

## **Text Analysis of Patent Abstracts** **Yvonne Tsai, National Taiwan University**

### **ABSTRACT**

Text analysis involves the deconstruction of information within a text. This includes text structure, text pattern, linguistic features, lexical analysis, and syntactic analysis. This research took as its starting point the bottom-up approach of analysing the lexical features, syntactic features, and textual features of patent abstracts for comprehensive coverage of text analysis. Several tools have been applied in the analysis of patent abstracts. This three-fold analysis of text outlined above embraces information on sentence statistics, segmentation statistics, word frequencies, lexical densities, and readability levels. It was found that English translated texts presented a more consistent use of short sentences than in the original Chinese texts, and a common usage of shorter words was also evident in the translated texts. While short sentences, short word length, and high repetitions of words characterised texts with easy readability, findings from the readability tests indicated that in order to understand patent abstracts without difficulty, readers should have received at least 14 years of education.

### **KEYWORDS**

Patent abstract, text analysis, readability, syntactic analysis, lexical analysis, lexical density, segmentation.

### **Introduction**

There are many ways to analyse texts: content analysis, textual analysis, and text analytics. These inter-related terms and concepts involve systematic approaches in deconstructing information within a text. The information for analysis usually includes text structure, text pattern, linguistic features, lexical analysis, and syntactic analysis. The structure of a text relates to text pattern and linguistic features, while linguistic features provide lexical and syntactic properties. This research began from a bottom-up approach by analysing the lexical features, syntactic features, and textual features of patent abstracts for comprehensive coverage of text analysis. This three-fold analysis of text further embraced information on sentence statistics, segmentation statistics, word frequencies, lexical densities, and readability levels.

As Olohan points out, “data-based or data-driven analysis of the text [...] is clearly aligned with the descriptive perspective” (2004: 8-9), thus, the integration of quantitative and qualitative analyses from this study helped to demonstrate the overall text typology of patent abstracts. In the study of translation, comparisons between the Chinese source text and the English translation are essential. The main purpose of comparing the two languages is to unveil the “correlations between the two sides of the relation” (Williams and Chesterman 2002: 51) and therefore, analyses of the text in both languages carry equal weight.

Several tools were applied in the analysis of patent abstracts. To start with,

part-of-speech tagging systems for different languages were adopted for the parsing of the Chinese and English texts. Parsing results provided part-of-speech markers and punctuation markers, which served as the basis for the study of lexical property, textual structure, and syntactic organisation. In order to investigate and compare the two languages, comparable corpora were compiled using the concordance tool AntConc 3.2.2w (Anthony 2008). Word statistics and lexical densities were generated from the concordance tool and textual analysis freeware, such as Topicalizer (Wilmsmann 2008) and Vocabulary Management Profiles (Youmans and Pauley 2008). Readability levels including *Flesch Reading Ease*, *Flesh-Kincaid Grade level*, *Gunning-Fog Index*, and *Automated Readability Index* were also calculated via an online interactive website (Editcentral.com 2008).

The first step in the text analysis process was to select Chinese texts as samples for this study. In order for the text samples to be representative, it was vital to set criteria for the selection.

## **1. Criteria for the selection of a 50-text corpus for text analysis**

Criteria for the selection of Chinese text samples included the subject field with the greatest demand for translation, the availability of patent abstracts written directly in Chinese as the source language, and practical concerns in relation to the feasibility of this research. Accordingly, the sampling procedure for Chinese texts adopted the following methods: the use of statistical reports to determine the category where most patents were granted, a search of online databases with key words for patent abstracts in a specific category, and limitations on the quantity of available records for representative data with significant value.

### **1.1. Selecting a subject field**

Inventions which require protection by patents could range from household furniture to weaponry. Since the attributes of inventions can be boundless, with large quantities of information contained in the patent documents, archiving patents in accordance with their subject field enhances accessibility, not only for the patent offices and patent applicants, but also for the general public. By the same token, the selection of a 50-text corpus for this research was developed from a general subject field to a more specific group with reference to an internationally recognised patent archiving system, the International Patent Classification (IPC) (WIPO 2006b).

#### **1.1.1. Categorisation of patents with International Patent Classification**

The World Intellectual Property Organization regulates the classification of patents with IPC in order to organise patents in such a way as to facilitate

users in locating the information they require quickly and precisely. According to IPC, patents are classified into eight broad categories, under which classes, subclasses, groups and subgroups are organised respectively in hierarchical order. These account for a vast quantity of data.

According to WIPO, the IPC that “speak(s) the lingua franca of the patent classification” (WIPO 2006a) is currently used in more than 100 countries in the world, including Taiwan. Consequently, the IPC was used as the basis of reference to other statistical reports throughout the selection process of the text corpus.

### **1.1.2. Statistics in relation to the filing of patents throughout the world**

The Patent Cooperation Treaty (PCT) Statistical Indicators Report (WIPO 2007) conducted by WIPO shows the number of PCT International Applications published in accordance with their technical field. The top 3 subject areas for the filing of patents during the period from January 2003 to October 2007 were: A61K (Preparations for Medical, Dental or Toilet Purposes) accounting for 9.2% of the total number of patent applications with 11,192 applications; G06F (Electric Digital Data Processing), which accounted for 6.4%, a total of 7,761 applications; and H04L (Transmission of Digital Information, e.g. Telegraphic Communication), 4.7%.

Among the eight broad categories, the first category of ‘Human Necessities’ (Section A) is seemingly the field which has the closest connection to the life experiences of individuals. However, the class ‘Preparations for Medical, Dental or Toilet Purposes,’ demands a high level of specialist domain knowledge in the field of medicine. The Second category, on the other hand, labelled ‘Electric Digital Data Processing,’ may be a more familiar subject field for both readers and translators, since advances in modern technology have led to a greater use of IT tools for the convenience these tools provide in our lives.

