The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 185-192


# Testing Distributional Hypothesis in Patent Translation

**Hsin-Hung Lin* and Yves Lepage***

**Abstract**

This paper presents a wordlist-based lexical richness approach to testing distributional hypothesis for genre analysis in translation studies. In recent years, there has been continuing interest in patent translation. However, there are only a few lay their interests on comparison between native and non-native writing. The proposed approach to terms distrubution of technical words contained in United States Patent and Trademark Office (USPTO) and Japan Patent Office (JPO) in terms of lexical variation, lexical density and lexical sophistication, in brief, highlights distributional similarity of technical genre, and in particular, distibutional difference of academic and general genres.

**Keywords:** Patent Translation, Native Characterization, Corpus, Co-Occurrence.

## 1. Introduction

As globalization has resulted in rapid greater economic growth, the challenges of interdisciplinary interaction in pursuit of precise patent writing have incredibly increased.

In Lin and Hsieh (2010a), English patent documents were statistically extracted and computationally examined from LexisNexis Academic, a database for legal professionals. They compiled a reference corpus of independent claim texts and lay the focus on their collocation features. Mutual information is attainable with the help of selectional collocation features underlining specific clausal types represented in natural language processing of patent specification.

While their work appears to fill a niche in the ESP (English for Specific Purposes) field (and particularly in the English for Occupational Legal Purposes), Lin and Hsieh (2010b) further compiled a modern patent language technical term list with statistical-retrieval methodologies as a mandatory

* Graduate School of Information, Production, and Systems, Waseda University, Japan.
E-mail: nobuhiro602@toki.waseda.jp; yves.lepage@waseda.jp

approach. The research content and statistical investigations assist patent attorneys expand the vocabulary size for the advancement of patent writing at an international level.

Lin and Hsieh (2011) proposed a mixed-method approach to detecting scholarly discourse in patent technical documents. The Patent Technical Word Corpus (hereafter PTWC), containing 16 million word tokens, was compiled to elucidate the underpinning principles in identifying discourse elements, text-structure components, and the location of references. Whereas most existing IPR (intellectual property rights) databases accessible for information retrieval, the creation of PTWC, based on corpus-statistics and text-processing technology, refines more decisive characteristics of terminological knowledge as potential contribution for evaluation of technical documents.

To characterize technical genre in translation studies, we use lexical richness based on technical wordlist to test distributional hypothesis.

## 2. Technical Terms Distribution

We firstly conduct a quantitative survey based on USPTO Glossary to rank the distribution of technical terms used in United States Patent and Trademark Office (USPTO) and Japan Patent Office (JPO) within the time period from year 2010 to 2013. Table 1 below presents the statistical results.

'Comprising', a term of art used in claim language which means that the named elements are essential in describing the invention, ranked the first in USPTO. According to USPTO Glossary, it is a transitional phrase that is synonymous with "including," "containing" or "characterized by;" is inclusive or open-ended and does not exclude additional, unrecited elements or method steps. On the contrary, 'consisting of', a transitional phrase that is closed and excludes any element, step, or ingredient not specified in the claim, ranked the $6^{th}$.

To characterize transitional phrases of technical genre in translation studies, we retrieved co-occurring information of 'comprising' and 'consisting of' to compare it with academic and general genres.

***Table 1. Distribution of patent technical words in USPTO***

| Rank | Term | Frequency | Rank | Term | Frequency |
|---|---|---|---|---|---|
| 1 | comprising | 3785213 | 11 | specification | 854667 |
| 2 | scope | 2459656 | 12 | continuation | 738785 |
| 3 | patent | 1603882 | 13 | dependent claim | 625886 |
| 4 | Group | 1306808 | 14 | composed of | 617353 |
| 5 | element | 1245265 | 15 | independent claim | 587926 |
| 6 | consisting of | 1165427 | 16 | representative | 518762 |
| 7 | drawing | 1015261 | 17 | benefit claim | 437599 |
| 8 | disclosure | 919881 | 18 | person | 383784 |
| 9 | application (patent) | 884470 | 19 | priority claim | 381352 |
| 10 | patent application | 884470 | 20 | interference | 341173 |

We give the survey of terms used in JPO in Table 2. It is noted that "comprising" ranked the first in distribution of USPTO and JPO, whereas "consisting of" ranked the 6$^{th}$.

