns, indications of theme, and indicators of style
	Rayson & Garside (2000)	LLR is calculated based on word frequencies and sizes of the studied and reference corpora
	Rayson et al. (2004)	● Critical values of LLR and significance level: 3.84: significant at 5% (95th percentile) level 6.63: significant at 1% (99th percentile) level 10.83: 0.1% (99.9th percentile) level 15.13: 0.01% (99.99th percentile) level ● LLR has higher reliability over chi-squared test, especially in case of highly unbalanced-sized corpora
Phraseology	Stubbs (2001)	● Phraseology is “the pervasive occurrence of phrase-like units of idiomatic language use” ● High-frequency words have a strong tendency to co-occur with restricted sets of collocation ● Notable co-occurrence relations include collocation and colligation
	Stubbs (2007)	● An n-gram is “a recurrent string of uninterrupted word-forms,” also referred to as “clusters,” “lexical bundles,” etc. ● PoS-grams: a continuous sequence of POS tags, can narrow down the object of study
	Biber & Conrad (1999)	Lexical bundles are: ● Definition: “extended collocations” with three or more sequential words and frequent recurrence across multiple texts in a register ● Usually not complete structures or fixed expressions
Pattern Grammar & Concordance	Sinclair (2003)	Concordance: ● Definition: “an index to the places in a text where particular phrases occur” with key phrases commonly presented at the center of a line of context

Table 2.2 (Continued)

Subject	References	Summary and/or Findings
Statistical Techniques in Corpus Analysis		
Pattern Grammar & Concordance	Sinclair (2003)	● Sampling and interpretation methodology: Select small batch of concordance lines distributed at fixed interval over all lines. Study conspicuous patterns to formulate hypotheses; revise, test, refine repeatedly.
	Hunston & Francis (2000)	● Pattern grammar: approach to formulating grammatical descriptions of lexical items by patterns ● Patterns: phraseologies frequently associated with (one sense of) a word's meaning ● A combination of words and structures meets the criteria of a pattern if it: (1) Occurs regularly (2) Is dependent on a word choice (3) Can be associated with the meaning of the combination
Corpora as Translation Reference Tool		
Typology and usage	Bernardini et al. (2003)	Common types of corpora used in translation: (1) Monolingual corpora: usually in target language, contains native-like means of expression (2) Comparable corpora: non-translational corpora in two or more languages selected with analogous criteria (3) Parallel corpora: original source texts and their translation; allow observation of translation strategies
	Pearson (2003)	● Parallel corpora aids translation of culture-specific, situationally-constrained expressions and supplements comparable corpora ● Comparable corpora provide the same functions as monolingual corpora
	Bowker & Pearson (2002)	Functions of corpora as reference tool: ● Monolingual corpora: (1) Verification of terminology, collocation, idiomatic usage (2) Inform writing style, conceptual explanations (3) Identification of equivalents by context search or clusters list ● Parallel corpora: (1) Linguistics information on translated texts (2) Identification of terminology equivalents
	Toyama (2011)	Parallel corpora can serve as bilingual dictionaries
	Quah (2006)	Parallel corpora provide usable TM for CAT tools
	Maia (2003)	Comparable corpora are easier to compile with higher quality texts than parallel corpora for monolingual text availability
	Baker (1995)	Comparable corpora in translation studies: original texts written in and translated texts into the same language; used for identifying patterns specific to translated texts

Table 2.2 (3)

Subject	References	Summary and/or Findings
Corpora as Translation Reference Tool		
Typology & usage	Laviosa (1998)	Study on English Comparable Corpus found lower lexical density, higher proportion and more repetition of high-frequency words, fewer lemmas in translational English
Corpus design	Bowker & Pearson (2002)	● Language for special purposes (LSP): language for communicating on a specialized subject field; LSP corpora are effective reference tools to translators ● Criteria recommended for LSP corpora include reasonably large size, full texts, a variety of authors, and electronic format
	Varantola (2003)	● Disposable (*ad hoc*) corpora: specialized corpora compiled for single translation assignment ● Cost-efficiency of corpora compilation was questioned due to difficulties including accessibility and reliability of many materials
	Frankenberg-Garcia & Santos (2003)	● Open-ended corpus: corpus that can be expanded to include any important texts to users ● *Compara*, a Portuguese-English parallel corpus, is open-ended with no pre-determined rules on types of included texts. Users can choose to work with selected sub-corpora as needed.
Computational Linguistics and Corpus Processing		
Computational linguistics	Mitkov (2003)	Definition: the field of studies "concerned with the processing of language by computers"
	McEnery (2003)	Corpora data are essential in the development and evaluation of many language processing applications, which are incorporated in and benefitting to corpus linguistics studies
Part-of-Speech (POS) Tagging		Advantages to corpus annotation: (1) Increases the ease of corpus exploitation (2) Records results of analyses for reuse without unnecessary repeat of analyses (3) Enables analyses to serve multiple functions (4) Makes explicit and increases objectivity of interpretations
	Voutilainen (2003)	● POS tagging: automatic assignment of grammatic tags by computer programs to indicate POS category ● General architecture and functions of taggers: (1) Tokenization/segmentation: identification of word boundaries and units allowing analyses (2) Assigning possible solutions by lexicon (3) Guesser narrows down possible ambiguity solutions (4) Resolving ambiguities by word and context info ● POS tags facilitate terminology identification because nouns and adjectives are likelier indicators of terms

Table 2.2 (4)

Subject	References	Summary and/or Findings
Computational Linguistics and Corpus Processing		
Part-of-Speech (POS) Tagging	Reppen (2010)	POS tagging allows disambiguation among polysemous words to facilitate analyses
Tokenization and segmentation	Mikheev (2003)	● Word-level segmentation is challenging with languages in which words are not delimited by a white space (e.g. Chinese), and statistical methods are often required ● Word segmentation is affected by the quality of sentence segmentation ● Sentence boundaries are commonly identified by a sentence terminal, blank space, capital letter sequence
Statistical machine translation (SMT)	EAMT (n.d.)	Machine translation (MT): application of computers to translating texts from one natural language to another
	Somers (2003)	● SMT: MT method adopting statistical means and examples of previously translated texts ● Statistical probabilities for determining appropriate translations in SMT: (1) Likelihood that particular source text words will give rise to particular target text words (translation model) (2) Possibility that generated words are in correct sequence in target language (language model)
Sentence alignment	Véronis (2000)	● Alignment: mapping and binding of source and target text units that translate each other, often by computer programs and can be performed on text units at different levels ● Major sentence alignment methods: (1) Lexical anchoring: using corresponding lexical elements as “anchor points” for identifying likely sentence alignments (2) Sentence length correlation: based on tendency of translated sentences to correlate highly with corresponding source sentences in length (3) Hybrid of (1) and (2) ● Common assumptions of alignment models: source and target texts will largely correspond: (1) Sentence by sentence (2) In approximately the same order (3) With very few multiple-sentence matches (4) With very few omissions or additions ● Higher accuracy in sentence alignment is demanded when results are used for word-level alignment
	Kay & Röscheisen (1993)	Theoretical basis for proposed lexical anchoring method: sentence pairs containing an aligned word pair will be appropriate sentence alignments
	Gale & Church (1991b)	● Statistical basis of proposed length correlation model: (1) Mean and variance of ratio of target text characters per source character

doi:10.6342/NTU201603145

Table 2.2 (5)

Subject	References	Summary and/or Findings
Computational Linguistics and Corpus Processing		
Sentence alignment	Gale & Church (1991b)	● Statistical basis of proposed length correlation model: (2) Probabilities of alignment types by corresponding number of sentences ● Sentence pairs involving addition or deletion were challenging with no successful results
Word & Phrasal alignment	Véronis (2000)	Phases of lexical alignment/extraction processes: (1) Detection of source and target text words/expressions (2) Mapping of detected expressions onto each other
	Brown et al. (1990)	Basis of proposed glossary compilation approach: correct word translation will occur significantly more frequently than other candidates in corresponding target text sentences
	Wu (1995a; 1995b)	Inversion transduction grammar (ITG) approach: employs probabilistic algorithms for grammatical structure extraction and inversed-order pairing of constituents
	Gale & Church (1991a)	Advantage of word-based bilingual concordance over sentence-based: enables identification of word correspondence without requiring input of possible corresponding words in both languages
	Dagan et al. (1993)	Word alignment programs save considerable time by providing results of terminology questions already solved by other translators
Corpus-based Studies on Legal English or Translation		
Major types of studies	Biel (2010)	(1) External variation: difference of legal language from language of general purposes or other LSPs (2) Internal variation: difference among legal genres (3) Temporal variation: current vs. historical legal languages (4) Cross-linguistic variation: studies across languages
External variation	Coulthard & Johnson (2007)	Variation of most frequent words in COMET (corpus of legal contracts) and British National Corpus (BNC): ● High distribution of function words: or, any, shall, etc. ● Higher lexical density
Genre variation	Gozdz-Roszkowski (2011)	Study of American Law Corpus (ALC) concluded that individual genres of legal English vary greatly in terms of linguistic features
Chinese & English legal contracts	Chen (2012)	● A method combining keyword analysis, n-gram extraction, SMT results, and CAT tool matching was found effective for obtaining translation equivalents or TMs from a bilingual comparable corpus ● A great deal of manual work was nevertheless required for correcting or filtering automatically generated results

Chapter 3 Method

On the basis of previous and studies as summarized in the previous chapter, this chapter outlines the methods and tools employed in conducting this study. Section 3.1 introduces the selection of corpora and means of their collection; Section 3.2 describes how the corpora are processed and annotated for analysis; finally, Section 3.3 summarizes the approaches with which the corpora are studied, as well as the computer tools employed in the process of analysis.

3.1 Corpus Selection

According to criteria recommended in previous studies, three sets of corpora were collected for this study with an attempt to achieve representativeness of a language variety, reasonably large size, full texts, and machine-readable format (McEnery & Wilson, 2001; Bowker & Pearson, 2002). Considering issues of practicality for working translators, accessibility of data is also prioritized. This selection therefore includes the Chinese texts of Taiwan statutes (Chinese corpora), the English translations of the same statutes (translational corpora), and non-translational English texts of United States statutes (English corpora), as summarized in Table 3.1.

Table 3.1

Corpus Features and Information

	Chinese Corpora	Translational Corpora	English Corpora
Content	Chinese source texts of Taiwan statutes	English translations of Taiwan statutes	Non-translational English texts of United States statutes
Size	2.2 million characters	1.9 million words	20 million words
Representativeness	Statutory language of Taiwan legislation	Language of Taiwan statute translations	Statutory language of U.S. legislation
Publication Time	Dec. 1929-Feb. 2015	Dec. 1929-Feb. 2015	Jul. 1862-Dec. 2014

doi:10.6342/NTU201603145

In addition to being a representative and significant type of legal texts, statutes are also accessible to the public in electronic format, and therefore relatively convenient for practicing translators to obtain by large quantity and in full. To achieve a certain degree of comparability between the translational and English corpora, only texts and translations of central-government legislation passed by the Taiwan parliament (Legislative Yuan) were chosen, as well as federal statutes enacted by U.S. Congress, while local-government statutes in Taiwan and U.S. state laws were not included.

As government regulations require all parliament-made laws in Taiwan to be translated into English, translations of legal texts at this level naturally include laws on a variety of subject matters. To maintain the balance among legislation on different subject matters, texts at the level of administrative regulations or below were not selected, as English translations may not be available unless the regulations concern foreign nationals, institutions or organizations, or are deemed necessary of translation by the relevant authorities (MOJ, 2015). The translational corpora in this study therefore comprise the translations of 510 laws in-effect at the time of collection, while the Chinese corpora consist of the corresponding original texts from the same 510 laws.

The English corpora were obtained from contents of the *United States Code* (*U.S.C.*), a compilation of “general and permanent laws” in statutes enacted by U.S. Congress. Selections for the Code are made by the Office of the Law Revision Counsel (OLRC), House of Representatives from newly-enacted bills, arranged according to subject matter as sections or statutory notes, and updated regularly. The latest version of the *U.S.C.* at the time of this study was the online version current through Pub. L. 113-296 (12/19/2014), except for Pub. L. 113-287, 113-291, and 113-295 (OLRC, 2015). The English corpora therefore comprise 52 of the existing 54 *U.S.C.* titles, excluding the two that were unavailable due to being repealed and reserved respectively.

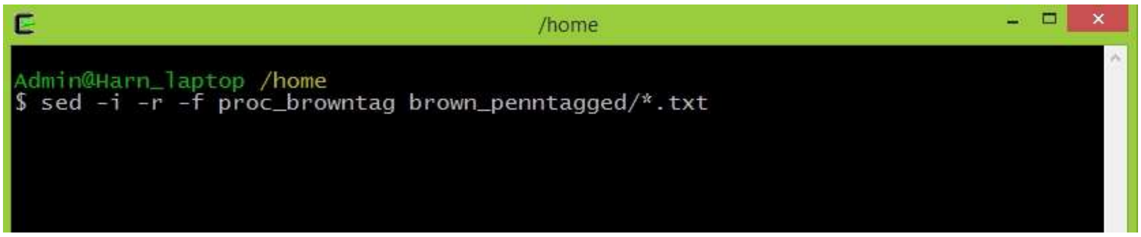
doi:10.6342/NTU201603145

Collection of the Chinese and translational corpora was completed with the help of the crawler *cURL* (Stenberg, 2015) from the Laws & Regulations Database of the Republic of China (http://law.moj.gov.tw/) maintained by the Ministry of Justice. The translational corpora were collected first; after excluding the texts of treaties, the corresponding Chinese corpora were next collected according to the translational corpora available. Legislations no longer in-effect were filtered out, identified with the “Abolished before reform” notations in the “Category” column. Parliament-made laws were distinguished from administrative regulations according to naming principles specified in the *Central Regulation Standard Act* (MOJ, 2015).

The English corpora were downloaded from the website of the Office of the Law Revision Counsel (http://uscode.house.gov/). The webpage format was chosen so further processing could be conducted semi-automatically according to information provided in the HTML (HyperText Markup Language) tags.

3.2 Corpus Processing and Annotation

To work with the considerable quantity of corpora selected for this study, several tasks of processing and annotation were performed from the command line interface in *Cygwin* (screenshot shown in Figure 3.1), a Linux-like environment for Windows that allows users to access many standard UNIX utilities (Red Hat, 2015). By making use of shell meta-characters, which are characters with special meanings in the command line for matching patterns in file names, the same line of command can be applied at once to



```
Admin@Harn_laptop /home
$ sed -i -r -f proc_browntag brown_penntagged/*.txt
```

Figure 3.1. Screenshot of *Cygwin* command line interface

doi:10.6342/NTU201603145

multiple files in command line interfaces (Lu, 2014) such as *Cygwin*. Summarized below are the steps of corpus processing taken and the tools employed, as presented also in Figure 3.2. The primary steps include text processing (3.2.1), part-of-speech tagging (3.2.2), as well as sentence and phrasal alignment (3.2.3).

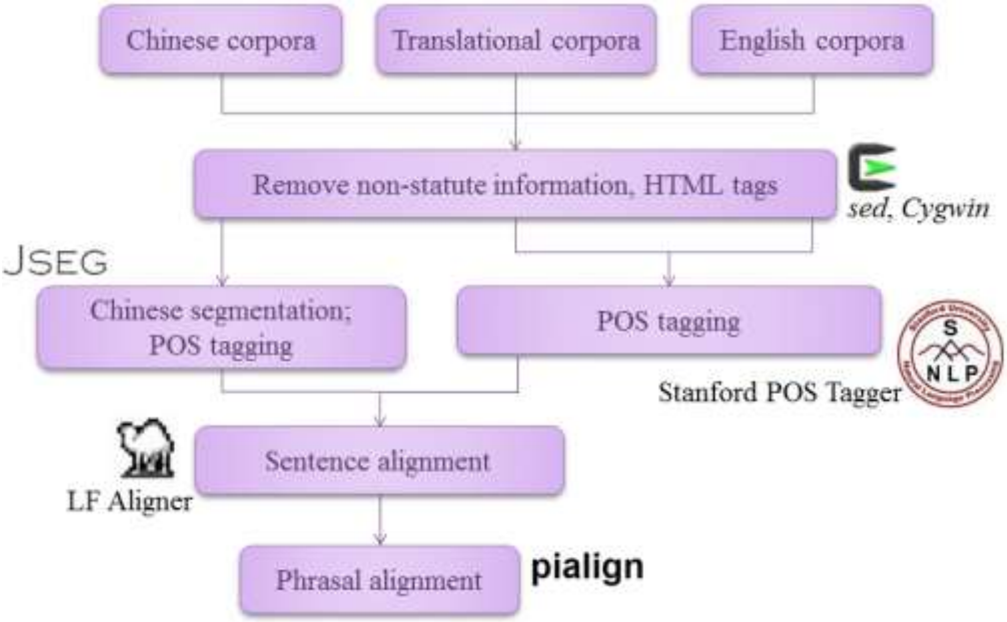


Figure 3.2. Flowchart of corpus processing procedures and tools

3.2.1 Text Processing

The retrieved and downloaded webpage files were processed in *Cygwin* environment to remove unnecessary information and convert from webpage to plain text format. As this study focuses on statute language alone, notes, source credits, tables and formulas in the *U.S.C.* are not included in the English corpora, while all tables and appendix file names were removed from the Chinese and translational corpora.