### **1.1.3. Statistics in relation to the filing of patents in Taiwan**

Article 2 of Taiwan Patent Act (TIPO 2004) classifies patents into three categories: invention patents; utility model patents; and design patents. The category of utility model refers to the invention of technical concepts applied to the form, construction or installation of an object, and only the title of the patent, rather than the abstract, requires translation. Hence this category is not considered in this research. The category of design patents relates to the invention of the visual representation of the shapes, patterns and colours of the invention, and is also excluded since abstracts are not readily available online for technical processing. The category of invention patents, defined as “any creation of technical concepts by utilizing the rules of nature” (TIPO 2004), contains patent abstracts in both the Chinese and the English language, and is therefore used for this

research.

At both national and international level, the 2006 patent application statistics of the Taiwan Intellectual Property Organization (TIPO 2007) indicated the categories where most patents were granted in 2006 in Taiwan. The statistical figures were generated according to the three types of patents mentioned above. As indicated by the statistics, a total of 49,315 patents were granted out of 80,988 applications in 2006. Amongst all the technical fields, the IPC codes for the 3 most prevalent areas in which patents were filed in Taiwan were H01 (Basic Electric Elements), H04 (Electric Communication Technique), and G06 (Computing; Calculating; Counting).

When compared to the PCT Statistical Indicators Report (WIPO 2007), there were evident overlaps in G06 and H04. As a result of the two sets of statistics, both national and international, G06Fi<sup>a</sup>Electric Digital Data Processing<sup>a</sup> was selected for this research.

#### **1.1.4. Supporting reasons for the selection of 'keyboard' for keyword search**

The predetermined subclass G06F remains too broad a selection for text sampling. The main reason for this is that under the hierarchical classification of the subclass tier, there are group and subgroup categories. In G06F, the groups vary from input arrangements for data transfer, to data conversion, data processing, digital computers, and security arrangements for protecting computers. Considerations in relation to the consistency of data and the familiarity level of the technical field concerned limited the search to a more specific range. In addition, three reports were referenced to support the selection of the 50-text corpus. These are the Global Information Technology Report (INSEAD 2008), the E-Communications Household Survey (TNS 2008), and the Digital Divide in Taiwan (RDEC 2007).

The Global Information Technology Report (INSEAD 2008) has been produced by the World Economic Forum on a yearly basis since 2001. The report covers economic activities in around 127 countries and assesses the influence of information and communication technologies on the various national economies. The most significant part of this report is the comprehensive networked readiness ranking, which is "a relative indicator of a nation's ICT excellence" (INSEAD 2004: 5). The index shows the global diffusion of ICT, or "the degree of preparation of a nation or community to participate in and benefit from ICT developments" (INSEAD 2004: 4). According to the latest Networked Readiness Index 2007-2008 (WEF 2008) rankings, the most networked ready country in the world is Denmark, the United Kingdom is the 12th, Taiwan the 17th, and China the 57th.

The E-Communications Household Survey (TNS 2008) is a special

Eurobarometer report<sup>i</sup> conducted by TNS Opinion and Social Institutes for the European Commission. The latest survey in 2008 pointed out that the majority of European households have a computer, and virtually half of the households in Europe have internet access. In countries with higher computer penetration, such as the Netherlands, 90% of households have a computer.

When expanding the horizon to a global view, the Global Information Technology Report (INSEAD 2008) shows Taiwan to be 7th in the world in terms of the number of internet users. As for the percentage of households with personal computer equipment, Taiwan ranked 18th. The Digital Divide in Taiwan (RDEC 2007) report conducted by the Research, Development and Evaluation Commission in 2007 revealed an average of 65.6% internet users and 71% computer penetration rate in the population aged over 12 in Taiwan. The figure equates to 14.07 million people who have access to computers. The internet dependency rate in Taiwan is also on the rise, from an average of 2.4 hours of surfing a day to 2.7 hours per day.

Since computers play a significant role in the lives of the majority of the world's population, it can be presumed that computer hardware is reasonably familiar to the users, especially the input device—the keyboard. In addition to the concern with technical field familiarity, any user could easily picture the shape of a keyboard, or understand the function of the keyboard, since the first step in learning to use a computer is knowing how to give instructions via the keyboard. This is the reason why patent abstracts related to keyboards under IPC category G06F are used for this research.

## **1.2. Searching for available patent abstracts in the Chinese language**

The online database of TIPO—Taiwan Patent Search (TIPO 2008)—contains a very large quantity of patent documents in both Chinese and English. This database was the only source of text sampling used in this study. Although it is reasonable to assume that the Chinese text is the source text while English is the translated version, it was found that some Chinese texts had been translated into Chinese from foreign languages rather than being originally written in Chinese. The statistics in relation to applicant nationalities in 2007 show that 15.35% of patent applications in Taiwan were filed by Japanese applicants. In cases where the patent was filed by a foreign company outside Taiwan, the Chinese version of the patent documents could be texts which have been translated into the Chinese language.

The language that the patent abstract is first written in is important for patent abstract translation analysis. Since it is widely believed that it is

not possible to accurately reproduce 100% of the source text in the translated text, partial distortion of the sentence structure, semantic unit, or syntax is very common. Translating translated text into a third language can be a difficult endeavour, and can be heavily dependent on the quality of the translated text, as comprehensibility always comes before translatability when the two are in issue. With this in mind, patent abstracts that had been translated into Chinese from other languages were excluded from the research.