***Table 2. Distribution of patent technical words in JPO***

| Rank | Term | Frequency | Rank | Term | Frequency |
|---|---|---|---|---|---|
| 1 | comprising | 629750 | 11 | applicant | 60293 |
| 2 | composed of | 371852 | 12 | drawing | 53469 |
| 3 | element | 272496 | 13 | person | 48893 |
| 4 | POWER | 272088 | 14 | IDS | 24946 |
| 5 | Group | 176103 | 15 | Control No. | 22905 |
| 6 | consisting of | 136992 | 16 | interference | 22445 |
| 7 | PAIR | 122746 | 17 | RE | 19777 |
| 8 | representative | 72519 | 18 | specification | 18102 |
| 9 | Request (PCT) | 70606 | 19 | classification | 16977 |
| 10 | application (patent) | 62027 | 20 | independent claim | 15513 |

## 3. Methodology

### 3.1 The Distributional Hypothesis

Sahlgren (2008:33) maintains that distributional approaches to meaning acquisition utilize distributional properties of linguistic entities as the building blocks of semantics. This hypothesis is often stated in terms like words which are similar in meaning occur in similar contexts (Rubenstein & Goodenough, 1965). In other words, words that occur in the same contexts tend to have similar meanings (Pantel, 2005).

### 3.2 Corpus Preparation

Transitional phrases in patent application were used to specify whether the claim is limited to only the elements listed, or whether the claim may cover items or processes that have additional elements. The most common transitional phrase used is the open-ended phrase "comprising". However, many claims use closed-ended language such as "consisting of".

In this regards, we retrieve co-occurring information containing "comprising" and "consisting of" from LexisNexis Academic for corpus preparation. Table 3 shows the structure for the corpus creation.

***Table 3. Genre-based co-occurrence corpus of transitional phrases***

| Genres | Native Writing | Non-Native Writing |
|---|---|---|
| Technical (Patent) | USPTO | JPO |
| Academic (Law Journal) | Canadian Legal Journals | HK Law Journal |
| General (Newspapers) | US Newspapers | Non-US Newspapers |

### 3.3 Lexical Richness

Lexical richness is a concept about one's lexical uses, which can be measured by lexical density, sophistication, and variation (Kao and Wang, 2014:54).

Kojima and Yamashita (2014:23) suggest that lexical richness measures primarily assess learners' vocabulary use. Lexical variation, the proportion between different words (types) and the total number (tokens) of words used in the text, is known as the type-token ratio (TTR).

Lexical density is defined as the percentage of lexical words in the text, for

example, nouns, verbs, adjectives, and adverbs (Laufer and Nation, 1995:309). Since only content words carry semantic meanings, a greater lexical density indicates more semantic information conveyed in a text.

Read (2000: 200) distinguishes dimensions of lexical richness, and one of these is lexical sophistication, which he defines as 'the use of technical terms and jargon as well as the kind of uncommon words that allow writers to express their meanings in a precise and sophisticated manner'. The proportion of words used at different frequency levels, in terms of K1, K2, AWL (Academic Word List), and off-list words, in the text. K1 and K2 words are the most commonly used first 1000 and 1001 to 2000 words, respectively, in English. Words beyond these K1, K2, and AWL are placed into the off-list level, where proper nouns, rare words, special terms, acronyms, abbreviations, incompletions, and even misspellings may be found.

## 4. Results and Discussion

In terms of lexical density, non-natives employed more semantic information than the natives, among all genres. In terms of lexical variation, non-natives employed more lexical diversity than the natives in technical and academic genres.

Academic Genre, in particular, HK Law Journal, containing most semantic information (83%), among the all, whilst general genre, Non-Us Newspapers, containing least lexical diversity, as we excluded technical genre for analysis.

### 4.1 Technical Genre

In technical genre, in particular, JPO (Patent Abstract of Japan), containing least advanced words (15.03%) in the texts, among all.

***Table 4. Lexical sophistication of "comprising" in technical genre***

| Word Level (%) | USPTO | JPO |
|---|---|---|
| K1 Words | 50.35 | 50.37 |
| K2 Words | 2.61 | 3.61 |
| AWL Words | 23.74 | 21.78 |
| Off-List Words | 23.31 | 24.24 |

Less vocabulary knowledge in K2, AWL, and Off-list words were employed in "consisting of", compared with that of "comprising". The natives employed more academic words in Table 4, more off-list words in Table 5.