Identification and removal of non-statute information was completed semi-automatically with the *sed* command in combination with (extended) regular expression. *Sed* is a commonly used text-processing command with string pattern search, substitute, and delete functions (Barnett, 2015). Regular expressions, basic and extended, are special characters that can be used for specifying patterns (Lu, 2014).

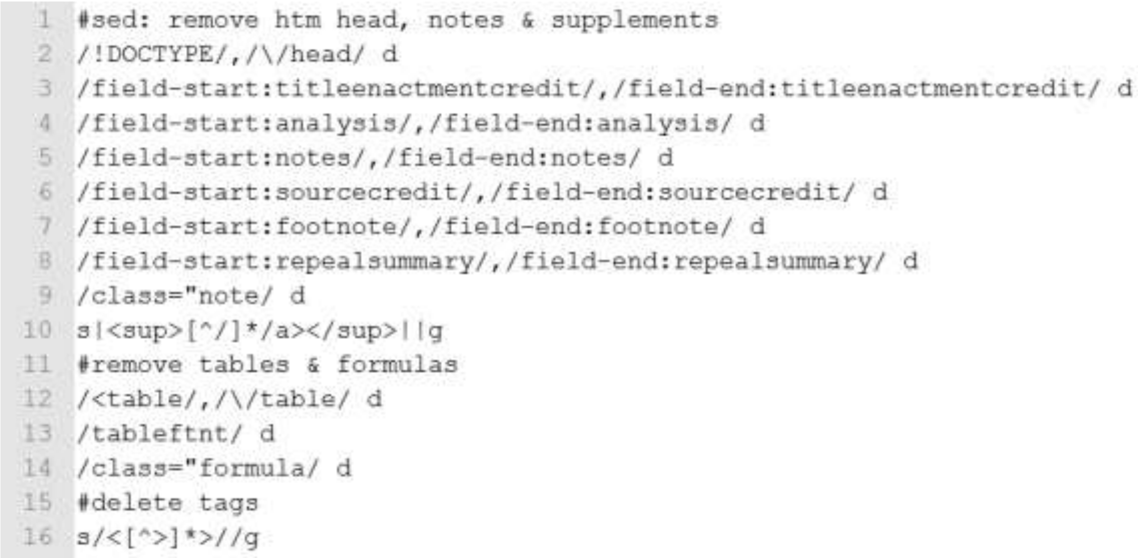
doi:10.6342/NTU201603145

Frequently used characters include positional anchors for specifying positions in a line; wildcards, such as the period, which matches any single character; characters for specifying various numbers of repetition; and expressions for character classes, such as alphabetic characters, digits, or all characters except specified exclusions.

Figure 3.3 shows a partial list of *sed* commands drafted for processing the English corpora, compiled into a file to be processed at once and performed on multiple files by specifying the “-f” and “-i” options. When information indicating notes, source credits, tables, or formulas, was found in the HTML tags, the appended “d” command instructs *sed* to delete the specified lines. For example, tags containing “class="note” indicate a line of notes, while tags containing “table” and “/table” indicate the start and ending lines of tables. Commands are therefore specified as follows to remove lines of notes and the ranges of lines from the start to ending lines of tables:

```
9   /class="note/ d
12  /<table/,/\/table/ d
```

To ensure intelligibility after processing, HTML decimal codes for special characters or symbols were then replaced with English letters, punctuation, and numbers.



```
1  #sed: remove htm head, notes & supplements
2  /!DOCTYPE/,/\/head/ d
3  /field-start:titleenactmentcredit/,/field-end:titleenactmentcredit/ d
4  /field-start:analysis/,/field-end:analysis/ d
5  /field-start:notes/,/field-end:notes/ d
6  /field-start:sourcecredit/,/field-end:sourcecredit/ d
7  /field-start:footnote/,/field-end:footnote/ d
8  /field-start:repealsummary/,/field-end:repealsummary/ d
9  /class="note/ d
10 s|<sup>[^/]*/a></sup>||g
11 #remove tables & formulas
12 /<table/,/\/table/ d
13 /tableftnt/ d
14 /class="formula/ d
15 #delete tags
16 s/<[^>]*>//g
```

Figure 3.3. Excerpt of *sed* commands for processing the English corpora

doi:10.6342/NTU201603145

For example, an apostrophe would be indicated by the string “&apos;” in HTML code. To display the punctuation form of apostrophes in the corpora, the substitute command of *sed* is specified to change occurrences of the string “&apos;” into an apostrophe symbol, and a global command “g” is tacked on to search the entire line for multiple occurrences, instead of moving on to the next line once an occurrence is found:

```
s/&apos;/'/g
```

To facilitate the subsequent part-of-speech tagging, lines were also combined when contents of the same sentence are spanned over more than one line; paragraphs containing more than one sentence were divided into multiple lines wherever possible.

A final text processing task is later performed after tagging and sentence alignment but prior to the extraction of phrasal alignments: all three sets of corpora were processed to edit out the list item markers at the beginning of lines, while in-text numerals were replaced with the hash symbol (#). These items were edited because while headings, listings, and numerals are useful in the sentence alignment process, they do not contribute to the primary object of subsequent analyses.

List item markers, which can take the forms of digits, roman numerals and alphabetical letters, often cause confusion for the POS tagger and therefore lead to mismatching in later analyses by computerized tools. For example, the tagger can be unable to annotate all items on the same numeral list consistently, resulting in some markers being tagged as cardinal number (CD), while others are deemed as list item markers (LS). The list item marker (a) is sometimes mistaken for a determiner (DT) or noun (NN); problems in distinguishing list item markers from other words go on to influence word frequencies, keyword analysis, n-gram/cluster frequencies, and concordance matches. In-text markers and numerals were replaced with the hash symbol (#) because this study is more interested in the general patterns associated with list item markers or numerals, rather than the actual marker or numeral that occurs. With all numerals represented by the same symbol, patterns are also more likely to surface.

Corpora data after text processing therefore add up to approximately 2.2 million Chinese characters of Chinese corpora, 1.9 million English words of corresponding translational corpora, and approximately 20 million English words of non-translational English corpora.

3.2.2 Part-of-Speech Tagging

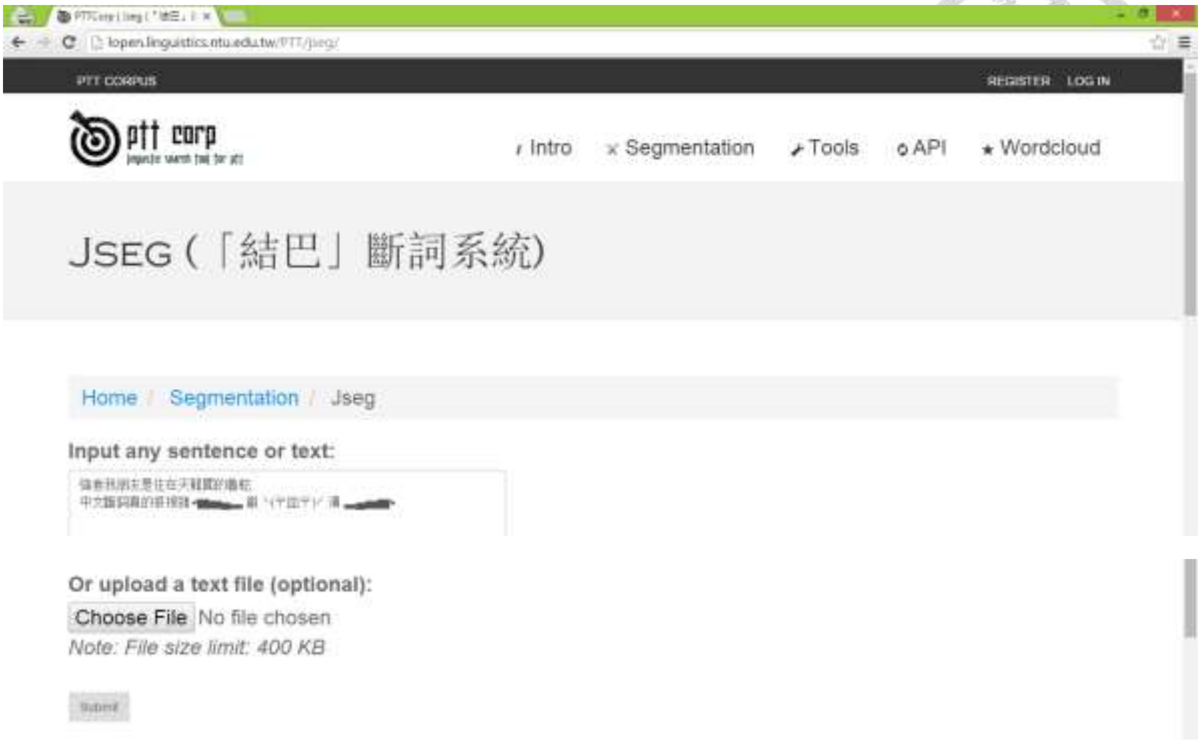
To prepare the Chinese texts for phrasal alignment, the Chinese statutes in this study were processed by *Jseg*, an automatic Chinese segmentator modified from *Jieba* (Sun, as modified by Liu, 2014). *Jseg* defines "word" boundaries and annotates the texts with POS tags. The program was trained with corpora from the Academia Sinica Balanced Corpus; algorithms of the Brill Tagger were incorporated to provide a POS-tagging feature trained on corpora from the Sinica Treebank.

For this study, the segmentator was accessed through the web interface of PTT Corpus (http://lopen.linguistics.ntu.edu.tw/PTT/jseg/), a dynamic corpus designed to automatically collect, update, and process data from the bulletin board system PTT (screenshot of PTT Corpus interface shown in Figure 3.4 on p. 48).

POS tagging of English texts in this study, including the English and translational corpora, were performed by the *Stanford Part-Of-Speech (POS) Tagger* 3.5.1 (Toutanova, Klein, Manning, & Singer, 2003). According to assessments by the developers, the tagger achieves per-position tag accuracy up to 97.24% with a model pre-trained on the Penn Treebank Wall Street Journal (WSJ) Corpus.

The tagset employed for denoting POS category is the Penn Treebank tagset (Santorini, 1990), originally designed for the large annotated corpus Penn Treebank of 4.5 million words in U.S. English (Marcus, Santorini, & Marcinkiewicz, 1993).

doi:10.6342/NTU201603145



Input any sentence or text:

[illegible]
[illegible] [redacted] [illegible] [redacted]

Or upload a text file (optional):

No file chosen

Note: File size limit: 400 KB

Figure 3.4. Screenshot of PTT Corpus web interface (http://lopen.linguistics.ntu.edu.tw/PTT/jseg/)

Developed based on the Brown Corpus (Francis & Kucera, 1964) tagset, the Penn tagset employs a reduced number of tags by eliminating redundancy, eliminating inconsistencies, encoding by syntactic functions, and avoiding indeterminacy (allowing for multiple tags). Instead of the original 87 in Brown, the Penn tagset comprises 36 POS tags and 12 tags for punctuation and currency symbols (Marcus et al., 1993). A list of the POS tags is shown in the Appendix.

To process the large quantities of texts in this study, the English Tagger was called in *Cygwin* environment and set to take each line as a sentence with the option "-sentenceDelimiter newline," considering that statutes contain many headings and listed items that are not always marked with line- or sentence-ending punctuation.

3.2.3 Sentence and Phrasal Alignment

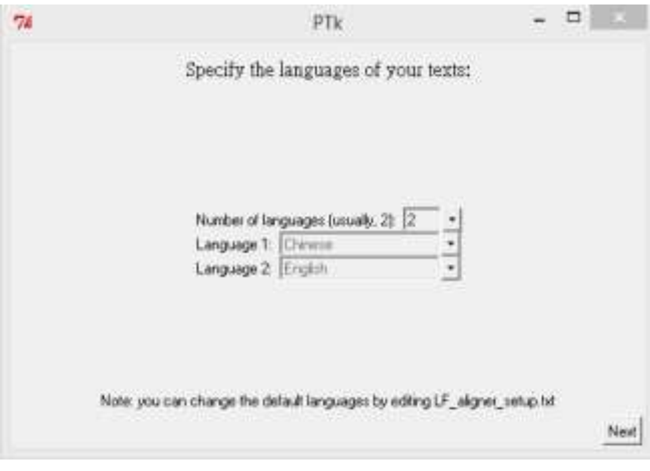
After removing POS tags to avoid reducing the effectiveness of the automatic sentence aligner, approximate sentence alignment is constructed between a portion of

doi:10.6342/NTU201603145

the Chinese and translational corpora. *LF Aligner* version 4.1 (screenshot shown in Figure 3.5) is a program intended for helping translators create translation memories (TM) from unaligned, previously completed work (Farkas, 2015). The tool incorporates the algorithm of *Hunalign*, an automatic bilingual aligner at sentence level (Varga et al., 2005), which identifies and bounds corresponding sentences based on sentence length and a dictionary, which the *LF Aligner* supplies for 32 languages and can be expanded or improved by the user if necessary.

Results generated by the *LF Aligner* are a list of aligned sentences in common TM format, with an Excel version provided by option. A cursory check was then conducted for misaligned sentences and manual adjustments were made to improve accuracy as much as possible, so as to enhance the chances of obtaining better results in the subsequent phrasal alignment process.

The aligned texts were then used to extract phrasal alignments using the automatic alignment tool *pialign* (Neubig, 2012), an ITG-based phrasal aligner used to create phrase tables for the translation model of an SMT system. The aligner adopts a context-free, language-independent, and fully statistical approach for calculating the probabilities of a phrase pair being the translation of one another. The statistics are then displayed with the potential phrase pair among the generated alignment results.



Specify the languages of your texts:

Number of languages (usually, 2): 2
Language 1: Chinese
Language 2: English

Note: you can change the default languages by editing LF_aligner_setup.txt

Figure 3.5. Screenshot of LF Aligner user interface

As *pialign* does not distinguish between words (characters, numbers) and punctuation, however, consecutive words in different sentence parts or even sentences, though separated by punctuation, can still be extracted as part of the same alignment entry. Because word or phrasal alignments are generally expected to be more useful to the translator, *pialign* results containing punctuations between the words were simply filtered out from the list of results.

3.3 Corpora Analysis

Making use of the corpora as compiled and processed with methods summarized in the previous sections, analyses were conducted in the ways and with the tools introduced below. Keyword analysis (3.3.1) identified indicators for potentially interesting directions for investigation; attempts were then made to identify terminology equivalents and translation units on the basis of selected keywords (3.3.2); usage patterns associated with stylistic features of the legal language were explored in 3.3.3; and finally, additional observations were attempted by using the translational and non-translational corpora in conjunction (3.3.4). A flowchart of the analysis process is shown in Figure 3.6.