### **1.3. Limiting the search to a specific time period**

The final criterion was limiting the search to a specific time period. The Taiwan Patent Search (TIPO 2008) website is updated three times a month. To date, patent bibliographic data on the English website covered a period from 01/01/1993 to 11/05/2008. Chinese patent abstracts could be accessed from as early as 1990 to the present. Since texts in both the Chinese and the English languages were essential in this study, historical data were not taken into consideration. Instead, a more recent timeframe was used for the search in order to comply with the constantly evolving properties of technological advancement.

The data for this research consisted of patent abstracts filed between 01/01/2006 and 01/08/2007. The search was limited to the IPC field of G06F in relation to 'keyboard', and was restricted to Taiwanese applicants or inventors. A total of 50 patent abstracts in both the Chinese and the English language were collected for initial text analysis.

## **2. Syntactic analysis**

Syntax is believed to be of "decisive importance for the choice of processing (translation) strategy" in languages with diversified structure such as Chinese and English (Kirchhoff 2002: 113). Syntactic analysis can be defined as the parsing of parts of speech in order to examine the grammatical structure of sentences. Parsed texts have semantic values and many different forms. Segmenting texts in accordance with their constituent parts of speech is a common form of annotation, and is often seen as the first step in a more comprehensive syntactic annotation. For a small number of texts, parsing can be completed manually. For larger quantities of texts, manual parsing is not only time-consuming but also labour-intensive.

Several online resources are available to handle texts of considerable quantities. This is why CLAWS (UCREL 1993), or the Constituent Likelihood Automatic Word-tagging System, is often employed in research studies, including this one. This part of speech tagging software was developed by the University Centre for Computer Corpus Research on Language at Lancaster in the early 1980s, and has been applied in various

research projects over the years, notably in tagging the 100-million-word collection of written and spoken language samples of the British National Corpus.

CLAWS assigns a part of speech tag to each word or word combination in a text and a phrase marker to each sentence of the corpus, and has an accuracy rate of 96% to 97% in judging text types and categories. In order to successfully determine the possible parts of speech of words, CLAWS has a constantly updating database that analyses lexicon and an idiom list of multi-word combinations. From 132 basic tags to the latest count of over 160 tags, the tagset has been revised and enriched several times over the years. The current standard tagset is the C7 tagset (UCREL), and since the C7 tagset contains larger tags, it was applied in this research.

The Chinese texts were processed through the Chinese Word Segmentation System with Unknown Word Identification (CKIP 2004) developed by the Language and Knowledge Processing Group of the Academia Sinica. The basis of the word segmentation process depends on the lexicons, morphological rules for quantifiers and repetitions of words. This application processes Chinese texts by matching word combinations of a sentence with a lexicon, and since there are no delimiters to space individual Chinese characters in the Chinese language structure, it is not easy to accurately locate the correct tag without any ambiguities. While many other programs try to solve segmentation ambiguities, this system identifies unknown words in a text, which, according to Academia Sinica, account for 3% to 5% of a given text.

The Chinese Word Segmentation System with Unknown Word Identification was the first word segmentation system equipped with the identification of new words and jargons and the prediction of syntactic category. There is a total of 100,000 entries of tagged lexicon and 46 tags in the tagset. This tagset contains 43 part-of-speech tags, 3 Chinese language specific features such as 的 (de), 是 (shi), and foreign words. The current version was reduced from the original 178 syntactic categories of Chinese Knowledge and Information Processing lexicon in 1993.

## **2.1. Parsing translated texts with CLAWS**

A free CLAWS trial service is offered on the CLAWS website (UCREL 1993), with the choice of tagging with the C5 or C7 tagset. The texts were respectively inputted in the space provided on the website, which automatically includes POS tags next to individual words in the output. Tags after an underscore following a word can be consulted for their representation on the tagset.

Among a cumulative sum of 6,138 words in 50 patent abstracts, there were 64 tags generated from the CLAWS POS tagging system. In the C7 tagset, nouns were subdivided into 22 different types, though only three were tagged in the given text. Other major tags being further classified into more specific tags included 31 verb tags, 19 types of pronoun, 14 adverb groups, 7 conjunctions, and 4 prepositions. In the tagged text, nevertheless, there were only 16 verb tags, 9 adverbs, 6 pronouns, 5 conjunctions and determiners, 4 prepositions, 3 nouns and adjectives, and others. The five taggers with highest occurrences were singular common nouns (NN1), with 1,800 hits; articles (AT), with 1,073 hits; general adjectives (JJ), with 578 hits; general prepositions (II) with 366 hits; and coordinating conjunctions (CC) with 244 hits.

## **2.2. Parsing source texts with the Chinese Word Segmentation System with Unknown Word Identification**

Chinese word segmentation is a pre-requisite in the analysis of Chinese texts. The reason lies in the fact that unlike English, where every word stands alone with a specified meaning in a sentence, in Chinese, phrases are the smallest meaningful unit and the fundamental unit of the Chinese sentence. For example, the phrase 元素(element) is composed of two Chinese characters: the first character 元 means primary, basic, unit, or dollar, and the second character 素 represents plain, or uncoloured (Inventec 2008). It is only when the two characters are combined together into one unit that the meaning of 'element' can be derived, or these two words would be of no significance in the sentence if viewed separately. Segmenting words into meaningful phrases, and part of speech tagging of the phrases, can facilitate research.

The online demo system is copyrighted in the National Digital Archives Program (CKIP 2004). The text was inputted in the blank space, and four results were generated following its submission. The results included the text file of the input text, the process of unknown word identification, tagging results inclusive of unknown words, and the unknown word list. Below is a sample of tagging results inclusive of unknown words. Tags are displayed in brackets next to the parsed phrase.

