CHAPTER 1

INTRODUCTION

1.1 Natural Language Processing

Natural Language Processing (NLP) refers to the computerized approach to analyzing text that is based on a set of theories and technologies (Liddy, 2001). The author defined NLP as a "theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications".

The element of *levels of linguistic analysis* in the definition above is further detailed into several levels of language: phonology, morphology, lexical, syntactic, semantic, discourse and pragmatic (Liddy, 2001). While all levels contribute to meaning, most people think that meaning is determined only at the level of semantics.

1.2 Machine translation and Word Sense Disambiguation

Machine translation was the first computerized application related to NLP (Liddy, 2001). Early systems developed for machine translation used simple dictionary-lookup for suitable words for translation and merely reordered the words after translation in order to obey the rules of the target language, disregarding the lexical ambiguity present in natural language. In the late 1970's however, attention shifted to semantics, discourse and communication.

One such concern in semantics is Word Sense Disambiguation (WSD), where much research has been done. Wilks and Stevenson (1996) described WSD as an "intermediate task", which is necessary at one level or another to accomplish most NLP tasks.

WSD involves determining the meaning (*sense*) of a given word in a text thus playing an important role in NLP. The process usually involves two steps: (1) establishing all the available senses for the word; and (2) assigning the word to the appropriate sense. An example can be illustrated with the sentence "*Can you give me a **hand**?"* Table 1 below lists all the senses for *hand*, the definitions and examples.

Sense	Definition	Example
1	(Noun) the extremity of the superior limb	*The child washed his hands.*
2	(Noun) hired laborer on a farm or ranch	*A ranch hand*
3	(Noun) something written by hand	*His hand was illegible.*
4	(Noun) ability	*She wanted to try her hand at knitting.*
5	(Noun) the cards held in a card game by a given player at any given time	*Amanda didn't hold a good hand all evening.*
6	(Noun) one of two sides of an issue	*On the other hand*
7	(Noun) a rotating pointer on the face of a timepiece	*The longer hand counts the minutes.*
8	(Noun) a unit of length equal to 4 inches used in measuring horses	*The horse stood 23 hands.*
9	(Noun) a card player in a game of bridge	*We need a 4th hand for bridge.*
10	(Noun) a round of applause to signify approval	*Please give the performers a big hand.*
11	(Noun) the terminal part of the forelimb in certain vertebrates (e.g. kangaroos and apes) "	*The kangaroo's forearms seem undeveloped but the powerful five-fingered hands are skilled at feinting and clouting.*
12	(Noun) physical assistance	*Give me a hand with the housework.*
13	(Verb) place into the hands or custody of	*Could you hand me the pepper, please?*
14	(Verb) guide or conduct or usher somewhere	*John hand the old man to the bus.*

Table 1. Senses, definitions and examples of *hand*

The above senses for the word *hand* are taken from the Princeton WordNet (Fellbaum, 1998), which will be discussed further in 3.1. As presented, *hand* has a total of 14 available senses, including two verb senses. The second step of WSD now requires us to select an appropriate sense that fits the sentence "*Can you give me a hand?"* In this case the most suitable sense is (12). If given the context, speakers of the language will be able to understand the sentence. However, a computer system would need a mechanism of using the context to determine the meaning if it were to depend on semantic interpretation (McCarthy, 2009).

The assignment of senses to words often relies on two major sources of information: (1) the context of the word being disambiguated, including information contained within the text or discourse in which the word appears and non-linguistic information about text such as situation, etc. and (2) external knowledge sources, such as lexical, encyclopedic or other resources that provide useful data to associate words with senses (Ide & Véronis, 1998).

WSD is not only a necessity for language understanding applications; it is also helpful in applications like machine translation where language understanding is not the main focus (Ide & Véronis, 1998). Machine translation systems benefit from progress in WSD as it helps to generate more human-like and accurate translation. Other applications that benefit from WSD include text processing, speech processing, information retrieval and hypertext retrieval, grammatical analysis and content and thematic analysis.

1.3 Translation divergences and translation shifts

Current trends in computational linguistics research tend to deal with the issue of how to represent meaning. The knowledge that languages present the same information in various ways is widespread. The differences in language are often addressed in translation studies, where the study of translation shifts has a long-standing tradition (Cyrus, 2006). In machine translation, these mismatches are studied as translation divergences (Dorr, 1993) or translation shifts in corpus linguistics.