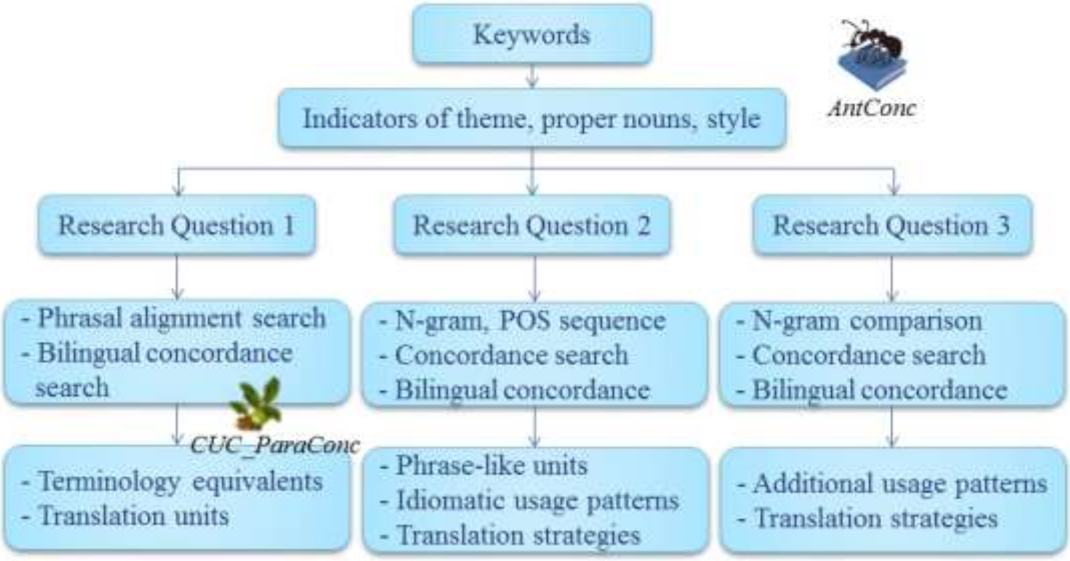


Figure 3.6. Flowchart of analysis process and tools

doi:10.6342/NTU201603145

3.3.1 Keyword Analysis

To quickly grasp an idea of the possible indications of proper nouns, theme, and style, keyword analysis (Scott, 2000) was respectively conducted on the processed English and translational corpora. The aim was that by automatic comparison of frequency data, interesting points for further exploration will emerge (Flowerdew, 2012) to facilitate subsequent analyses. The reference corpus chosen for this study is the Brown Corpus (Francis & Kucera, 1964), a general corpus of approximately 1 million English words (features summarized in Table 3.2).

Despite its limitation in size and time coverage, the Brown Corpus is the most accessible and practical choice to non-academic users when compared to other general English corpora. The corpora are available in full texts, enabling the necessary manipulation to suit the needs of different studies and corpus-based approaches. To facilitate comparison between the specialized and reference corpora, the Brown Corpus was re-tagged with the *Stanford Tagger* and Penn Treebank tagset before the analysis process in this study.

Keyword analysis was performed by the keyword list tool of *AntConc* 3.4.3

Table 3.2

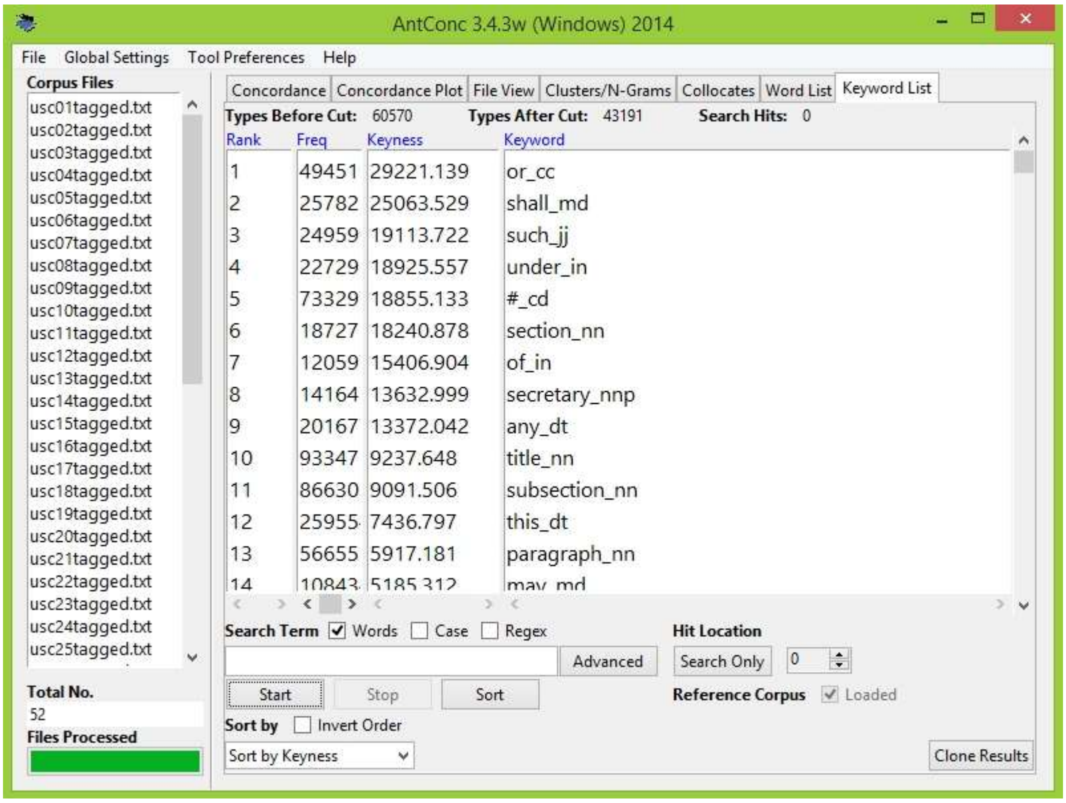
Specifics of the Brown Corpus

	Brown Corpus
Content	Non-translational English texts of United States press reportage, editorial, and reviews; religion; skill and hobbies; popular lore; belles-lettres; government and house organs; academic knowledge; fiction (general, mystery, science, adventure, and romance); and humor
Size	1 million words
Representativeness	General English of the United States
Publication Time	Jul. 1958-Jan. 1962

doi:10.6342/NTU201603145

(Anthony, 2014a), a freeware which incorporates a number of tools for conducting corpus-based research (Anthony, 2014b). Settings were adjusted to include the underscore "_" and tag as part of the words. Keyword lists were generated respectively for the English and translational corpora based on log-likelihood ratio (LLR), the default and recommended significance test for calculating "keyness," or keyword strength (Anthony, 2014b; Dunning, 1993; Rayson et al., 2004). The user interface of the *AntConc* keyword list tool is shown in Figure 3.7.

After excluding results containing numbers, punctuation, and other symbols, keywords with keyness below the critical value 15.13 were also omitted, retaining only keywords that can be deemed with 99.99th percentile certainty to be a significant difference between legal corpora and the reference corpus (Rayson & Garside, 2000). The frequency threshold of keywords was set at 3 occurrences, adopting the criteria recommended in Scott and Tribble (2006). Keywords that occur exclusively in the



Corpus Files

usc01tagged.txt
usc02tagged.txt
usc03tagged.txt
usc04tagged.txt
usc05tagged.txt
usc06tagged.txt
usc07tagged.txt
usc08tagged.txt
usc09tagged.txt
usc10tagged.txt
usc11tagged.txt
usc12tagged.txt
usc13tagged.txt
usc14tagged.txt
usc15tagged.txt
usc16tagged.txt
usc17tagged.txt
usc18tagged.txt
usc19tagged.txt
usc20tagged.txt
usc21tagged.txt
usc22tagged.txt
usc23tagged.txt
usc24tagged.txt
usc25tagged.txt

Total No. 52

Files Processed

Types Before Cut: 60570 Types After Cut: 43191 Search Hits: 0

Rank	Freq	Keyness	Keyword
1	49451	29221.139	or_cc
2	25782	25063.529	shall_md
3	24959	19113.722	such_jj
4	22729	18925.557	under_in
5	73329	18855.133	#_cd
6	18727	18240.878	section_nn
7	12059	15406.904	of_in
8	14164	13632.999	secretary_nnp
9	20167	13372.042	any_dt
10	93347	9237.648	title_nn
11	86630	9091.506	subsection_nn
12	25955	7436.797	this_dt
13	56655	5917.181	paragraph_nn
14	10843	5185.312	may_md

Search Term ☑ Words □ Case □ Regex

Hit Location 0

Reference Corpus ☑ Loaded

Sort by □ Invert Order

Sort by Keyness

Figure 3.7. Screenshot of *AntConc* keyword list tool. The interface shows the keyword list generated for the English corpora.

translational corpora were also identified by comparing the translational keyword list against the English corpora frequency list.

Part-of-speech distributions of the keywords were calculated by totaling the token frequencies of keyword part of speech. For easier observation, POS tags were roughly categorized into nouns, verbs, adjectives, adverbs, prepositions, determiners, conjunctions, modals, pronouns, and foreign words. POS distributions of translational keywords and English keywords were calculated separately. Part-of-speech distribution was also tallied for keywords that occur exclusively in the translational corpora.

The keyword lists of the English and translational corpora, particularly the top-ranking keywords and along with keyword part-of-speech distributions, were used for making preliminary observations as to what translators will likely come across in working with legal texts. Selected keywords that may be of particular interest were then explored through other statistical techniques and with computational linguistics tools so as to address the research questions proposed in Chapter 1.

3.3.2 Terminology Equivalents and Translation Units

As suggested in previous studies, parallel corpora are an effective tool for identifying terminology equivalents (Bowker & Pearson, 2002; Toyama, 2011), with word alignments being even more advantageous than sentence alignments in the case of bilingual concordance programs (Gale & Church, 1991a). Based on the knowledge that nouns and adjectives provide more likely indicators of terms than words of other categories (Voutilainen, 2003), this study selected noun and adjective category keywords for identifying terminology equivalents and associated translation units from the phrasal alignment and sentence alignment results.

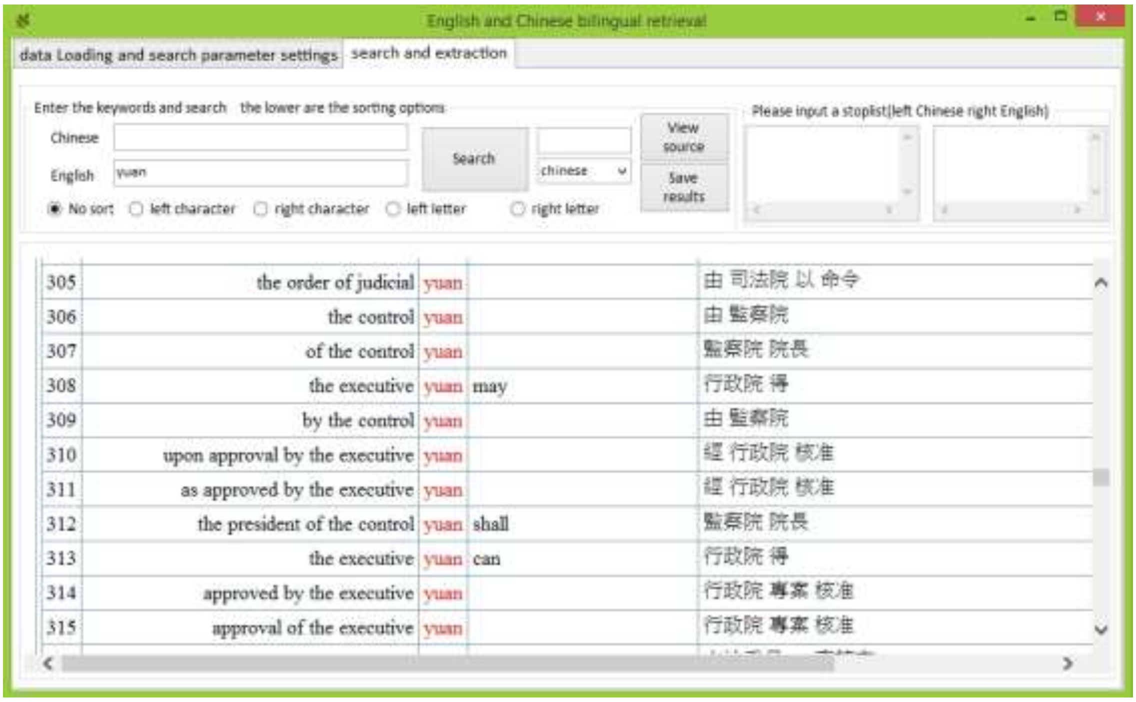
Two types of keywords were selected for searches of terminology equivalents. Indicators of theme or “aboutness” were selected from content words among the

doi:10.6342/NTU201603145

high-frequency keywords of the translational corpora. Indicators of proper nouns were obtained from the list of translational keywords that do not occur at all among the English corpora.

With the exception of phrasal alignment search on entries containing keywords specific to the translational corpora, bilingual searches in either phrasal alignment results or sentence-aligned parallel corpora were conducted in *CUC_ParaConc V0.3* (N. Cheng, 2013), a screenshot of which is shown in Figure 3.8. *CUC_ParaConc* is a parallel-corpus retrieval program that accepts parallel corpora aligned at any level as supplied by the user, and supports bilingual and multilingual search functions with monolingual or multilingual search words (Cheng & Hou, 2012).

Due to the quantity of translational keywords absent from the English corpora as well as limitations of corpora size and software capacity, a batch search for the proper noun indicators is handled with the *sed* command in *Cygwin* environment. The search items were listed with the print command “p” and processed with the “-n” option, which


English and Chinese bilingual retrieval
data Loading and search parameter settings
search and extraction
Search
View source
Save results
Please input a stoplist(left Chinese right English)

Enter the keywords and search the lower are the sorting options

Chinese

English yuan

● No sort ○ left character ○ right character ○ left letter ○ right letter

chinese

305	the order of judicial	yuan		由 司法院 以 命令
306	the control	yuan		由 監察院
307	of the control	yuan		監察院 院長
308	the executive	yuan	may	行政院 得
309	by the control	yuan		由 監察院
310	upon approval by the executive	yuan		經 行政院 核准
311	as approved by the executive	yuan		經 行政院 核准
312	the president of the control	yuan	shall	監察院 院長
313	the executive	yuan	can	行政院 得
314	approved by the executive	yuan		行政院 專案 核准
315	approval of the executive	yuan		行政院 專案 核准

Figure 3.8. Screenshot of *CUC_ParaConc* bilingual search and retrieval interface. The results shown are those of a phrasal alignment search.

doi:10.6342/NTU201603145

prevents *sed* from outputting lines unless a “print” request is supplied. Matching results containing the specified search items were then copied to a designated output file.

Possible terminology equivalents obtained from phrasal alignment search, either through *CUC_ParaConc* or the *sed* string-matching function, were then studied and selected in terms of their correctness or usability. By using partially correct alignments or abbreviations, searches were attempted to identify the full corresponding equivalent or proper noun through sentence-based bilingual concordance. Also, because phrasal alignment results generated by *pialign* can range from lengths of single to several tokens, some of the results include other co-occurring words and extended collocation. Once amended in the same way, these search results provide lengthier translation units associated with key terminology that are readily usable.

3.3.3 Exploring Stylistic Features

As pointed out in Scott (2000), indicators of style identified through the keyword approach often appear to be function words with unusually high frequencies, therefore not likely ideal candidates for identifying terminology equivalents. However, translators require more than bilingual dictionaries to complete their jobs, and some of the ways in which corpora can serve as useful reference tools include informing writing style and idiomatic usages (Bowker & Pearson, 2002).

Phraseology, as Stubbs (2001) pointed out, is an important subject of linguistics, and corpora can facilitate study on these recurring, multi-word phrasal units of natural-sounding language use. This study therefore investigated n-grams and concordances associated with style indicators as an approach to exploring useful collocation, colligation, and other usage patterns in legal English.

N-grams and monolingual concordance were studied with the aid of the *AntConc* n-gram/cluster tool and concordance tool, respectively. POS-tagged versions of the

doi:10.6342/NTU201603145

English corpora were used for studying n-grams to better observe colligation patterns and POS sequence where relevant. To exclude n-grams spanning different sentences or sentence parts, the corpora were processed to have line breaks are inserted after punctuation marks. During analysis, the line break replacement option was cancelled in the settings of the *AntConc* n-gram/cluster tool. The number of texts containing the found n-gram entries is also provided by the tool, helping to eliminate results that may be specific to only certain topics or authors (law drafters, translators). A possible starting point for analyses is n-grams of three or more sequential words, occurring at least 20 times per million words across five or more different texts, as recommended for lexical bundles in Biber and Conrad (1999), Biber, Conrad, and Cortes (2004).