The segmentation results collected from the system showed 5,689 parsed Chinese phrases segmented from a cumulative sum of 10,200 Chinese characters among 50 patent abstracts. 42 tags were applied. Like the tagged results in the English translated texts, the part of speech with the highest occurrence was the common noun (Na), with 1,818 hits. As there were no quantifiers for nouns in Chinese, only one tag was used for common nouns. The second most frequent tag from the top was the transitive verb (VC), with 662 hits, followed by the numeral (Neu). However, the numeral '一' (yi, one) alone accounted for more than 80% of all numeral tags, 14% of which represented the ordinal number 'first'

while the remaining were cardinal numbers.

A comparison of the POS tag results is presented below in Table 1.

Chinese				English		
Common noun (例如：鍵盤)	Na	1818	1	1800	NN1	singular common noun (e.g. keyboard, invention)
base form of lexical verb (例如：摺疊、輸出)	VC	662	2	1073	AT	article (e.g. the, a)
Numerals (例如：一、第一、三十七)	Neu	361	3	578	JJ	general adjective (e.g. foldable, portable, local)
Function words (例如：的、之)	DE	308	4	366	II	general preposition (e.g. to, on, at, up)
Preposition (例如：在、與、以、於)	P	300	5	244	CC	coordinating conjunction (e.g. and, or)

**Table 1. Comparisons of POS tag in Chinese and English texts**

### 2.3. Sentence statistics

Sentence length has been applied in various analyses of texts. It is one of the measures Irizarry (1990) used for her stylistic analysis of Spanish narratives. In a study of the writing styles of Hemingway, Galsworthy and Faulkner, Whissel (1994) adopted sentence length as one of her 50 evaluating criteria. Malmkjaer also applied sentence length in her study of stylistics, describing it as the “consistent occurrence in the text of certain items and structures, or types of items and structures, among those offered by the language as a whole” (Malmkjaer 2002: 510). Last but not least, the study of Devlin and Tait (1998) shows the effect of sentence length on readability scores, where shorter sentences with more common words tend to receive higher readability scores.

English sentence length in the 50 patent abstracts in question varied from a minimum of 13.14 words per sentence to 125 words per sentence. The range of 111.86 and the standard deviation of 24.12 indicated inconsistencies in the length of sentences. However, with an average of 35.71 English words per sentence, only 8 abstracts had an average of more than 50 words per sentence, two of which were over 100 words. This was supported by the word counts in 50 English texts. Though the word count of English translations deviated from 55 to 281 words, only 6 abstracts contained more than 160 words, three of which were over 200.

With regard to the Chinese sentence length, segmented Chinese phrases

were used in the calculation instead of unitary Chinese characters. Take the single character 鍵 (key) for example. Among a total of 325 instances of 鍵 (key) in the 50 texts, there were 189 instances of 鍵盤 (keyboard), 86 instances of 按鍵 (press key), and as many combinations as there could be with the word 鍵(key). Some of the examples included 鍵碼 (key code), 按鍵組 (keyboard assembly), 數字鍵 (numeric key), 鍵盤組 (keyboard), 鍵帽 (key cap), 鍵盤區 (keyboard regions), 音樂鍵 (musical key), 鍵入 (input), 編輯鍵 (editing key), and 字鍵 (key). Without segmenting the characters, a lot of effort would be required to differentiate meanings in distinct combinations. As is indicated in Olohan, "an annotated corpus is likely to lend itself to more automatic analysis than an untagged corpus" (2004: 63), segmented Chinese phrases with better matched characters and more appropriate meaning can facilitate the analytical process.

The word count of Chinese segmented words ranged from 35 to 246 words. The wide range of 211 words resulted in a big standard deviation of 43.78. With an average word count of 113.8, only four abstracts contained fewer than 50 words, and two abstracts exceeded 200 words. The majority lay between 100 and 200 words. In view of Chinese sentence length, the range and standard deviation showed similar results with segmented word count. The average words per sentence were 62.37, which varied from 11.5 to 229 words per sentence. Similarly, while 46% of 50 abstracts contained a shorter sentence length of below 50 words, 14% were over 100 words per sentence and only one abstract contained more than 200 words per sentence.

It is evident in the sentence length statistics that English sentence length had a smaller standard deviation and range, in addition to a lower maximum number of words per sentence count value. Despite the fact that the number of sentences in both Chinese and English texts varied from 1 to 8, with a range of 7, the number of segmented Chinese words per sentence was 1.75 more than that of the English translation. Moreover, around 84% of English sentences consisted of fewer than 50 words, whereas only 46% of Chinese sentences length contained fewer than 50 words. In comparison, the average word per sentence count showed that the translated texts were more consistent in the use of short sentences than the Chinese source texts.

#### **2.4. Segmentation statistics**

Segmentation statistics provide information on the division of sentences or the breaking within a sentence, and can be measured by punctuation marks. A collective sum of 657 segmentations was seen in 50 Chinese patent abstracts, with an average of 13.14 segmentations used per text. English translations carried a slightly lower segmentation value of 543, and a lower average of 10.86. Punctuation marks found in translated

English patent abstracts included the comma, period, semicolon, colon, and parenthesis. This result showed a close resemblance to the punctuation marks employed in the Chinese texts, with the exception of an exceptional 頓號 (pause) in Chinese, which is often replaced by commas in English.

#### **2.4.1. Punctuation markers of commas, periods, and Chinese pauses (、)**

In the analysis of segmentation statistics, I found in my data that the most commonly seen punctuation marks were commas and periods which divided the briefest segments. According to the statistics, more commas were applied in Chinese texts while periods were used more in English texts. I also found strong correlations between commas and periods in the Chinese texts and the Chinese sentence length. The frequent use of commas and limited use of periods in Chinese texts explained a longer sentence length than that in the English texts. As indicated in the sentence statistics, English translations showed a more consistent use of short sentences, which was further supported by the correlations found between periods and sentence length in the English texts.