***Table 5. Lexical sophistication of "consisting of" in technical genre***

| Word Level (%) | USPTO | JPO |
|---|---|---|
| K1 Words | 62.37 | 64.89 |
| K2 Words | 0.29 | 0.34 |
| AWL Words | 19.43 | 19.74 |
| Off-List Words | 17.92 | 15.03 |

### 4.2 Academic Genre

In academic genre, HK Law Journal, containing most advanced words (33.28%) in the texts, among all.

***Table 6. Lexical sophistication of "comprising" in academic genre***

| Word Level (%) | Canadian Legal Journal | HK Law Journal |
|---|---|---|
| K1 Words | 59.02 | 51.02 |
| K2 Words | 5.88 | 3.07 |
| AWL Words | 13.47 | 12.63 |
| Off-List Words | 21.63 | 33.28 |

As shown in Table 6 and Table 7, non-natives employed more off-list words in academic legal genre, whereas the natives employed more K1 and K2 words in academic legal genre.

***Table 7. Lexical sophistication of "consisting of" in academic genre***

| Word Level (%) | Canadian Legal Journal | HK Law Journal |
|---|---|---|
| K1 Words | 59.74 | 50.74 |
| K2 Words | 6.12 | 2.97 |
| AWL Words | 13.07 | 14.24 |
| Off-List Words | 21.07 | 32.05 |

### 4.3 General Genre

As can be seen in Table 8 and Table 9, the non-natives employed more K2, AWL, and Off-list words but less K1 words in general genre.

***Table 8. Lexical sophistication of "comprising" in general genre***

| Word Level (%) | US Newspapers | Non-US Newspapers |
|---|---|---|
| K1 Words | 63.34 | 51.97 |
| K2 Words | 3.69 | 5.47 |
| AWL Words | 11.66 | 16.73 |
| Off-List Words | 21.30 | 25.84 |

***Table 9. Lexical sophistication of "consisting of" in general genre***

| Word Level (%) | US Newspapers | Non-US Newspapers |
|---|---|---|
| K1 Words | 63.37 | 53.48 |
| K2 Words | 4.03 | 7.71 |
| AWL Words | 12.61 | 16.42 |
| Off-List Words | 19.99 | 22.09 |

In short, K1 words were employed more by the natives in academic and general genres, whilst less used in technical genres.

## 5. Conclusion and Future Work

There is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latter (Sahlgren, 2008:33). In terms of distribution statistics, the technical genre reveals more distributional and meaning similarity.

In summary, lexical richness is a valid and reliable measure to characterize genres. For future research, we seek to investigate the origin differences between syntagmatic and paradigmatic relations to further refine the preliminaries of the present study.

## References


Kao, S. M, & Wang, W. C. (2014). Lexical and organizational features in novice and experienced ELF presentations. *Journal of English as a Lingua Franca, 3*(1), 49-79.

Kojima, M., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions. *System, 42*, 23-33.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical density in FL written production. *Applied Linguistics, 16*, 307-322.

Lin, H. H., & Hsieh, C. Y. (2010a). Collocation Features of Independent Claim in US Patent Documents: Information Retrieval from LexisNexis. *ROCLING XXII: Conference on* Computational Linguistics and Speech Processing, pp. 296-310. Taiwan: Academia Sinica.

Lin, H. H., & Hsieh, C. Y. (2010b). The Specialized Vocabulary of Modern Patent Language: Semantic Association in Patent Lexis. *PACLIC 24: Pacific Asia Conference on Language, Information, and Computation*, pp. 417-424. Japan: Waseda University Press.

Lin, H. H., & Hsieh, C. Y. (2011). Characteristics of Independent Claim: A Corpus-Linguistic Approach to Contemporary English Patents. *International Journal of Computational Linguistics and Chinese Language Processing, 16*(3-4), 77-106.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Conference of the Association for Computational Linguistics*, pp. 125–132). Morristown, NJ, USA: Association for Computational Linguistics.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8* (10), 627-633.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics, 20*(1), 33-53.