Monolingual English concordances obtained by the *AntConc* concordance tool were sampled with the method provided in Sinclair (2003), aiming to extract samples evenly distributed over all texts in the corpora. A batch of 25 samples was taken for each object of study; the first sample is selected at random among the 4% of all generated concordances, and each sample afterwards is selected automatically after skipping 4% of concordance hits since the previous selected instance. The 4% gap between concordance samples was calculated for each searched item by dividing the number of all found samples by 25.

Analyses were then attempted following the instructions of Sinclair (2003) and Hunston and Francis (2000). The sampled concordances were observed for conspicuous patterns that surface on either side of the queried keyword. Endeavors were made to formulate hypotheses on usage patterns associated with the style indicator in question, taking into account the part of speech of the combination of words as well as the meaning of the keyword. A summary was then attempted to describe the idiomatic usage of the identified patterns in a legal context.

doi:10.6342/NTU201603145

The above sampling method was also applied to bilingual concordance lines from the parallel corpora, extracted instead with *CUC_ParaConc* and studied in a similar fashion for translation strategies associated with the selected style indicators. Revealing strategies that previous translators have used to overcome constraints imposed by the source texts is an important function that parallel corpora serve in the translator training environment which supplements the features of comparable corpora (Bernardini et al., 2003; Pearson, 2003).

3.3.4 Utilizing Translational and Non-translational Corpora

Statistical machine translation systems rely on not only a translation model for identifying the appropriate word sets translated from the input text, but also a language model which ascertains the correct word sequence in the target language (Somers, 2003). Similarly, translators can turn to parallel corpora for identifying terminological equivalents (Bowker & Pearson, 2002), but monolingual corpora in the target language are often found useful for providing information on “native-like” means of expression (Bernardini et al., 2003). It is therefore deduced that resources of parallel and monolingual corpora can be used in combination to provide more comprehensive information to the translator.

As proposed in Baker (1995) and confirmed in Laviosa (1998), comparison of translational corpora and non-translational corpora will reveal patterns specific to translated texts. In addition to the keyword lists (as described in Subsection 3.3.1), therefore, terminology equivalents, translation units, and style-related usage patterns identified were also used in this study as starting points of comparison between the translational and English corpora. By conducting n-gram and concordance searches associated with the identified equivalents and patterns, efforts were made to identify additional phrase-like units and information that can aid the process of legal translation.

doi:10.6342/NTU201603145

In the case of terminology equivalents and translation units, comparisons of their frequencies in the translational and English corpora will help verify which of the corresponding word sets are likely preferred or more common in legal English, while possibly uncovering similar usable phrase-like units. Concordance searches can further ascertain the contexts in which these word sets are used and whether or not these contexts are similar to one another or associated with specific phrasal units.

Comparison between bilingual and English concordance lines containing the same patterns, whether terminology or style related, will help determine if certain translation strategies are preferred over others when aiming to achieve idiomatic usage appropriate in legal English. It is also possible that additional translation strategies will be deducible from comparable English concordance results that are not apparent by simply observing the translational corpora.

Based on an initial keyword analysis, therefore, the above methods were used to identify and explore terminology equivalents, translation units, style-related patterns, other phraseology features, and translation strategies from within the parallel and non-translational corpora, with an aim to summarize useful information and provide insights to legal translators. The results obtained from this process will be presented and discussed in the next chapter.

doi:10.6342/NTU201603145

Chapter 4 Results and Discussion

Using the methods and computerized tools introduced in the previous chapter, this study investigated the keywords, terminology equivalents, and selected patterns in user-compiled corpora, focusing on features that may be of interest to the legal translator. The results are elaborated in this chapter and discussed with an aim to address the research questions proposed in Chapter 1. Section 4.1 summarizes the findings and observations concluded from keyword analysis of the translational and English corpora. On the basis of those observations, Section 4.2 explores the identification of terminology equivalents associated with high-frequency keywords; Section 4.3 turns to usage patterns related to stylistic features of legal texts. Lastly, Section 4.4 will explore the use of translational and non-translational corpora in conjunction for additional information on legal English.

4.1 Keyword Analysis and Preliminary Observation

By comparing the translational and English corpora against the Brown Corpus, the *AntConc* keyword list tool identified 2,909 translational and 3,032 non-translational keywords at 15.13 log-likelihood ratio or above. The 15.13 threshold indicates a 99.99th percentile certainty that the keywords identified are significantly over-represented in the specialized corpus (Rayson & Garside, 2000). Among the identified keywords, all exceed the threshold of at least 3 occurrences, and 1,481 are found on both lists. The top 20 keywords in the translational and English corpora are shown in Table 4.1 (on p. 60).

Part-of-speech distribution was observed to be fairly similar between keywords of the translational and English corpora. As seen in Figure 4.1 (on p. 61), nouns constitute the highest proportion of tokens among both sets of keywords, yet function words account for more than 40% of keyword tokens. Based on the conclusions of Scott (2000), it is likely that indicators of style would be identifiable among the high-ranking

doi:10.6342/NTU201603145

Table 4.1

Top 20 Keywords in Translational and English Corpora

Translational Corpora					English Corpora			
Keyword	POS	%	Keyness	Rank	Keyword	POS	%	Keyness
shall	MD	1.91%	31727.44	1	or	CC	2.41%	29221.14
article	NNP	1.67%	29672.49	2	shall	MD	1.26%	25063.53
or	CC	2.14%	18749.2	3	such	JJ	1.22%	19113.72
the	DT	9.34%	9621.818	4	under	IN	1.11%	18925.56
paragraph	NN	0.47%	8244.984	5	section	NN	0.91%	18240.88
be	VB	1.67%	8035.006	6	of	IN	5.89%	15406.9
authority	NN	0.47%	7633.576	7	secretary	NNP	0.69%	13633
competent	JJ	0.43%	7523.038	8	any	DT	0.98%	13372.04
preceding	VBG	0.32%	5510.283	9	title	NN	0.46%	9237.648
may	MD	0.62%	5271.416	10	subsection	NN	0.42%	9091.506
of	IN	4.88%	5140.082	11	this	DT	1.27%	7436.797
person	NN	0.27%	3621.288	12	paragraph	NN	0.28%	5917.181
by	IN	1.07%	3405.231	13	may	MD	0.53%	5185.312
central	JJ	0.23%	3241.17	14	the	DT	8.08%	5026.09
accordance	NN	0.18%	3079.058	15	states	NNPS	0.33%	4578.839
regulations	NNS	0.18%	3056.9	16	for	IN	1.58%	4451.831
act	NN	0.22%	3049.965	17	united	NNP	0.31%	4046.806
apply	VB	0.19%	2977.322	18	by	IN	1.01%	3918
provisions	NNS	0.18%	2938.495	19	chapter	NN	0.20%	3717.662
authorities	NNS	0.17%	2797.166	20	term	NN	0.20%	3668.407

Note. Parts of speech of keywords are represented by tags, the descriptions of which are provided in the Appendix.

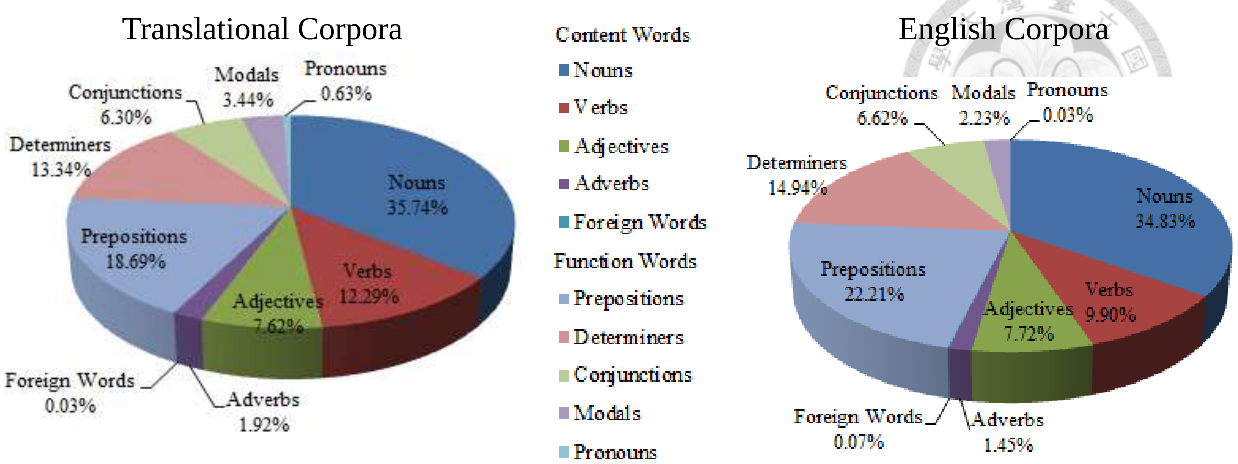


Figure 4.1. Keyword part-of-speech distributions. Distributions are shown for keywords in the translational and English corpora.

function words in the keyword lists, while content words, nouns in particular, will be associated with proper nouns or indicative of common themes in legal texts.

Also found among the 2,909 translational keywords were 154 that are not included at all in the English corpora, and 36 which are annotated with a different POS tag when occurring in the English corpora. Part-of-speech analysis of the former 154 keywords (Figure 4.2) shows that up to 90% of their occurrences are tokens of nouns or proper nouns in either singular or plural form. A hypothesis was therefore formed that keywords occurring exclusively in the translational corpora correspond to or compose the English translations of terms specific to Taiwan, and can be used to identify existing, common terminology equivalents or translation units associated with proper nouns.

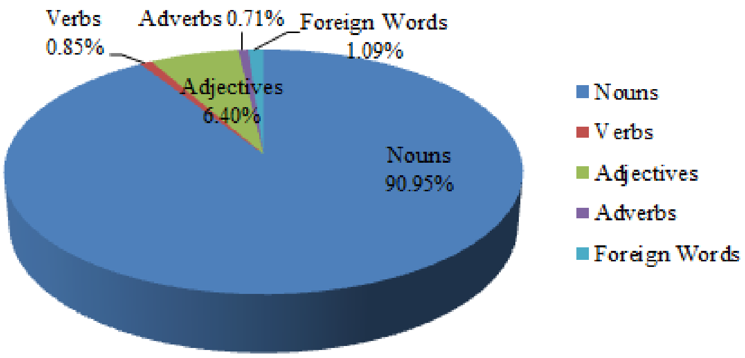


Figure 4.2. POS distribution of keywords unique to translational corpora.

4.2 Terminology Equivalents and Translation Units

According to observations made from keyword analysis results in the previous section, two types of content words were found that may be of particular interest to the legal translator for identifying terminology equivalents. The first type are content words that rank high among the translational keywords, including words referencing grammatical agents such as a person or relevant authorities, as well as words that point to certain provisions. The second type consists of nouns among the translational keywords that do not occur in the English corpora and likely correspond to translations or partial translations of proper nouns specific to Taiwan. The following subsections will address the first research question by exploring the identification of terminology equivalents and translation units based on these two types of content words.

4.2.1 Theme-related Terminology

From preliminary observations, the top keywords that can be associated with grammatical agents frequently referred to in legal provisions include the nouns “person” as well as the singular and plural forms of “authority.” By utilizing the bilingual concordance program *CUC_ParaConc*, searches in the phrasal alignment results quickly confirm that the above keywords likely constitute fixed translations for some frequently recurring terminology. A preliminary search for the word “person” yielded 1,339 entries of possible terminology candidates; the number doubles when a search for the same word was conducted among the sentence alignment results, which would also require considerable more time to sift through due to the lengths of the results.

As an example, Table 4.2 (on p. 63) shows 20 automatically extracted samples of phrasal alignment results containing the word “person.” While exactly correct matches are relatively few in this particular batch of examples (three out of 20, excluding the usage in the sense “body” in sample 13), several terminology equivalents were quite

doi:10.6342/NTU201603145

Table 4.2

Sample of Alignment Results Containing "Person"

No.	Chinese	English	Precision
1	人	**person** ~~in~~	△
2	已 +	~~the person~~ no longer	╳
3	~~之~~人	**person**	△
4	**第三人**	~~in a~~ **third person**	△
5	任何 +	any person ~~shall~~	△
6	**委託人**	**trust person**	○
7	**參選人**	**person planning + + + +**	△
8	外國 法人	**alien legal person** ~~are~~	△
9	**受保護人**	**the protected person**	○
10	身體 健康 + +	a person + + + +	╳
11	各 款 **人員**	each **person** whom + + + +	△
12	相對人 為	person to whom administrative guidance is	╳
13	**身體** 、 **物件**	**person** , **property**	○*
14	+ + ~~或 其他 得為 訴訟 當事人~~	+ person ~~or unincorporated~~	╳
15	攜帶 刀械 ++	any person who carries knives	△
16	**負責** 之人 ~~有~~	**person in charge**	△
17	~~但~~ **財團法人** ~~經~~	**foundational judicial person** ~~where~~	△
18	~~依本法 規定~~ **參加 行政 程序** 之人	**person intervening into administrative procedures**	△
19	**學校 財團法人** ~~於 申請~~	**school 's judicial person** ~~during the~~	△
20	**歸化** 人 之 **未婚 未成年 子女**	**unmarried minor children of a naturalized person**	○

Note. Correctly matching words and phrases are shown in bold; plus signs (+) indicate missing elements; strikethroughs are applied to unnecessary elements. Symbols are used in the Precision column to indicate correct (○), partially correct (△), and incorrect (╳) matches; an additional asterisk (*) indicates that the search word is used in a different sense than in the other results.

doi:10.6342/NTU201603145

easily obtained from the 12 partially correct entries. From the correct and partially correct samples, therefore, it can be deduced that the keyword “person” corresponds to the Chinese character “人” in a majority of cases, while the following eight terminology equivalents can be acquired, along with two additional terminology equivalents that are not directly associated with the search word:

●	第三人	third person	[JJ NN]
●	委託人	trust person	[JJ NN]
●	外國 法人	alien legal person	[JJ JJ NN]
●	受保護人	the protected person	[DT VBN NN]
●	負責 之人	person in charge	[NN + IN NN]
●	財團法人	foundational judicial person	[JJ JJ NN]
●	學校 財團法人	school 's judicial person	[NN POS JJ NN]
●	歸化 人	naturalized person	[JJ NN]
●	行政 程序	administrative procedures	[JJ NN]
●	未婚 未成年 子女	unmarried minor children	[JJ JJ NN]

Though POS tags were not available in the search and study of parallel corpora due to alignment technicalities, the above terminology entries are found to be fairly similar in structure. With the exception of one “noun-prepositional phrase” structure, as is labelled at the end of said entry, most of the terminologies consist of one or two modifiers (adjectives or past participle verbs in the English translation, with one example of noun plus possessive ending) preceding a noun. These results confirm the observation that nouns and adjectives provide more likely indicators of terms than words of other categories (Voutilainen, 2003). Their grammatical structure can be summarized in Penn tagset format as:

(DT) (JJ) JJ/VBN NN

Some partially correct phrasal alignments, while unable to yield terminology equivalents in themselves, can nevertheless be utilized to identify terminology equivalents in searches among sentence alignment results. One example of such a case is Sample 7 in Table 4.2 above. With the information "參選人" and the words "person planning" in the corresponding English sentence, the complete and correct translation unit was identified as:

擬 參選人 (a) person planning to participate in campaign

The fact that "person" often translates the Chinese character "人" in the parallel corpora can also be used to further narrow down search results among phrasal alignment entries. By supplying both a Chinese and English keyword in *CUC_ParaConc*, the number of possible candidates for terminology equivalents is cut down to 417, which eliminates more than two-thirds of phrasal alignment entries that are likely mismatches or information unrelated to the keyword in discussion.