As with the comma, 頓號 (the Chinese pause) is used to mark off discrete elements within a sentence. There are two main functions of the Chinese pause. In particular, it can separate a list of words with the same part of speech, usually nouns or phrases, as in the following sentence taken from one patent abstract:

一種鍵盤電路, 包括一芯片、一按鍵電路及一識別碼生成電路。

A keyboard circuit comprises a chip, a key-press circuit and an ID generating circuit.

Another function is when it is used after ordinal numbers or sequential categories such as first, second, and so on, as in the following example from one of the patent abstracts:

第一壓合裝置可將該第一、二治具的第一、二框件壓合形成多數架橋單元,

The first press-bonding device can press/bond the first and second frame members on the first and second tools to form plural bridging units.

Like its name, the sesame shaped pause (、) in Chinese enables readers to pause once in a while as they read the text. It is also a way to link juxtaposed words or phrases into groups (教育部國語推行委員會 2007), for the purpose of producing succinct and concise sentences.

From my investigation, I found correlations between pauses in the Chinese texts and commas in the English texts. Of 78 pauses in the 50 Chinese abstracts, only three abstracts contained more than five pauses. The only text with more than ten pauses had a maximum number of 14

pauses. The total Chinese segmentation number of that particular text was double the sum of English punctuation marks. In the Chinese text, there were only two periods, yet five commas and 14 pauses, which was indicative of lexically dense sentences.

The number of commas and pauses in the Chinese text were not reproduced in the English translated text. Instead, there were fewer commas but more periods in the English text. The division of two long sentences into six short sentences replaced numerous commas and pauses in the Chinese text. This is one way of handling pauses found in my research, which is to divide what was linked with pauses in the Chinese text with periods in the English translation.

#### **2.4.2. Characteristic semicolons and colons in patent abstracts**

Punctuation marks that were found characteristically in patent abstracts were semicolons and colons, usually for an uninterrupted description of all the components in an object. Semicolons and colons were also found to be statistically correlated with sentence lengths in both languages. In spite of this, semicolons were used a lot more in the Chinese texts than the English translations. Among 50 texts, at least one semicolon was present in 17 Chinese abstracts, whereas 94% of the English texts contained no semicolons. In both Chinese and English texts, only one abstract had six semicolons and two colons. Colons were used a lot less frequently than semicolons. Nevertheless, colons were used in one quarter of the Chinese texts, yet only three included colons in the translation.

The parenthesis was the remaining punctuation mark employed in patent abstracts in my study. Two main uses of the parenthesis were for sequencing information and for providing foreign words or explanations. Parenthetical expressions of foreign words are commonly seen in specialised texts with unfamiliar and specialised terminologies, and patent abstracts are no exception. The provision of foreign translation facilitates the translation process and benefits the translator. However, most of the parentheses found in the study were for the sequential numbering of descriptions, especially for the inclusion of explanations for respective items within one invention (Appendix III). Parentheses were only used in seven Chinese texts and six English translations.

To summarise, Chinese texts demonstrated a more diversified use of punctuation marks. Although more segmentation was seen in the Chinese texts, English texts exceeded Chinese texts in the use of the period. This reaffirmed the findings of short sentence features in the English texts.

### **3. Lexical analysis**

Lexical analysis investigates word length, word frequencies, keyword in context, and lexical density, or type-token ratio. Word length is easy to understand from its literal meaning—the number of characters in a word. Word frequencies provide statistical evidence for stylistic features by presenting the occurrence of each word, inclusive of its respective parts of speech. The use of corpus tools enables searches for words or phrases with co-texts displayed in the output. The word being searched is also referred to as keyword, and the display of keyword with co-texts is called a concordance line. The aligned and sorted result is what constitutes “Key Word in Context.”

Lexical density refers to the amount of content words in a text or corpus. More precisely, lexical density measures the percentage of lexical words to grammatical words in a corpus. Another commonly used metric in measuring lexical variations and identifying repetitions is the type-token ratio. Type refers to the number of different words used in the corpus, in other words, “lemmatized word count” (Irizarry 1990: 268). Token is the number of words in the corpus, so for example, a text of 500 words long is said to have 500 tokens. If a corpus consisted of 500 tokens and 250 types, the ratio between types and tokens would be 50% in this example. This tells us something about the relationship between the total number of running words in a corpus and the number of different words used.

Various text analytical freeware is easily accessible online. The most notable one is AntConc 3.2.2w (2008), a popular concordancing software program developed by Anthony (2008) at Waseda University, Japan. AntConc 3.2.2w can process more than one corpus file, generate key word in context concordance lines and concordance plots, and analyse word clusters, collocates, word frequencies, and keywords. One of the unmatched advantages of this software is that while most freeware only processes European languages, AntConc 3.2.2w is able to process Asian languages such as Chinese, Japanese and Korean.

Another freeware applied in this research is Topicalizer, developed by Wilmsmann (2008). Topicalizer processes plain text and provides word, sentence and paragraph count, collocations, lexical density, keywords, readability, and so on. The benefit of Topicalizer is that, in addition to frequent words, frequent phrases can be listed and organised into frequent two-word phrases up to five-word phrases, with a separate list for the inclusion of stop words. Stop words are common words such as ‘about,’ ‘again’ or ‘become’ in computer search engines. As these common words are of lesser relevance to the search, stop words are the words that are usually filtered out in the search.

The last tool employed is the Vocabulary Management Profiles (VMP) from the University of Missouri (2008). The four main functions provided on the VMP website are vocabulary management, fractal dimension, type-token

statistics, and concordances and word frequencies. Text files can be uploaded for analysis, and graphs can be generated along with downloadable output. However, the computed output is presented in plain text form, and without columns and charts, copious figures can be less reader-friendly and rather daunting.

### **3.1. Word statistics**

The word statistics in the analysis showed a mean word length of 4.15 English characters, which typified translated patent language as a whole. The small standard deviation of 0.43 denoted the common usage of short words in the translated texts. The longest word was the word 'electroluminescence,' with 19 characters. However, only one instance was found of the word 'electroluminescence.' The shortest word was the word 'one,' with 26 instances. Among the longest words in each translated patent abstract, 8 were compounds with a noun-verb combination. Most compound words in the texts were hyphenated. An example of this is 'light-emitting.' Another feature of the formation of long words was the presence of affixes. The longest word in all the texts, 'electroluminescence,' is a good example of this feature.

The word length in the Chinese texts ranged from a minimum of one character to four characters. This was partly due to the unique word formation (see Section 6.2.3) of the Chinese language, and partly due to the intrinsic nature of using phrases as the smallest meaningful unit. Within the 50 texts, there were five words with four characters. In spite of this, many more four-character-words could be formed by combining words together into noun phrases. It can be inferred from the repetitions of frequent words found in compound words that word length statistics in the Chinese text may carry less value in terms of representativeness.

#### **3.1.1. Type-token ratio**

The ratio of type over token provides lexical variations. Lexical variations can be computed by dividing the number of different words by the number of running words in a text, and repetitive use of words would lower the ratio as types are counted for differences. As more repetitions occur in a text, the type-token ratio decreases. In longer texts, where the probability of repetitions is higher, the ratio would be lower than in shorter texts. For this reason, lexical varieties are calculated with a standardised 1,000 as the ratio for successive tokens in a text and provided in mean value. However, since the patent abstracts were structured with 350 words at most, a ratio of 100 is sufficient.

The ratio of types to tokens in the 50 English patent translations was presented in the type-token curve generated from VMP (Youmans and Pauley 2008). Youmans (1990) considers this type-token curve as a type-

token vocabulary curve that estimates the vocabulary size of a text with the support of complicated statistical calculations (Carroll 1968; Carroll, Davis and Richman 1971). The curve starts as a straight line, with types = tokens, until the first occurrence of repetition. When the number of tokens continues to exceed the number of types, the rate at which the curve rises slows down. The final number of types would then provide information on the total vocabulary used.

The Chinese type-token ratio of the 50-text corpus was 48.85. The English translated texts received a lower type-token ratio of 35.37. The low type-token ratio found in my study in the English texts displayed high frequencies and low varieties in word usage. This is also indicative of high repetitions, of which some are redundant (see Section 6.2.4).

### **3.1.2. Lexical density**

Lexical density can be used as an indicator of text type by measuring the number of content words used in a text. It is perceived that written texts tend to be more planned and more formal than spoken texts, and thus it is reasonable to assume that written texts are lexically denser than spoken texts. Stubbs (1996: 73) considers written texts in general have more than 40% lexical density. Variations in text-typological differences in the lexical densities of written texts have also been observed by Stubbs (*ibid.* 73-4), where non-fiction texts have higher lexical densities than fiction texts. This figure is not too far from the measures provided by UsingEnglish.com (King & Flynn 2008), where texts with more information loads have a higher lexical density of around 60-70%, and low lexically dense texts have around 40-50% lexical density.

Technical texts are not only non-fiction texts, but also non-fiction texts with a heavy information load. For this reason, the lexical densities of technical texts may be considerably higher, depending on how lexical items are distributed in the grammatical structure. However, in my findings, the average lexical density of the 50 English patent abstract translations was 36%. This figure was not only lower than the low lexically dense text of 40-50% suggested by UsingEnglish.com (King & Flynn 2008), it was also below the 40% density threshold of Stubbs (1996) for written texts. The result can be supported by the study of Laviosa (2002: 60-62) on writing styles, where translated texts exhibit lower lexical densities than the source texts. The low density feature in translation is also indicative of the simplification feature in translation (see Section 2.3.2.5).

### **3.1.3. Frequency list**

A frequency list is a list of all the words that appear in a corpus, with the number of times each word occurs in the text. The list can be used to

distinguish common expressions or detect rare usage (Kenny 2001). Frequent words in the Chinese texts were generated from AntConc 3.2.2w (Anthony 2008) as it is one of the few freewares that processes Chinese. The English texts were analysed by Topicalizer (Wilmsmann 2008) to list frequent words. One of the benefits of using Topicalizer (*ibid.*) was that, in addition to frequent words, information on frequent two-word phrases up to five-word phrases was also listed. In my text analysis, the ten most frequent words appearing in the 50 texts were:

Rank	Chinese	Word
1	一 (one)	Keyboard
2	該 (demonstrative adjective)	Key
3	鍵盤 (keyboard)	Device
4	的 (of/possessive)	Unit
5	之 (of/possessive)	Circuit
6	按鍵 (key)	First
7	裝置 (device)	Input
8	單元 (unit)	Computer
9	於 (in/at)	Connected
10	係 (is)	Second

**Table 2. The most frequent words in the Chinese and English texts**

In the English corpus, 'keyboard' was the most frequent single word with 190 instances in the 50-text corpus, followed by 'key' and 'device,' with 62 instances. As the texts were searched and the search word 'keyboard' was selected, it could be inferred that the main thread of the invention would be keyboard-related. Since patent abstracts introduce technical inventions, most of the frequent words found were related to the domain of information technology. Two exceptions were the ordinal numbers 'first' and 'second.'