Authorities are another type of grammatical agents frequently seen in legal texts, indicated by the singular and plural forms of the word "authority," both of which identified among the top 20 translational keywords. The above search methods found that the two keywords often translate the term "機關" (establishment, institution). Relevant terminology equivalents thus obtained include:

●	(中央) 主管 機關	(central) competent authority
●	行政 機關	administration/administrative authority
●	衛生 主管 機關	(competent) health authority
●	當地 主管 機關	local competent authority
●	監督 機關	supervisory authority

● 外交 機關	foreign affairs authorities
● 政府 公產 管理 機關	public property management authorities
● 各級 政府 機關	all levels of government authorities

Similar to the results associated to “person,” terminology equivalents containing the Chinese word “機關” as well as the singular or plural forms of “authority” in their English components are mostly structured as noun phrases with the keyword as head, preceded by modifying adjectives or nouns, which often provide information regarding the powers or responsibilities of the authority. The exception is the final example, which presents as a “noun phrase-prepositional phrase” (JJ NNS + IN NN NNS).

Terminology equivalents identified in this way also co-occur with extended collocation to form translation units, many identifiable from phrasal alignment results.

High-frequency content words that point to references of legal provisions in the translational keyword list include the nouns “article,” “paragraph,” “regulations,” “act,” and “provisions.” As with the case with person- and authority-related terminology, a brief search for each of the above nouns revealed recurring Chinese equivalents in four of the five cases that will likely narrow down search results for obtaining extended translation units. Both “條例” (ordinance) and “法” (law) were found to frequently give rise to “act,” but Chinese phrasal units associated with “regulations” were too varied to narrow down the search results without imposing limitations. Table 4.3 (on p. 67) shows a selection of alignment entries that include extended collocations, identified by the five translational keywords above and their frequent Chinese correspondences below:

● article:	第# 條	article #
● paragraph:	項	
● act:	條例; 法	

doi:10.6342/NTU201603145

- provisions:　規定

Table 4.3

Selection of Alignment Results Concerning Legal Provisions

No.	Chinese	English Translation
1	適用 第# 條 之 規定	**article** # **shall apply** to
2	第# 條 規定	the **provisions** stated in **article** #
3	依 第# 條 辦理	conducted in **accordance** with **article** #
4	第# 條所 規定 之	set forth in **article** #
5	第# 條之# 之罪	the offenses prescribed in **article** #
6	第# 項 各 款	subparagraphs of **paragraph** #
7	前 項 續聘	reappointment as specified in the **preceding paragraph**
8	第# 項 教育 課程	course indicated in the first **paragraph**
9	第# 項 證據 資料	evidence materials stipulated in **paragraph** #
10	前 項 重利	usurious interest under the **preceding paragraph**
11	其 建築 管理 辦法	**regulations** on management of these buildings
12	履約 爭議 調解 規則	**regulations** governing the mediation of disputes
13	本法 所稱	referred to in this **act**
14	依 勞工 保險 條例	according to the labor insurance **act**
15	本 條例 未 規定	matters not provided for in this **act**
16	除本法 另 有 規定 外	unless otherwise provided by this **act**
17	優先 適用 本 條例	this **act shall** prevail
18	本法 所定 事項	matters described in this **act**
19	本法 施行 細則	enforcement **regulations** for this **act**
20	依 下列 規定 計算	calculated pursuant to the following **provisions**

Note. Words in bold indicate a ranking among the top 20 keywords of the translational corpora; other recurring usage patterns associated with the queried words are underlined.

The above selection confirms that phrasal alignment results containing law-referencing keywords can often be found to include extended collocation. These keywords and co-occurring phrase-like units form translation units which share similarities in structure and usage on the part of the English translations. Apart from the examples associated with “regulations,” which tend to take supplementary information succeeding itself, structures of the co-occurring collocation appear to fall roughly under into three categories:

- model-verb phrase: shall apply to (適用); shall prevail (優先 適用)
- prepositional phrase including keyword
- past participle verb or gerund preceding prepositional phrase including keyword

The third category, in particular, was found in association with an assortment of different verbs that can be viewed as synonyms. Alternatively, phrase-like units including “stated in,” “set forth in,” “prescribed in,” and other similar-meaning VBN-preposition (past participle verb plus preposition) combinations can be described as commonly preceding a noun phrase containing a keyword associated with legal provisions. This type of phrase-like units often form translation units in which the corresponding Chinese includes some variation of “依 [...] 規定” (according to provisions [...]), though sometimes reference is implicit in the Chinese.

On a side note, the above examples also evidence that strong connections often exist between multiple items among the top-ranking translational keywords. The word “preceding,” which often modifies “article” or “paragraph” to reference previously stated provisions, for instance, is also identified among the top 20 translational keywords. The same applies to the word “accordance,” which recurs in the phrasal unit “in accordance with” prior to provision-referencing nouns, as well as the words “shall”

doi:10.6342/NTU201603145

and "apply," oftentimes found after the nouns. The need to frequently reference sources of legal provisions, along with the co-occurrence tendencies of the said keywords, may contribute to their significantly higher frequencies in legal texts.

4.2.2 Taiwan-specific Proper-nouns

As was observed in Section 4.1, out of the 2,909 translational keywords identified, 154 are unique to Taiwan statutes, up to 90% of which are nouns or proper nouns in singular or plural. The higher ranking keywords (as excerpted in Table 4.4) show a tendency to indicate existing translations or abbreviations of government agencies or

Table 4.4

Excerpt of Keywords Specific to Taiwan Laws

Keyword	POS	Rank	Frequency	%	Keyness
yuan	NNP	35	1855	0.10%	1747.32
nt$	NN	57	1293	0.07%	1217.944
roc	NN	338	306	0.02%	288.237
motc	NNP	435	243	0.01%	228.894
macau	NNP	690	145	0.01%	136.583
bas	NN	744	134	0.01%	126.222
councilors	NNS	745	134	0.01%	126.222
expiry	NN	799	123	0.01%	115.86
ntd	NNP	842	114	0.01%	107.382
chunghwa	NNP	886	107	0.01%	100.789
cmo	NNP	936	100	0.01%	94.195
broking	NN	1214	72	0.00%	67.821
prestation	NN	1218	72	0.00%	67.821
dgt	NNP	1317	64	0.00%	60.285

doi:10.6342/NTU201603145

organizations in Taiwan. For example, the word “Yuan” is used for the five highest-ranked government bodies in Taiwan, as in “Executive Yuan” or “Legislative Yuan;” MOTC stands for the Ministry of Transportation and Communications, and DGT refers to the “Directorate General of Telecommunications” under the MOTC.

By comparing the 118 unique nouns or proper nouns against the phrasal alignment results in *Cygwin* environment, 2,067 entries were automatically extracted that contain at least one of the 118 keywords. Like in the case of theme-related keywords explored in the previous subsection, not all the alignment results are immediately usable, though the precision rate of the below examples is rather high as compared to that of the theme-related samples associated with “persons.” Table 4.5 (shown on p. 71) contains 1% (20 samples) of the extracted entries, selected automatically at a fixed interval; in this particular batch, six of the samples are correct matches, four are incorrect matches, while the remaining 10 are partially correct, either containing excessive words or missing certain elements as compared to the corresponding Chinese.

The correct and partially correct matches, again, help to identify terminology equivalents of and translation units containing proper nouns after selection and manual correction, sometimes with further searches in the sentence alignments conducted through *CUC_ParaConc*. Examples of such cases include samples 4 and 16, which can be revised after the concordance search to yield the following translation units:

4	報請 行政院 核定 之	submitted to the Executive Yuan for approval
16	應 分配 之 當選 名額	allocated quota of electees

The number of extracted results, manageable for human review, produced approximately 200 entries of terminology equivalents and translation units useful for future reference, excluding most of the entries that provided repeated information. The

Table 4.5

Sample of Alignment Results Containing Keywords Specific to Taiwan Laws

No.	Chinese	English Translation	Precision
1	之	under executive yuan	╳
2	**中華民國**	of **the roc**	△
3	~~之~~ **行為**	~~electee conducts~~ **the action** ~~prescribed in~~	△*
4	報請 **行政院**	+ to **the executive yuan**	△
5	其所	broking agencies	╳
6	**由 行政院** 以 命令 **定**	**determined by the executive yuan** + +	△
7	**中華民國 領域**	**territory of the roc**	○
8	**立法委員 選舉**	**election of members of the legislative yuan**	○
9	**當鋪 業**	**pawnshop**	○
10	**仲介** 或	**broking** ~~**or** the~~	△
11	**參事** #人	+ **councilors** ~~and #~~	△
12	**由 行政院 訂定**	**decided by the executive yuan**	○
13	行政院 及 所屬 +	+ + of agencies **of the executive yuan**	△
14	**教師 、 教保員**	**teachers , educare givers**	○
15	收到 被	functionary	╳
16	分配 之 當選 +	+ + of electees ~~distributed to~~	△
17	當地國	draftee	╳
18	依 金門 馬祖 東沙 南沙 地區	+ **kinmen , matsu , dongsha and nansha** +	△
19	**移轉 於 當鋪 業**	**transferred to the pawnshop** +	△
20	**花蓮縣 政府 及 臺東縣**	**hualien county government and taitung county**	○

Note. Correctly matching words and phrases are shown in bold; plus signs (+) indicate missing elements; strikethroughs are applied to unnecessary elements. Symbols are used in the Precision column to indicate correct (○), partially correct (△), and incorrect (╳) matches; an additional asterisk (*) indicates that the match does not include the intended search word.

doi:10.6342/NTU201603145

majority of these translation units are associated with government agencies or organizations (excerpt shown in Table 4.6), relevant government posts, locations in

Table 4.6

Alignment Results of Government Agencies and Organizations

Chinese	English Translation
內政部	Ministry of the Interior (**MOI**)*
公務員 懲戒 委員會	Public Functionary Disciplinary **Sanction** Commission
司法院	the Judicial **Yuan**
外交部	Ministry of Foreign Affairs (**MOFA**)*
立法院	Legislative **Yuan**
全國 營造業 工地 主任 公會	National Construction Industry **Jobsite** Directors Union
考試院	Examination **Yuan**
行政院	Executive **Yuan**
行政院 人事行政局	Personnel Administrational Executive **Yuan**
行政院 公共 工程 委員會	Public Construction Commission, Executive **Yuan**
行政院 原住民族 委員會	Council of Indigenous Peoples, Executive **Yuan**
行政院 海岸 巡防署	Coast Guard Administration, Executive **Yuan**
行政院 勞工 委員會	Council of Labor Affairs, Executive **Yuan**
行政院 新聞局	Information Office of the Executive **Yuan**
法務部 矯正 署	**MOJ** Agency of Corrections
客家 委員會	**Hakka** Affairs Council
國防部	Ministry of National Defense (**MND**)*
採購 申訴 審議 委員會	Complaint Review Board for Government Procurement (**CRBGP**)*
監察院	Control **Yuan**

Note. Boldface indicates the keywords specific to the translational corpora used to identify the translation units. An asterisk (*) indicates information obtained through additional search.

doi:10.6342/NTU201603145

Taiwan, and relevant terms or immediate context of the extracted terms or proper nouns. With some of the agencies, only the English abbreviations were extracted by the phrasal alignment software, in which case bilingual concordance searches were relied on to obtain the full translations, as indicated with the asterisks in the table above.

An example of relevant terms being extracted in the vicinity of proper nouns is shown in Table 4.7. The samples selected are entries containing the term “司法院” (Judicial Yuan), one of the five highest-ranked government bodies in Taiwan. Contexts in the alignment results obtained this way also reveal terminology information on posts or personnel associated with the Judicial Yuan (“justice,” “president of the Judicial Yuan”), relevant agencies (the “Ministry of Justice” was originally established under the Judicial Yuan, and still exercises jurisdiction over judicially-related affairs), and jurisdictional matters that fall under the Yuan’s responsibilities (“interpretation of the Judicial Yuan,” “transfer judges”).

Table 4.7

Alignment Samples Containing “Judicial Yuan”

Chinese	English Translation
司法院	Judicial Yuan
司法院 **大法官**	**justices** of the Judicial Yuan
司法院 及 **法務部**	Judicial Yuan and **Ministry of Justice**
司法院 **院長**	**President** of the Judicial Yuan
司法院 得 **調派 法官**	the Judicial Yuan may **transfer judges**
司法院 以 **命令 定 之**	the Judicial Yuan shall **mandate**
司法院 **解釋**	**interpretation** of the Judicial Yuan

Note. Words in bold indicate the additional information obtained from context.

4.3 Stylistic Features: Modals

Keyword analysis, as seen in Section 4.1, has shown that function words account for 40% of keywords in both the translational and English corpora, and that these function words likely include indicators of stylistic features in legal texts (Scott, 2000). Among the function words that rank within top 20 of the translational and non-translational keywords, modal are a significant component near the top of both lists. “Shall,” in particular, is ranked 1st and 2nd among the two keyword lists respectively, while “may” is ranked the 10th and 13th.

In addition to their significant frequency difference in statutory texts and in the Brown Corpus, occurrences of the words “shall” and “may” also account for considerably larger portions among modal verbs in the specialized corpora than their occurrences in the reference corpus. From the percentages of their frequencies as shown in Table 4.8, it is apparent that usage of the majority of modals is fairly distinctive in the legal context. In the following subsections, usage patterns of “shall” and “may” are studied through n-grams and concordance samples to explore their differences from general-purpose language use and relevant information usable to the legal translator.

Table 4.8

Percentage of Frequent Modals in the Reference, Translational and English Corpora

Modal	Brown Corpus	Translational Corpora	English Corpora
Would	20.34%	0.29%	3.00%
Will	16.82%	3.96%	4.17%
Can	13.28%	3.26%	0.98%
Could	12.00%	0.35%	0.38%
May	9.76%	21.13%	26.54%
Shall	2.01%	64.80%	63.14%

doi:10.6342/NTU201603145

4.3.1 Shall

The specificity of “shall” usage in legal contexts can be observed by comparison to its usage in a general corpus. A search for trigrams beginning with a modal “shall” identified just over 100 combinations in the non-translational English corpora that meet the criteria of 20 occurrences per million words and across five texts. Taking into account the POS tags of the n-gram results, the grammatical structures of frequent “shall” trigrams are strikingly homogeneous.

As an example, Table 4.9 below shows a list of the eight most frequent trigrams beginning with “shall.” The basic structure of the n-grams can be represented as:

- shall (not) VB

A majority of the combinations has the word “be” occurring as the verb in base form succeeding “shall,” either directly or after an adverb “not.” This pattern is then followed by a past-participle verb to form a passive voice structure, or by an adjective:

Table 4.9

Top Eight Trigrams Beginning with “Shall” in English Corpora

MD	(RB)	VB	Collocate	POS	Frequency	Range
shall	not	be			9981	52
shall	not	apply			5622	50
shall		be	made	VBN	4905	50
shall		be	treated	VBN	3735	44
shall		be	subject	JJ	3152	48
shall		be	construed	VBN	3097	49
shall		be	deemed	VBN	2983	50
shall		submit	to		2722	46

Note. The range column refers to document frequency, or number of documents the n-gram occurs in; the total number of documents is 52.

doi:10.6342/NTU201603145

- shall (not) be VBN
- shall (not) be JJ

This same POS sequence is shared by all of the most frequent “shall” trigrams.

The Brown Corpus, on the other hand, does not contain a single modal “shall” trigram that meets the criteria of 20 occurrences per million words across five texts, being a general corpus instead of representing any one specific register. From the trigrams that could be found to occur in at least two texts, it was observed that while they share the same POS sequence as legal English trigrams, “shall” in the Brown Corpus seems to be less often succeeded by passive voice and tends to co-occur with vocabulary of a different register.

The first of the two differences observed above are confirmed by using concordance searches, which showed that the structure “be + VBN” occur 7% more frequently after a modal “shall” in the legal English corpora, including occurrences with and without an adverb in between. No additional results were identified for reversed subject and modal order followed by passive voice usage in the general corpus.

Concordance and bigram searches also discovered that collocation of “shall” apparently differs in general purpose language and legal language. For example, the phrase-like units “shall apply” and “shall not apply,” which recur frequently throughout the legal English corpora (245 and 281 occurrences per million words), were found only once in the Brown Corpus and not at all outside the “Miscellaneous: Government & House Organs” category. Co-occurrence of the verb “forget” with “shall,” though rare in the Brown Corpus, was not found at all in the English corpora. Co-occurrence of the words “show” (VB), “find” (VB), and “never” (RB) are also rarer in proportion in the non-translational legal texts.