The findings in relation to the most frequent two-word phrases up to five-word phrases were also very much IT-centered. Some of the IT-unrelated frequent phrases included 'the present invention' (15 instances), 'the present invention relates to' (5 instances), 'at least one' (14 instances), 'a plurality of' (19 instances), and 'first and second' (11 instances).

Among the ten most frequent words in the Chinese texts, there were only four nouns. The remaining most frequent words were function words, prepositions, referents, and ordinal numbers. These words could be classified as stop words in the Chinese texts. See Section 6.2.4 for more analysis of the frequent words in the Chinese texts.

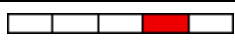
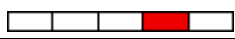
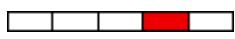
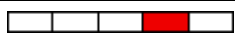
#### **4. Readability - textual analysis**

In this study, readability levels were computed with the use of Flesch

Reading Ease (Flesch 1948), the Flesh-Kincaid Grade Level (Kincaid, Fishburne, Jr. Rogers, and Chissom 1975), the Gunning-Fog Index (Information and Services 2004), and the Automated Readability Index (Senter 1967). The Flesch Reading Ease Formula is considered one of the oldest and most accurate measures to assess the difficulty of English written texts, and has been applied in many U.S. governmental agencies. The Flesh-Kincaid Grade level was originally devised for the U.S. Navy. The Automated Readability Index was designed for U.S. Air Force technical materials. The Gunning-Fog Index was developed by an American businessman to measure reading ease. In these readability tests, word length and sentence length were used as the main measurements but with different weightings.

Texts with higher scores on the Reading Ease test would receive a lower score on the Grade level tests, since texts which are easier to read on average require less schooling. The Gunning-Fog Index (Information & Services 2004) specifies the number of years of education a person should receive in order to understand a text without difficulty. Likewise, the Automated Readability Index (Senter 1967) and the Flesh-Kincaid Grade level (Kincaid, et al. 1975) test use U.S. grade levels to infer the years of education a reader requires to understand a text. As in Flesch Reading Ease (Flesch 1948), higher scores indicate that the texts are more reader-friendly. An example is Time magazine, which has a readability score of 52, and is considered to be best understood by people of high school level and above.

Web-based readability programs such as Edit Central (Editcentral.com 2008) provide automatic computation of text readability by transferring text complexities into scores. According to the output generated online, there were 923 complex words among a total of 6,138 words in the 50 texts. However, the results of the number of complex words included repetitions.

Flesch reading ease score	 47.5
Automated readability index	 15.5
Flesch-Kincaid grade level	 13.1
Gunning fog index	 16.4

**Figure 1. Readability test results from Edit Central (Editcentral.com 2008)**

Of these tests, an average readability level of 14.5 indicated that in order to easily understand these texts, a person should have received at least two years of college education. With regard to the Flesch Reading Ease (Flesch 1948) test, a readability score of 47.5 suggested that patent abstracts were less readable than, say, Time magazine. Although patent abstracts are intended for, and are targeted at the general public, a readability score of 47.5 is definitely not considered as reader-friendly. For readability results of all the texts in this research, please refer to

## Appendix II.

**5. Conclusion**

From the syntactic analysis, lexical analysis, and textual analysis of selected texts, the following conclusions could be drawn. First of all, a more consistent use of short sentences was displayed in the English translated texts than in the Chinese texts. A common usage of shorter words was also evident in the translated texts. Second, the translated texts exhibited low variation yet high frequencies of word usage. In terms of segmentation, the Chinese texts demonstrated a more diversified use of punctuation marks. While short sentences, short word length, and high repetitions of word characterised texts with reading ease, findings from the readability tests, particularly the Gunning-Fog Index, indicated that in order to understand patent abstracts without difficulty, readers should have received at least 14 years of education.

**Bibliography**

- **Anthony, Laurence** (2008). *AntConc* (Version 3.2.2w). Tokyo: Laurence Anthony.
- **Carroll, John B.** (1968). "Word Frequency Studies and the Lognormal Distribution." Paper presented at the Conference on Language and Language Behavior. University of Michigan's Center for Research on Language and Language Behavior.
- —, **Davis, Peter, and Richman, Barry** (1971). *Word Frequency Book*. New York: American Heritage.
- **CKIP** (2004). "Chinese Word Segmentation System with Unknown Word Identification." Academia Sinica. On line at <http://rocling.iis.sinica.edu.tw/CKIP/engversion/wordsegment.htm> (consulted 10.12.2009)
- **Devlin, Siobhan and Tait, John** (eds). (1998). "The use of a psycholinguistic database in the simplification of text for aphasic readers." John Nerbonne (ed.) 1998. *Linguistic Databases. Lecture Notes*. Stanford California: CSLI Publications, 161-173.
- **Editcentral.com** (2008). "Style and Diction." Online at <http://www.editcentral.com/gwt/com.editcentral.EC/EC.html> (consulted on 10.06.2008).
- **European Commission** (2009). "European Commission Public Opinion" (27.02). Online at [http://ec.europa.eu/public\\_opinion/index\\_en.htm](http://ec.europa.eu/public_opinion/index_en.htm) (consulted 28.02.2009).
- **Flesch, Rudolph** (1948). "A new readability yardstick." *Journal of Applied Psychology* 32, 221-233.
- **Impact Information Plain-Language Services** (2004). "Judges Scold Lawyers for Bad Writing." *Plain Language at Work Newsletter*. Online at <http://www.impact-information.com/impactinfo/newsletter/plwork08.htm> (consulted 10.12.2009)
- **INSEAD** (2004). *The Global Information Technology Report 2003-2004*. New York: Oxford University Press.