Study of sampled concordance lines confirmed the basic grammatical structure succeeding “shall” that was identified during n-gram analysis. Contextual information drawn from the concordance samples also made it possible to further elaborate on the pattern so that it applies to a wider range of situations. In addition to the possibility of containing an adverb “not” within the “shall + VB” structure, it was observed that a selection of other adverbs may be used in place of “not.” This pattern is discernable from the following concordance samples containing “shall” followed by an adverb (extended information are omitted so as to show the sentence components directly pertaining to the use of “shall”):

```
1  expenses paid to or on behalf of [...]  shall  not  exceed the aggregate of
2 the postmaster at the place [...] who  shall  promptly  notify the sender of said
3 if found to [...], the State inspector  shall  so  notify the postmaster at the
```

In addition to “so” and “promptly,” which occur in the above examples, bigram searches also identified “also,” “only,” “immediately,” and “annually” as some of the more frequent adverbs to be found in the same pattern. A sampling of concordance lines containing the “shall-adverb” pattern further confirms that the basic structure of “shall” usage can be revised to include more adverbs other than “not,” while revealing a few more adverbs that apply to this pattern:

```
1   voluntary agencies and cooperatives  shall  also  be eligible to receive
2  such investigation , and the Secretary  shall  immediately  begin a study of --
3 Board of Governors and the Corporation  shall  jointly  issue final rules implement
4 The leave described in this paragraph  shall  only  be available during a single
5    , which modifications or revisions  shall  thereafter  be treated as a part of
```

- shall (RB) VB: “immediately begin,” “jointly issue”
- shall (RB) be VBN: “be treated”
- shall (RB) be JJ: “be eligible,” “be available”

A comparison of the previously sampled concordance lines against concordances

doi:10.6342/NTU201603145

sampled from the Brown Corpus quickly show a distinctive trait of “shall” usage in legal texts with regard to sentence subjects; whether in active or passive voice, all the sentence subjects (indicated with waved underline in the concordance samples) have thus far been noun phrases and none of them first or second person pronouns, which are quite often found in the Brown Corpus outside the government documents category:

```
1         . When it comes to this , I  shall  prefer emigrating to some country
2  is the strength of my life ; of whom  shall  I be afraid '' ? ( Psalm 27 : 1
3    core for all undergraduates ? Or  shall  we permit early specialization
4      was right when he said , `` We  shall  never negotiate out of fear
5    experiment . Mathematically , we  shall  not distinguish the experiment
6  jumping platform , aku . Later , you  shall  know it better . Is it not well-
```

In *The Cambridge Grammar of the English Language,* Huddleston and Pullum (2002) associated the meaning of “shall” when taking a first or second person subject with indication of speaker’s guarantee, instruction-seeking questions, and non-deontic uses denoting futurity, consequence, or volition/determination. The above sample 4 from the Brown Corpus can be seen as an example of speaker’s guarantee; samples 2 and 3 are instruction-seeking questions; samples 1 and 5 can be interpreted as expressing volition; while sample 6 most likely denotes futurity.

In total, more than 36.9% of the “shall” occurrences identified through bigram searches are directly preceded by first or second person pronouns in the Brown Corpus. “Shall” is also succeeded by first or second person pronouns in some cases, such as in the questions mentioned above. The combinations “we shall,” “shall we,” “shall I,” and “you shall” are not found at all in the legal English corpora, however; only the combination “I shall” occurs twice (out of almost 258 thousand “shall” occurrences), both of which being in oaths.

The vast majority of “shall” usages in the legal English corpora fall under the constitutive or regulative use of "shall," described by *The Cambridge Grammar of the English Language* as occurring frequently in legal or quasi-legal documents and associated with third-person subjects. These usages, as was observed in the legal English concordance lines above, appear to be in association with the prescribing of directives or obligations to take a specified (course of) actions.

The English legal corpora enabled identification of usage patterns of "shall" as well as their differences from general-purpose language use. On the other hand, parallel corpora offered an opportunity to study the strategies that legal translators have employed when the modal is needed.

Bilingual concordance samples showed that "shall" is also used with third person subjects in the translational corpora, indicating the employment of the word in its constitutive or regulative sense. The modal is frequently found when the corresponding Chinese segment either comprises the Chinese words "應" (should, is to be), "不得" (shall not), or conveys a sense of prescribing rules without an explicit Chinese equivalent present. For example:

1	訴願 有 理由 者. 受理 訴願 機關 **應**以 決定 撤銷 原 行政 處分 之 全部 或 # 部. 並 得視 事件 之 情節. 逕為 變更	appeal is sustainable, the agency with jurisdiction of administrative appeal **shall** revoke the administrative action as a whole or
2	所管 公有 土地. 非經 該 管區 內 民意 機關 同意. 並經 行政院 核准. **不得** 處分 或 設定 負擔 或為 超過 #年 期間	public lands under the jurisdiction of the Municipal, or County (City) Government **shall not** be disposed of, or encumbered, or leased
3	本館 掌理 下列 事項 .	The NTM **shall** be in charge of the following matters

As "shall" translates the prescription of laws or directives by direct statement or specifying a prohibition, with or without a particular Chinese word to explicitate the meaning, translators often insert the "shall" when this regulative sense is implied, in addition to using the modal as an equivalent of sorts to "應" and "(不)得." Conversion

doi:10.6342/NTU201603145

to passive voice is another strategy often observed in translating legal stipulations; from the examples below, the contexts in which such a strategy are considered most likely involve reversed sentence structures or covert agents:

1	前 項 甄試 審查 委員會 委員. 由 司法院 指派 人員 並 遴聘	The Judicial Yuan **shall** appoint [...] to be members of the Review Committee
2	學士、碩士、博士 學位 由 大學 授予 .	degree, master's degree, and doctor's degree **shall** be conferred by universities.
3	且 審判 程序 尚未 終結 或 違反 組織 犯罪 防制 條例 案件 者 外. 至遲 **應**於 資料 製作 完成 時起 #年 內 銷毀 之.	Data preserved as per the proviso to paragraph # **shall** be destroyed no later than # year after they are needed unless they are required for
4	前 項 委任 **應** 提出 委任書 狀於 檢察 官 或 司法 警察 官. 並 準用 第# 條	A power of attorney **shall** be presented to public prosecutor or judicial police officer
5	前 項 情形. **應** 將 委託 事項 及 所 依據 之前 項 規定 公告 之. 並 刊登 於	case as described in the preceding Paragraph, the authority in charge **shall** make a public announcement specifying the matters delegated

Though all conveyed in active voice in the Chinese source text, translators have opted for a switch to passive voice in three of the above five examples. The first two examples, while explicitly specifying "司法院" (Judicial Yuan) and "大學" (universities) as the overt agents in the Chinese sentences, have moved the direct object up to the beginning of the sentences, inserting the preposition "由" before the overt agent, which would have started the sentences had they retained the traditional subject-verb-object structure. The translator therefore had to either revert back to the traditional order as in sample 1, or change to passive voice in order to move up the verb complement, as was done in sample 2.

Meanwhile, the translators were prompted to opt for passive voice in two of the latter three examples as no overt agent was stated in the Chinese source text. Interestingly, however, sample 5 adopted a strategy of supplying an overt agent for the English translation, possibly deduced from the extended context in nearby sentences, in order to retain the active voice. This strategy is arguably applicable to other cases where directives are set forth without mention of the overt agent. If such information can be uncovered in the nearby context, sample 4 might, for instance, be translated in the active voice by supplying an overt agent such as “the applicant” or “the principal.”

4.3.2 May

In the same way as “shall,” the other frequently used modal “may” can, too, be examined for legal-text usage, patterns, and translation strategies using n-gram and concordance searches.

A total of 18 trigram combinations containing the modal “may” were found in the non-translational English corpora along with three types in the Brown Corpus that meet the criteria of 20 occurrences per million words over five texts. The variety aspect aside, grammatical structure or POS sequence of “may” trigrams appear at first glance to be fairly similar to the findings on “shall.” As seen from the eight most frequent trigrams beginning with “may” in the legal English corpora and Brown Corpus (shown in Table 4.10 on p. 82), with only one exception in the general English corpus, structures of the majority of trigrams can be represented as:

- may (RB) VB / be + VBN/JJ

Upon closer inspection, however, it was observed that though they appear in very similar structures, “may” uses in the top trigrams might in fact be associated with two different senses of the word. The most frequent combinations in the two sets of corpora, “may be necessary” and “may have been,” for example, most likely indicated a sense of possibility, while “may not exceed” and “may enter into” in the legal context could point to permission or prohibition. With the majority of the trigrams, however, it was difficult to discern one way or the other without more context. Further verification is required, which can be achieved through concordance searches.

doi:10.6342/NTU201603145

Table 4.10

Top Eight Trigrams Beginning with "May" in English Corpora and Brown Corpus

English Corpora								Brown Corpus						
MD	(RB)	VB	Collocate	POS	Freq.	Range	Rank	MD	(RB)	VB	Collocate	POS	Freq.	Range
may		be	necessary	JJ	4194	50	1	may		have	been	VBN	36	31
may	not	be			3571	50	2	may	not	be			33	30
may		be	made	VBN	2477	50	3	may		be	a	DT	26	21
may		be	used	VBN	2083	45	4	may		be	made	VBN	15	12
may	not	exceed			1489	44	5	may		have	to	TO	13	11
may		be	provided	VBN	1009	44	6	may	not	have			11	11
may		enter	into	IN	954	45	7	may	or	(CC)	may	(MD)	11	9
may		be	required	VBN	940	45	8	may	also	be			11	8

Note. The range column refers to document frequency, or number of documents the n-gram occurs in; the total number of documents is 52 texts of legal English corpora and 500 in the Brown Corpus.

Sampled concordance lines from the English corpora showed that a large portion of the "may" usages indeed resemble the pattern observed above, which is similar to the use of "shall." Samples 1 and 2 below are examples of such a structure, in which a verb in base form (active voice) or "be + VBN" combination (passive voice) succeeds the modal "may" to denote a permissible option or entitlement, while samples 3 and 4 demonstrate insertion of an adverb into the same structure to indicate denial of permission or granting of conditional permission for a course of action.

1	members of the naval service of [...]	**may**	be assigned to United States commands
2	, and such individual	**may**	sue in a State or Federal court of
3	, such activities	**may**	not be undertaken after the effective
4	support agreement under [...]	**may**	only be used when the Secretary
5	calculated in [...] , as the case	**may**	be , shall be based on the number of
6	perform such [...] , as the Chairman	**may**	assign to them , and , upon request

A considerable number of concordance lines, like the above samples 5 and 6, however, reveal another pattern of “may” usage in legal English contexts. In such cases, “may” is preceded by an “as” and noun phrase (NP) and succeeded by a base form verb to denote the sense of possibility. The entire pattern, often inserted within a sentence in the form of a supplementary condition, conveys the meaning that there are multiple possibilities to a determining factor, depending on the outcome or situation as specified. This pattern can be expressed as follows:

- as NP may VB / be + VBN

Examination of the structure “may be + JJ,” which appears to be a recurring pattern in the legal English corpora as the trigram frequencies would indicate, also found examples of “may” usages in both its permission giving and possibility sense. Its permission granting use, such as in sample 1 below, turned out as being rather rare with a “be-adjective” combination. The majority of the samples are associated with the possibility sense. Samples 2 and 3 take on a pattern similar to the above “as NP may VB” combination to express the idea of a condition that differs according to circumstances. In many occasions, elaborative information on the adjective is further provided in the form of a tacked on prepositional phrase, as seen in samples 1, 3, and 5:

- as may be + JJ + (prepositional phrase)

The fourth and final samples can be seen as interesting extensions of the above usage pattern. From their contexts, both seem to mean something along the lines of “to the extent that” the succeeding condition is possible.

```
1  a general aviation airport with [...]  may  be exempt from having to accept schedu
2   the production of any documents as    may  be reasonable and necessary , shall
3  exemption of any securitization , as   may  be appropriate in the public interest
4  provisions of [...] shall , so far as  may  be practicable , apply to any bonded
5                     ; and insofar as    may  be consistent with the performance of
```

The patterns discussed above are rather infrequent in the Brown Corpus, as was also discovered through concordance sampling. While largely still following the "MD (RB) VB / be + VBN" pattern identified earlier, the majority of "may" usages in this general language corpus are in its sense of possibility. Samples 2-5 below all fall under this category; samples 3 and 5 provide usage examples of the more frequent trigrams "may have been" and "may or may (not)" observed at the beginning of this subsection. Sample 6 yields an entirely different usage of "may," in the sense of expressing a wish or prayer, which has not been observed in the legal English corpora.

```
1                      He says : `` We  may  further grant to those of her ( Poetry
2   , there is no telling how far it  may  go . Inmates might even demand the
3          the decision in retrospect  may  have been a wise one .
4  absurdity of that contention . You  may  have misgivings about certain aspects
5    development . Though this may or  may  not be good biology , it does aptly
6       I now offer this to you , and  may  this food fill up the ten quarters of
```

Another difference between "may" usages in the specialized and general corpora, as revealed in the concordance samples, is their co-occurrence tendency with first, second, or third person subjects. The tendency for "may" to co-occur with third person subjects in legal English is supported by bigram frequencies. While the first and second person pronouns "I," "we," and "you" all rank among the 10 most frequent collocates preceding "may" in the Brown Corpus, they are very scarcely found in the same position in the legal English corpora. Only one occurrence of "I may" was found in an oath, as well as 32 occurrences (0.03% of all collocates preceding "may") of "you may," which occurred in disclosure statements or notices demanded by law in certain situations of contract-signing or judicial procedure. It is surmised that this occurrence tendency may be associated with an assumption that target audiences for these texts would include non-experts of legal language.

Bilingual concordances show that “may,” consistent with its use in the non-translational legal texts, is often used to translate the granting or denial of permission and frequently corresponds to “得” (or “不得”) in Chinese. As seen in the concordance sampled below, the translations generally follow the basic “may (RB) VB / be + VBN” pattern that recurs in the non-translational corpora. Sample 6 is an example of the less frequently found usages of “may” expressing possibility in the parallel corpora. The two “as (NP) may VB” patterns indicating possibility are also more infrequently found in the translational corpora.

1	本會 因 業務 需要 . 經 行政院 核准 . **得** 聘用 顧問 或 研究員 .	The Council **may** appoint advisers or researchers as needed for the performance of its functions
2	以 議價 方式 辦理 之 採購 . **得** 免收押標金 .	where there is only # supplier invited for tendering , the bid bond **may** be waived .
3	大學 置 校長 #人 . 綜理 校務 . 負校務 發展 之責 . 對外 代表 大學 .	A university **may** appoint # president responsible for the overall management of the
4	私立 高級 中等 學校 **不 得**以 地 名為校名 .	Private senior high schools **may not** use place names as their school names.
5	對於 同# 被告 因 債權 及 擔保 該 債權 之 不動產 物權 涉訟 者 . **得** 由 不動產 所在地 之 法院 合併 管轄 .	In matters relating to [...], an action **may** be initiated against the same defendant in the court for the place where the real property is
6	其他 保護 子女 、 被害人 或 其他 家庭成員 安全 之 條件 .	Any other conditions that **may** be required to ensure the safety of children , victim and other

The above samples 3 and 4 show a certain degree of overlapping in the functions of “may” and “shall.” Where the regulative sense of a provision is implied in the Chinese instead of explicitly stated with the characters “應” or “得,” some interpretation is required on the translator’s part to determine whether to express the instructions in terms of an order, or to state it more mildly as though an entitlement, as was done in sample 3. Similarly, “不得” in sample 4 can arguably be interpreted as a prohibition and expressed in stronger terms (such as “shall not”) instead of the politer denial of permission approach that was adopted here.

The same dichotomy of active or passive voice that arises in “shall” usages also exists with the modal “may.” Samples 2 and 5 present examples of conversion to passive voice where the source segments were expressed in active voice, in sample 2 because no overt agent was given in the Chinese and in sample 5 to follow the word order of the source texts more closely. It is possible, therefore, that the same strategies of opting for active voice when an overt agent is specified or identifying the overt agent from context may still be applicable.