- **INSEAD** (2008). *Global Information Technology Report 2007-2008*. Geneva: World Economic Forum.
- **Inventec** (2008). *Dr. Eye Dictionary*. Taipei: Inventec Corporation.
- **Irizarry, Estelle** (1990). "Stylistic Analysis of a Corpus of Twentieth-Century Spanish Narrative." *Computers and the Humanities* 24(4), 265-274.
- **Kenny, Dorothy** (2001). *Lexis and creativity in translation : corpus-based study*. Manchester, UK: St. Jerome Publishing.
- **Kincaid, John P. et al.** (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.
- **King, Adam and Flynn, Richard** (2008). UsingEnglish.com. Online at <http://www.usingenglish.com/glossary/lexical-density-test.html> (consulted 09.09.08).
- **Kirchhoff, Hella** (2002). "Simultaneous Interpreting: Interdependence of variables in the interpreting process, interpreting models and interpreting strategies." (D. Sawyer, Trans.). Franz Pöchhacker and Miriam Shlesinger (Eds). *The Interpreting Studies Reader*, New York: Routledge, 110-119.
- **Laviosa, Sara** (2002). *Corpus-based Translation Studies: theory, findings, applications*. Amsterdam, New York: Rodopi.
- **Malmkjaer, Kirsten** (ed.) (2002). *The Linguistics Encyclopedia* (2 ed.). London: Routledge.
- **Olohan, Maeve** (2004). *Introducing Corpora in Translation Studies*. New York: Routledge.
- **RDEC** (2007). 2007 Digital Divide in Taiwan. Taipei: Research, Development and Evaluation Commission, Executive Yuan.
- **Senter, R. J. and Smith, E. A.** (1967). *Automated Readability Index*. Ohio: Aerospace Medical Division.
- **Stubbs, Michael** (1996). *Text and Corpus Analysis*. Oxford: Blackwell.
- **Patents Act** (2004). Online at [http://www.opsi.gov.uk/acts/acts2004/ukpga\\_20040016\\_en\\_1](http://www.opsi.gov.uk/acts/acts2004/ukpga_20040016_en_1) (consulted 10.12.2009).
- **TIPO** (2007). *Annual Patent Statistics*. Online at [http://www.tipo.gov.tw/patent/patent\\_report/95年專利統計.pdf](http://www.tipo.gov.tw/patent/patent_report/95年專利統計.pdf) (consulted 10.12.2009)
- — (2008). *Taiwan Patent Search, 2008*. Online at [http://twpat2.tipo.gov.tw/twcgi/ttsweb?@0:0:1:twpat2\\_e@@@0.13090362023455942](http://twpat2.tipo.gov.tw/twcgi/ttsweb?@0:0:1:twpat2_e@@@0.13090362023455942) (consulted 10.12.2009).
- **TNS Opinion & Social** (2008). *E-Communications Household Survey*. Online at [http://ec.europa.eu/information\\_society/policy/ecomms/doc/library/ext\\_studies/hold\\_07/eb68\\_2infoecomm\\_full.pdf](http://ec.europa.eu/information_society/policy/ecomms/doc/library/ext_studies/hold_07/eb68_2infoecomm_full.pdf) (consulted 10.12.2009).
- **UCREL** (no date). *UCREL CLAWS7 Tagset* Online at

<http://ucrel.lancs.ac.uk/claws7tags.html> (consulted 22.04.2008).

- – (1993). *CLAWS part-of-speech tagger for English*. Lancaster: University Centre for Computer Corpus Research on Language.
- **WEF** (2008). *The Networked Readiness Index Rankings: World Economic Forum*. INSEAD, Cisco Systems.
- **Whissell, Cynthia** (1994). "A Computer program for the objective analysis of style and emotional connotations of prose. Hemingway, Galsworthy, and Faulkner compared." *Perceptual and Motor Skills* 79(2), 815-824.
- **Williams, Jenny and Chesterman, Andrew** (2002). *The Map : a Beginner's Guide to Doing Research in Translation Studies*. Manchester: St. Jerome.
- **Wilmsmann, Björn** (2008). Topicalizer. Online at <http://www.topicalizer.com/> (consulted 12.11.2007).
- **WIPO** (2006a). *Frequently Asked Questions about the International Patent Classification (IPC)*. Geneva: WIPO.
- – (2006b, 01/01/2006). *International Patent Classification (IPC)*. Online at <http://www.wipo.int/classifications/ipc/en/> (consulted 10.12.2009).
- – (2007). *WIPO Patent Report: Statistics on Worldwide Patent Activities (2007 Edition)*. Geneva: WIPO.
- **Youmans, Gilbert** (1990). "Measuring Lexical Style and Competence: The Type-Token Vocabulary Curve." *Style* 24, 584-599.
- – **and Pauley, Nathan** (2008). *Vocabulary Management Profiles*. Columbia: University of Missouri.
- **教育部國語推行委員會** (National Languages Committee of Ministry of Education) (2007). 《重編國語辭典修訂本》網路版 (Revised Chinese Dictionary, web version). Online at <http://dict.revised.moe.edu.tw/index.html> (consulted 20.10.08).
- **教育部國語推行委員會** (National Languages Committee of Ministry of Education) (2008). 《重訂標點符號手冊》修訂版 (Punctuation Handbook, revised version). 台北: 中華民國教育部 (Taipei: Ministry of Education).

## Biography

Yvonne Tsai is a T and I specialist in the Department of Foreign Languages and Literature at National Taiwan University. Her research interest focuses on translation quality of patent abstracts. She can be contacted at [yvtsai@ntu.edu.tw](mailto:yvtsai@ntu.edu.tw)



<sup>i</sup> Eurobarometer consists in a series of surveys performed by the European Commission since 1973, with public opinion reports published on a regular basis to monitor public opinion in the Member States (European Commission).