4.4 Translational and Non-translational Corpora in Conjunction

In the process of keyword analysis, terminology identification, and studying stylistic features, some differences have already surfaced between the corpora of translational and non-translational legal English. These differences are further discussed in the subsections below in terms of how they may provide more information on terminology equivalents or insights into writing style for the legal translator.

4.4.1 Extended Terminology Information

As was observed in Section 4.2, a considerable number of terminology equivalents were identifiable through theme-related keywords. While the terminology section focused mainly on translational keywords, such as the words “authority” or “authorities” for “機關” (establishment, institution) and “article” instead of “section” when referencing legal provisions, it was also noticed that occurrence tendencies of those keywords may differ in the translational and non-translational corpora.

In the case of “authority/authorities,” which identified many terminology equivalents associated with governmental departments, the same words were more

doi:10.6342/NTU201603145

frequently found to refer to the “power” or “jurisdiction” of authority departments or law in the English corpora. High-frequency examples include “under the authority” (1,006 occurrences) of specified provisions, “pursuant to the authority” (152 occurrences) of a legal source, or “have the authority” (269 occurrences). Usage of “authority/authorities” to indicate an authority department was also found, as in “local authority/authorities” (74 and 149 occurrences), “competent authority” (78 occurrences), or “taxing authority” (53 occurrences), but occurrences are rather rare in comparison.

Instead, the noun “agency,” which ranks 21st among the English corpora keywords, was found to be more frequently used in reference to authority departments. Frequent n-grams ending in “agency/agencies” include: “(local) educational agency/agencies” (3,559 and 1,329 occurrences), “federal agency/agencies” (2,015 and 1,826 occurrences), “state agency” (1,620 occurrences), and “department or agency” (1,323 occurrences), among others.

“Agency” was also found to translate “機關” (establishment, institution) in the translational corpora. The word ranks at 32th (noun, 2,371 occurrences) and 209th (proper noun) among the translational keywords, its associating n-grams including “government agency/agencies,” “regulatory agency,” “immigration agency,” “governing agency,” and “patent agency,” among others. However, with the exception of “government agency/agencies,” the above examples occur only in very limited numbers of texts, at most 7 out of the selected 510.

By comparison of the translational and non-translational corpora, therefore, it can be deduced that “authority/authorities” in Taiwan statute translations are to some degree comparable to “agency/agencies” in U.S. statutes. The words “agency” and “agencies” can, in turn, be used to conduct phrasal alignment and bilingual concordance search, and thereby identify more terminology equivalents for translators’ use. By using

doi:10.6342/NTU201603145

"agency/agencies" as identifiers in the legal English corpora, meanwhile, frequent collocation may be found that can apply to the use of "authority/authorities" when translating new terms associated with governmental establishments.

Search of additional terminology can also initiate from prominent features of the non-translational corpora. Among the top English corpora keywords, for example, another content word could be observed to indicate agents of authority, namely "secretary," which was tagged as proper noun in more than 99% of its occurrences. A search in the phrasal alignment results found that in the parallel corpora, "secretary" occurs in entries corresponding to "主任秘書," which is translated as "secretary-general" or "chief secretary." "Secretary-general" is also found in the English corpora, but occurrences are distinctively rare: 4 occurrences are found in association with the United Nations, and 3 with an organization, indicating that the majority of usages are highly specific to the U.S. legal context.

The most frequent words to co-occur with "secretary" show up in the English corpora as "secretary of the interior, "secretary of defense," "secretary of the treasury," "secretary of agriculture," "secretary of state," "secretary of transportation," "secretary of labor," "secretary of commerce," and "secretary of health," occurrences all numbering over a thousand and found in over 70% of the texts. By searching in the translational corpora with the above word strings following "secretary," several more words referencing agents of authority were located, including "ministry," "department," "council," "bureau," and "minister," which serve as useful search words for identifying additional terminology equivalents through phrasal alignment search.

4.4.2 Writing Style

Identification of terminology equivalents can extend to results of translation units, as seen in the discussions on translational keywords associated with referencing legal

doi:10.6342/NTU201603145

provisions. While a number of extended collocations were identified in 4.2.1, along with (partial) Chinese correspondences in some cases, it is less apparent as to which combinations of extended collocation might be more preferable.

Using the non-translational English corpora for verification of appropriate collocation and idiomatic usage, an approach suggested in Bowker and Pearson (2002), the extended collocation or partial translation units identified in 4.2.1 were compared against frequencies of n-grams associated with “section” or “subsection” in the non-translational corpora. The words “section” or “subsection” were chosen for the preliminary comparison for their comparability to “article” in the translational corpora.

As can be seen in Table 4.11, frequencies in the English corpora provided initial confirmation that some extended collocation may be more appropriate than others

Table 4.11

Extended Collocation and Frequency in English Corpora

(Partial) Translation Unit		Frequency	Range
(No equivalent)	as specified in	148	22
依 [...] 辦理	in accordance with	3598	49
規定	provided for	676	32
依 [...] 規定	pursuant to	2976	46
所稱	referred to in	2624	46
規定 之	set forth in	1257	44
適用	shall apply to	154	27
規定	stated in	15	5
N/A	as defined in*	5975	48
N/A	as provided in*	3167	50

Note. The range column refers to document frequency, or number of documents the n-gram occurs in; the total number of documents is 52.

within a comparable context. It was confirmed that phrase-like units such as “in accordance with” and “pursuant to” would likely be preferable over “stated in,” for example, when preceding the word “article.” A more unexpected discovery was “shall apply (to),” which was rather infrequent in the same context, though known to recur frequently in the non-translational corpora. It is surmised, therefore, that this combination is most likely appropriate elsewhere, and further concordance study on its usage may be advisable. N-gram frequencies have also identified two additional phrase-like units applicable to similar contexts, namely “as defined in” and “as provided in,” which the selected terminology equivalents did not happen to include.

With regard to stylistic features, non-translational corpora can, too, provide a complementing function to parallel corpora by providing examples of idiomatic usage required for translating contexts identified from bilingual concordances.

From discussions on translation strategies involving usage of the modals “shall” and “may,” the previous section identified the choice of active or passive voice as a recurring issue in legal translation. The sampled concordances give the impression that there is a tendency for converting active voice to passive in the translation process where “shall” and “may” (but to a lesser extent) are involved. The non-translational legal texts, meanwhile, do not seem to exhibit a tendency for passive voice in the usage patterns of these two modals.

To confirm whether a tendency exists in either way, the frequencies of passive expression in the two corpora are estimated with concordance searches of “shall” and “may” with a “be + VBN” pattern. It was found that approximately 43.6% of “shall” occurrences in the translational corpora are associated with the passive voice pattern, as opposed to only 28.7% in the non-translational legal texts. The percentages of passive voice against all “may” occurrences are 30.7% in the translational texts and 27% in the

doi:10.6342/NTU201603145

non-translational corpora.

The above estimations suggest that passive voice may tend to be overused in translating with “shall” in the legal context. In the concordance samples of non-translational corpora, “shall” is rarely used with passive voice, particularly if an overt agent is specified. After narrowing down the concordance results, it was found that 3.4% of the “shall” occurrences followed by “be VBN” is also immediately succeeded by the preposition “by,” though some of these occurrences specify the means of performing said action instead of an overt agent. The translational corpora, however, include approximately 11.6% of “shall” passive occurrences that precede “by.” It is therefore advisable that translators opt for reordering the sentence and using the “shall” in active voice where an overt agent is identifiable. Even when the relevant agent is not readily apparent in the immediate context, the information may be discernable in nearby source segments, allowing the translator to insert the overt agent as sentence subject.

Finally, the large volume of non-translational corpora can provide abundant usage examples that are not available in the translational corpora. In the first batch of “shall” (modal) concordance sampled for this study, for example, two collocations had surfaced that were not revealed in bilingual concordance samples:

1 the Administrator , upon receipt of [...] **shall** make available [...] all records

2 Any Tribal Action Plan [...] this section **shall** provide for -- the establishment of

The combination “shall make available” in sample 1 does not occur at all in the translational corpora, though a pattern “shall make NP available” of similar meaning is found with 3 occurrences. The phrase-like unit “shall make available,” however, occurs 548 times in the non-translational corpora with a fairly fixed pattern:

1 The Administrator **shall make available** to the panel any information

2 The amount that a [...] agency **shall make available** for supplemental educational

3 the Secretary **shall make available** information in [...] to a State

4 The Administrator of [...] **shall make available** the amounts appropriated pursua

- shall make available to [recipient] [direct object]
- shall make available [direct object] to [recipient]* (omissible)
- shall make available for [purpose]

Similarly, upon narrowing down concordance search results to “shall provide” in the English corpora, two more patterns have surfaced that are of common use in legal English: “shall provide for” (1,449 occurrences) and “shall provide that” (861 occurrences). The former appears to prescribe an obligation “to take measures” for achieving an objective, while the latter specifies a requirement “to stipulate” or include provisions on a certain matter.

1 The Administration **shall** provide for a central registration

2 Each contract or agreement **shall** provide that any person who enters

The two patterns above, presumably useful for translators to learn and imitate, by indication of their frequency in U.S. statutes, each occur only once in the translational corpora. Like “shall make available,” therefore, they would likely be easily missed or dismissed if one were to work only with translational corpora.

Chapter 5 Conclusion

In conclusion of the results discussed in Chapter 4, this chapter sums up the major findings of this thesis and attempts to address the research questions proposed in the Introduction chapter as well as implications of those findings. Section 5.2 summarizes the limitations of this study, and concludes with suggestions for future research.

5.1 Summary of Findings and Implications

This study has explored a number of methods and computerized tools in compiling and analyzing a Chinese-English legal corpus of legislation incorporating both parallel and monolingual corpora. The aim was to compile a specialized corpus with means and tools that are sufficiently convenient, efficient, and “automated” to be manageable to an individual, before demonstrating how the compiled corpus can be utilized for a number of linguistic investigations.

The subsections below will summarize findings of the previous chapters and how they may provide answers as to the facilitation of identifying terminology equivalents with computerized tools (5.1.1), exploration of stylistic features and patterns in legal corpora (5.1.2), and combined usage of translational and non-translational corpora (5.1.3). Implications of these findings will be summarized in 5.1.4.

5.1.1 Terminology Equivalents and Computerized Tools

Methods and tools for keyword analysis, word segmentation, sentence alignment, phrasal alignment, and bilingual concordance have constituted an integral part of this study as they were employed to address the first research question:

(1) How can corpus-based approaches and available computerized tools be utilized to facilitate identification of terminology equivalents and translation units for legal translation?

doi:10.6342/NTU201603145

Having started all subsequent analysis based on the results of keyword analysis, this study successfully identified indicators of corpus theme, proper nouns, and stylistic features through keyword analysis, achieving the functions introduced in Scott (2000).

Selected keywords of the translational corpora were used to identify terminology equivalents from the parallel corpora, the compilation of which relied heavily on a sentence aligner to speed up the process. Sentence alignment then served as the material for automatic phrasal alignment, which also required word segmentation of the Chinese texts to be pre-processed by a segmentator.

The above processing procedures, facilitated by computerized tools, enabled the use of parallel corpora for identifying terminological equivalents, as proposed in Bowker and Pearson (2002), much in the way that one would use a bilingual dictionary, an analogy drawn in Toyama (2011). Searches of the keyword “person,” for example, led to the identification of a series of terminology equivalents, including those for “legal person,” “judicial person,” and “naturalized person,” among others. The keywords “authority” and “authorities,” on the other hand, identified equivalents for government establishments or institutions.

It was also demonstrated that search of terminology equivalents could be expanded to include extended collocation and therefore translation units. Keywords associated with references to legal provisions, such as “article” and “paragraph,” identified not only terminology equivalents but also extended collocation (“in accordance with,” “as specified in”), their corresponding Chinese, and therefore lengthier translation units.

The search process above confirmed the advantage that word alignment provides in bilingual concordance by enabling searches with input terms in only one language instead of two (Gale & Church, 1991a); the speed for compiling bilingual terminology

doi:10.6342/NTU201603145

lexicons can be significantly improved with only partial results of word or phrasal alignment as compared to sentence aligned concordance (Dagan et al., 1993).

Similarly, terminology equivalents were identified for proper nouns by using keywords of the translational corpora that do not occur in the non-translational corpora. The equivalents thus identified include those of government agencies or organizations, relevant government posts, locations in Taiwan, and relevant terms or immediate contexts of the above categories. Whether in the case of theme-related or proper noun equivalents, bilingual sentence concordance was sometimes helpful for supplying elements mistakenly excluded in phrasal alignment results, so as to obtain the complete terminology equivalent or translation unit.

5.1.2 Stylistic Features and Patterns

(2) How can user-compiled corpora be utilized to investigate stylistic features and patterns specific to the legal genre?

In addressing the second research question, the present thesis focused on the specialized usage and patterns of two distinctive indicators of style in legal contexts: the modals “shall” and “may.”

Comparison of legal corpora against a reference corpus, also referred to as study on external variation of the legal language (Biel, 2010), can reveal interesting features of legal English, as was demonstrated in Coulthard and Johnson (2007). By exploring the phraseologies of “shall” and “may” in the legal English corpora and Brown Corpus, this study attempted to establish the specificity of “shall” and “may” usage in legal contexts while identifying idiomatic usage patterns and translation strategies regarding the two modals that would be useful to legal translators.

Statistical techniques including n-gram and PoS-gram approaches (Stubbs, 2007), concordance sampling and investigation (Sinclair, 2003), as well as pattern grammar

doi:10.6342/NTU201603145

(Hunston & Francis, 2000) all proved to be effective methods for the research described above. A basic structure of “MD (RB) VB” (modal, optional adverb, verb base form) was deduced for usage of both modals. In legal English, “shall” is almost always used in its regulative sense; this inference was supported by the co-occurrence of “shall” usages with third person subjects. Examples showed that “may” was used to expresses both permission and possibility; the pattern “as (NP) may VB” was also identified in association with usage in the latter sense. The POS tags were especially useful in this process as they provided readily available information for discerning POS patterns and a means of conveniently narrowing down search results in computerized tools, as was predicted in McEnery (2003) and Reppen (2010).

Bilingual concordance allowed for observations on translation strategies (Pearson, 2003). “Shall” translates the prescription of laws or directives, often corresponding to the Chinese “應” and “(不)得,” but also frequently inserted in the English translation when the regulative sense is only implicit in the source text. Translated text segments involving “shall” usage also show a tendency to adopt the passive voice for source segments articulated in active voice. This strategy is mostly seen when no overt agent is stated in the source text or when translating into active voice would result in a greater change in word order of the sentence. It is also possible to translate a sentence with covert agent into active voice; the agent might be identifiable through nearby contexts and inserted into the translation.

Translation of an implicit regulative sense or “不得” sometimes requires interpretation on the translator’s part to determine whether “shall” or “may” is more appropriate. As seen from the concordance samples, translators may find the source text ambiguous as to whether the provisions intended to state an order or an entitlement, or if they express a prohibition or a denial of permission.

doi:10.6342/NTU201603145

5.1.3 Translational and Non-translational Corpora

While comparison between translational and non-translational corpora was proposed in Baker (1995) as a means to identify patterns specific to translated texts, this study focuses on what the differences indicate about non-translational legal texts for purposes of addressing the third research question:

(3) How can translational and non-translational corpora be utilized in combination to discover additional information for aiding the process of legal translation?

Differences between the two sets of corpora can be applied to the identification of more terminology equivalents. An example of such a difference arises in referring to governmental institutions, which were commonly translated into “authority/authorities” but more frequently referred to with the terms “agency/agencies” in non-translational corpora. Collocation of “agency/agencies” in the legal English corpora would therefore likely be applicable to “authority/authorities” when translating new terms associated with governmental establishments.

Alternatively, features of the non-translational corpora may serve as indicators for identification of additional terminology. The English corpora keyword “secretary” (proper noun), for example, leads to identification of words used in similar contexts in the translational corpora such as “ministry,” “department,” “council,” “bureau,” and “minister,” which serve as useful identifiers for additional terminology equivalents.

With regard to writing style, Subsection 4.4.2 made use of the advice in Bowker and Pearson (2002) that monolingual corpora are useful for verifying whether phrase-like units are appropriate in terms of idiomatic usage. N-gram search in the non-translational corpora provided preliminary estimations on frequencies of partial translation units previously identified. From their occurrence numbers, it was deduced that phrase-like units such as “in accordance with” and “pursuant to” were likely more appropriate than “stated in” when preceding the word “article.” Previously unidentified collocation applicable to this context may also surface, such as the units “as defined in” and “as provided in” discovered in this study.

A difference in employing active or passive voice with the modals “shall” and “may” was observed and confirmed by concordance search. It was found that passive voice seemed to co-occur with a large percentage of “shall” usages in the translational corpora, considerably larger than the percentage in non-translational texts. It might therefore be advisable that translators reduce the conversion to passive voice when translating with “shall,” particularly if an overt agent is specified in the corresponding source text. The same strategies may also be applied to translating with “may,” though the tendency does not seem significant for conversion to or overusing passive voice with “may.”

Finally, non-translational corpora provide ample usage examples for the legal translator to learn from and imitate, which would provide supplementary information to the applicable contexts identified through parallel corpora. From concordance samples of “shall,” for example, frequently occurring usage patterns were identified that would likely to have been easily missed or dismissed if one were to work only with translational corpora.

5.1.4 Implications

In the process of examining keywords, terminology equivalents, stylistic features, and translation strategies, usage patterns and information were identified that will likely prove useful to the legal translator. While only a handful of examples were given in this study to illustrate how this legal corpus could be utilized to suit the needs that may arise during translation, it has been demonstrated how the individual and combined uses of parallel and monolingual corpora can allow the translator a variety of options for

doi:10.6342/NTU201603145

obtaining the information required for decision making. With the aid of the specialized corpus, legal translators will be better equipped to make informed, appropriate decisions in their translation process.

Meanwhile, it is also anticipated that the parallel and monolingual corpora compiled for this study, as well as the phrasal alignment results automatically extracted, will remain useful as reference material to future translation projects in the legislative genre. Results generated in the forms of key terms, (verified) terminology equivalents, frequently used word strings, other patterns of usage, and observed translation strategies, while perhaps small in quantity at a time, can also accumulate to become more complete, structured translation resource over time as the corpora continue to be utilized.

Finally, it is hoped as well that the approaches of corpus compilation and analysis employed in this study can, perhaps in part, be applied to other specialized fields of translation to benefit future work in this discipline.

5.2 Limitations and Suggestions for Future Research

Due in part to the aim that this study set out to achieve, availability was prioritized in a number of decisions regarding the selection of both corpora and tools. The attempt to increase efficiency and facilitate automated procedures, while helping to keep the compilation manageable to an individual, also led to the selection of legislation as object of study, instead of other less represented genres of legal language. Selection criteria were rudimentary as to what legislation to include, as were categories in the corpus structure, which makes it less feasible to work with subsets of the corpora.

The reference corpus selected for keyword analysis was the Brown corpus, which, while representative of general purpose English in the United States and effective for the needs of this study, was by standards scholars generally recommend (e.g. Bowker & Pearson, 2002) too small in size compared to the legal corpora used in this study.

doi:10.6342/NTU201603145

Computerized tools employed in this study were all freeware accessible to the public, and their available functions served as determining factors as to the types of analyses to be conducted. This study, therefore, focused on phraseologies or patterns associated with selected lexical items, and did not investigate key keywords, syntax, sentence length, or employ empirical approaches beyond lexical frequency, n-gram frequency, and concordances. Frequent navigation between different tools and subsets of corpora had been required in the analysis process, partially because each tool has its specific functions with requirements on the input corpora to match. Translators would benefit a great deal if the appropriate tools or methods could be found to further integrate the tool functions and corpora that can aid the translation process as a reference tool.

For future research, it is suggested that user-compiled corpora and methods similar to those adopted in this study be explored with smaller sets of corpora, preferably with more defined sampling or selection criteria, in part to avoid technological issues of software limitation, but also to further facilitate compilation and analysis. Ideas could perhaps be drawn from designs of open-ended corpora, *ad hoc* corpora, and flexible use of user-defined sub-corpora: small quantities of corpora compiled over time could accumulate into a more sizeable corpus; with the appropriate categorization, such as in the way TM or glossaries would be compiled and categorized in translation practice, subsets of corpora can be selected for reuse according to the translation task at hand.

Also, as this study attempted only a preliminary investigation into several methods of corpus-based analysis, there were a number of features observed, as well as other potential uses for the same corpora, that have not been fully explored. In terms of linguistic features, for example, significant differences also exist among usage of

doi:10.6342/NTU201603145

modals other than “shall” and “may” in the legal translational corpora, legal English corpora, and reference corpus. Distributions of pronouns among keywords of the translational and English corpora also show a statistical difference that may lead to interesting discoveries upon closer examination. In terms of potential uses for the corpora, the Chinese corpora has not yet been studied in much detail, nor have several aspects of comparable corpora, both in sense of original Chinese corpora paired with non-translational English corpora, and for studying the inherent features of translational language, by the definition of comparable corpora in Baker (1995). All the above are potential aspects that may be worth exploring in future research.

doi:10.6342/NTU201603145

References

Anthony, L. (2014a). AntConc (3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Anthony, L. (2014b). *AntConc help file*. Version 001. Laurence Anthony. Retrieved from http://www.laurenceanthony.net/software/antconc/releases/AntConc344/help.pdf

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target, 7*(2), 223-243.

Barnett, B. (2015). Sed - An introduction and tutorial by Bruce Barnett. *The Grymoire - home for UNIX wizards*. Retrieved from http://www.grymoire.com/Unix/sed.html

Bernardini, S., Stewart, D., & Zanettin, F. (2003). Corpora in translator education: An introduction. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 1-14). UK: St. Jerome Publishing.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (2004). Conversation text types: A multi-dimensional analysis. In G. Purnelle, C. Fairon, & A. Dister (Eds.), Proceedings from JADT 2004: the 7th International Conference on Textual Data Statistical Analysis (pp. 926-936). Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181-189). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Biel, Ł. (2010). Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In C. Heine & J. Engberg (Eds.), *Reconceptualizing LSP*, Online proceedings of the XVII European LSP

doi:10.6342/NTU201603145

Symposium 2009.

Bondi, M., & Scott, M. (Eds.). (2010). *Keyness in texts*. Philadelphia: J. Benjamins.

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. London; New York: Routledge.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roossin, P. (1990). A statistical approach to language translation. *Proceedings of the 12th Conference on Computational Linguistics,* 71-76. Stroudsburg: Association for Computational Linguistics.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1993). Aligning sentences in parallel corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 169-176. Berkeley.

Cao, Deborah. (2007). *Translating law*. Clevedon: Multilingual Matters.

Chen, P. (2012). *Using comparable specialized corpora with machine translation for extracting n-gram translation equivalents: A case study of Chinese and English contracts* (Doctoral dissertation). Available from National Digital Library of Theses and Dissertations in Taiwan.

Cheng, N. (2013). CUC_ParaConc (0.3) [Computer Software]. Beijing, China: Communication University of China. Available from http://ling.cuc.edu.cn/chs/News_View.asp?NewsID=244

Cheng, N., & Hou, M. (2012). Parallel corpus retrieval technology research. *Computer Engineering and Applications, 48*(31), 134-139.

Cheng, Y.-C. (2013). *A Corpus-based analysis of character usage in Chinese transliteration: A case study of newspapers in Taiwan and Mainland China*

doi:10.6342/NTU201603145

(Master’s thesis). Available from National Digital Library of Theses and Dissertations in Taiwan.

Church, K. W., & Gale, W. A. (1991). Concordances for parallel text. *Using corpora: Proceedings of the Eighth Annual Conference of the UW Centre for the New OED and Text Research*, 40-62. Oxford.

Coulthard, M. & Johnson, A. (2007). An introduction to forensic linguistics: Language in evidence. London; New York: Routledge.

Dagan, I., Church, K. W., & Gale, W. A. (1993). Robust bilingual word alignment for machine aided translation. *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1-8. Ohio.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61-74.

European Association for Machine Translation. (n.d.) What is machine translation?. Retrieved from http://www.eamt.org/mt.php

Farkas, A. (2015). LF Aligner (4.1) [Computer Software]. Available from http://sourceforge.net/projects/aligner/

Flowerdew, L. (2012). *Corpora and language education*. New York: Palgrave Macmillan.

Francis, W. N., & Kucera H. (1964). *Brown corpus*. Retrieved from http://www.nltk.org/nltk_data/packages/corpora/brown.zip

Frankenberg-Garcia, A., & Santos, D. (2003). Introducing Compara: The Portuguese-English parallel corpus. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 71-89). UK: St. Jerome Publishing.

Gale, W. A., & Church, K. W. (1991a). Identifying word correspondences in parallel texts. *Proceedings of the DARPA Workshop on Speech and Natural Language*,

doi:10.6342/NTU201603145

152-157. Stroudsburg: Association for Computational Linguistics.

Gale, W. A., & Church, K. W. (1991b). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, 177-184. Berkeley.

Garner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35*, 1-24.

Gozdz-Roszkowski, S. (2011). *Patterns of linguistic variation in American legal English: A corpus-based study*. Frankfurt am Main; New York: Peter Lang AG.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge, U.K.; New York: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Philadelphia, PA: John Benjamins Publishing.

Ji, M. (2012). Hypothesis testing in corpus-based literary translation studies. In M. P. Oakes & M. Ji (Eds.), *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research* (pp. 53-72). Philadelphia, PA: John Benjamins.

Kay, M., & Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics, 19*(1), 121-142.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta: Translators' Journal, 43*(4), 557-570.

Lee, D. Y. W. (2010). What corpora are available?. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 107-121). London; New York: Routledge.

Liu, T.-J. (2014). *PTT Corpus: Construction and applications* (Master's thesis). Available from National Digital Library of Theses and Dissertations in Taiwan.

doi:10.6342/NTU201603145

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Netherlands: Springer.

Maia, B. (2003). Training translators in terminology and information retrieval using comparable and parallel corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 43-54). UK: St. Jerome Publishing.

Marcus, P., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. Technical Report MSCIS-93-87, Department of Computer and Information Science, University of Pennsylvania.

McEnery, T. (2003). Corpus linguistics. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 448-463). New York: Oxford University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Mikheev, A. (2003). Text segmentation. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 201-218). New York: Oxford University Press.

Ministry of Justice. (2015). Laws & Regulations Database of the Republic of China. Available from http://law.moj.gov.tw/

Mitkov, R. (Ed.) (2003). *The Oxford handbook of computational linguistics*. New York: Oxford University Press.

Neubig, G. (2012). pialign (0.2.4) [Computer Software]. Retrieved from http://phontron.com/pialign/download/pialign-0.2.4.tar.gz

Neubig, G., Watanabe, T., Sumita, E., Mori, S., & Kawahara, T. (2011). An unsupervised model for joint phrase alignment and extraction. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 632-641. Portland.

Office of the Law Revision Counsel. (2015). *United States code*. Available from

doi:10.6342/NTU201603145

http://uscode.house.gov/browse.xhtml

Pearson, J. (2003). Using parallel texts in the translator training environment. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 15-24). UK: St. Jerome Publishing.

Quah, C. K. (2006). *Translation and technology*. New York: Palgrave Macmillan.

Rayson P., Berridge D., & Francis B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Fairon, & A. Dister (Eds.), Proceedings from JADT 2004: *The 7th International Conference on Statistical Analysis of Textual Data* (pp. 15-34). Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, 9*, 1-6. Stroudsburg: Association for Computational Linguistics.

Red Hat, Inc. (2015). *Cygwin user's guide*. Retrieved from https://cygwin.com/cygwin-ug-net/cygwin-ug-net.html

Reppen, R. (2010). Building a corpus: What are the key considerations?. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 31-37). London; New York: Routledge.

Rossini Favretti, R., Tamburini, F., & Martelli1, E. (2007).Words from BOnonia Legal Corpus. In W. Teubert (Ed.), *Text corpora and multilingual lexicography* (pp. 11-30). Amsterdam; Philadelphia: John Benjamins Publishing Company.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Šarčević, S. (1997). *New approach to legal translation*. The Hague: Kluwer Law

doi:10.6342/NTU201603145

International.

Scott, M. (1997). PC analysis of key words – And key key words. *System*, *25*(2), 233-245.

Scott, M. (2000). *WordSmith tools help manual*. Version 3.0. Mike Scott and Oxford University Press.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Philadelphia: J. Benjamins.

Shuttleworth, M., & Lagoudaki, E. (2006). Translation memory systems: technology in the service of the translation professional. Paper presented at 1st Athens International Conference of Translation and Interpretation, Athens, Greece. Retrieved from http://project2007.hau.gr/telamon/files/MarkShuttleworth_ElinaLagoudaki_PaperAICTI.pdf

Sinclair, J. (2003). *Reading concordances: An introduction*. London; New York: Pearson/Longman.

Somers, H. (2003). Machine translation: Latest developments. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 512-528). New York: Oxford University Press.

Stenberg, D. (2015). cURL (7.41.0) [Computer Software]. Available from http://curl.haxx.se/

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford; Malden, MA: Blackwell Publishers.

Stubbs, M. (2007). An example of frequent English phraseology: Distributions, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89-105). Amsterdam; New York: Rodopi.

Toyama, K. (2011). *Brief introduction to Bilingual KWIC for Taiwan Laws* [PowerPoint

doi:10.6342/NTU201603145

slides]. Retrieved from http://www.slidefinder.net/t/taiwanlii-workshop-toyama english20110607/32657725

Toutanova, K., Klein D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of HLT-NAACL 2003*, 252-259.

Varantola, K. (2002). Disposable corpora as intelligent tools in translation. *Cadernos de Tradução, 1*(9), 171-189.

Varantola, K. (2003). Translators and disposable corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 55-70). UK: St. Jerome Publishing.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., & Nagy, V. (2005). Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, 590-596.

Véronis, J. (2000). From the Rosetta Stone to the information society: A survey of parallel text processing. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora* (pp. 1-24). Boston: Kluwer Academic Publishers.

Voutilainen, A. (2003). Part-of-speech tagging. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 219-232). New York: Oxford University Press.

Wu, D. (1995a). Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. Technical Report HKUST-CS95-30, Department of Computer Science, University of Science and Technology.

Wu, D. (1995b). Grammarless extraction of phrasal translation examples from parallel texts. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, 2*, 354-372. Leuven, Belgium.

Appendix

Table A.1

The Penn Treebank POS Tagset and Tag Descriptions

No.	Tag	Description	No.	Tag	Description
1.	CC	Coordinating conjunction	19.	PRP$	Possessive pronoun
2.	CD	Cardinal number	20.	RB	Adverb
3.	DT	Determiner	21.	RBR	Adverb, comparative
4.	EX	Existential *there*	22.	RBS	Adverb, superlative
5.	FW	Foreign word	23.	RP	Particle
6.	IN	Preposition/subordinating conjunction	24.	SYM	Symbol (mathematical or scientific)
7.	JJ	Adjective	25.	TO	*to*
8.	JJR	Adjective, comparative	26.	UH	Interjection
9.	JJS	Adjective, superlative	27.	VB	Verb, base form
10.	LS	List item marker	28.	VBD	Verb, past tense
11.	MD	Modal	29.	VBG	Verb, gerund/present participle
12.	NN	Noun, singular or mass	30.	VBN	Verb, past participle
13.	NNS	Noun, plural	31.	VBP	Verb, non-3rd person singular present
14.	NNP	Proper noun, singular	32.	VBZ	Verb, 3rd person singular present
15.	NNPS	Proper noun, plural	33.	WDT	*wh*-determiner
16.	PDT	Predeterminer	34.	WP	*wh*-pronoun
17.	POS	Possessive ending	35.	WP$	Possessive *wh*-pronoun
18.	PRP	Personal pronoun	36.	WRB	*wh*-adverb

Note. Adapted from "Building a Large Annotated Corpus of English: The Penn Treebank," by P. Marcus, B. Santorini, and M. Marcinkiewicz, 1993, Technical Report MSCIS-93-87, p. 5. Copyright 1993 by the Department of Computer and Information Science, University of Pennsylvania.

doi:10.6342/NTU201603145