In translation theories, Vinay and Darbelnet (1958/1995) were prominent with their approach to translation shifts. They theorized four types of translation shifts:

1. Transposition – a change in word class e.g. *scared*, an adjective in English, translated to *びびる bibiru* "to feel frightened", a verb in Japanese.
2. Modulation – a change in semantics e.g. *thumb* to *指 yubi* "finger" in Japanese
3. Equivalence – completely different translation, but the meaning is still apparent e.g. proverbs
4. Adaptation – change of situation due to disparities in culture e.g. in Japanese society, the phrase *いただきます itadakimasu* is said before a meal. The expression is used to show appreciation to the plants and animals that gave their lives for the meal about to be consumed. In addition, it is also used to give thanks to the people involved in the whole process, which ranges from farmers to the preparer of the meal. As this practice is not common in the English society, there is no exact translation for the phrase. Instead, it is often loosely adapted as "let's eat" or "thank you for the meal".

Catford (1965) described *shift* as "departures from formal correspondence" between source and target language. The author distinguished *formal correspondence* from *translational equivalence.* The former exists between source and target texts that occupy nearly the same place in their respective languages. *Translational equivalence* exists when two texts essentially translations of each other. In other words, if the translation equivalents are not the exact correspondents, a *translation shift* is said to have occurred.

Previous literature on contrastive analysis among languages has also shown that various types of discrepancies can occur between a source and target language (Vinay & Darbelnet, 1958/1995; Marello, 1989; Dorr, 1993). These lexical divergences between various pairs of languages substantiate that lexicalized conceptual hierarchies are not universal among all languages. Bentivogli and Pianta (2000) provided a list of common lexical singularities between a source and target language. The circumstances where translation shifts occur are as follows:

1. Synthetic deviations occur when a translation equivalent does not have the same synthetic structure as the source language (Dorr, 1993).
2. Lexical deviations occur when the source and target languages lexicalize a similar concept with a different word or phrase. Lexical gap, which is an

absence of a translation equivalent of a word from the source language, is another form of lexical deviation (Vinay & Darbelnet, 1958/1995; Marello, 1989).

3. Differences in connotation happen when the translation equivalent is unable to reproduce all the implicatures and connotations expressed by a word from the source language (Brown, Mendes & Natali 1995).
4. Lastly, denotation differences occur when the denotation of a word from the source language partially overlaps with the denotation of the translation equivalent (Lo Cascio *et al,* 1995).

Studies on translation shifts are important for translators to assist them in choosing suitable translation strategies and to avoid translation mistakes as well as unwarranted influence of the source to target language (Baker, 1996; Teich, 2003). While translation shifts may have been a common research area in translation studies, they have not yet been studied extensively and systematically in corpus linguistics (Cyrus, 2006).

A corpus-based study of translation shifts will be potentially interesting for linguists as well as translation theorists, who can evaluate the phenomenon empirically. With this in mind, parallel and comparable corpora play an essential role in such studies, as they not only serve as the base of intuitions with actual examples but also allow the quantitative studies of translation shifts in different genres and other domain-specific shifts (Čulo et al., 2008).

Studies on translation shifts can also help to improve machine translation systems, as the occurrence of translation shifts is a major challenge in achieving human-like translation (Ahrenberg, 2007).

This study aims to not only investigate the translation shifts and describe them but also to contribute to the area of machine translation research, as data from the study will be released. This data can then be used in training and improving machine translation systems so that their output can resemble human translation.

This paper is structured as follows. In Chapter 2 we review previous research that has worked on the relevant topics on annotation and translation mismatches. Chapter 3 introduces a detailed documentation of the resources and methodology taken in this study. The results of the study is presented in Chapter 4 and discussed in Chapter 5. Lastly, we conclude the study and suggest future works in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

Several studies have been working with parallel corpora to investigate translation shifts and cross-lingual divergences. Bentivogli, Forner and Pianta (2004) examined the effectiveness of a cross-lingual annotation transfer methodology in the MultiSemCor Corpus, an English-Italian parallel corpus based on the English SemCor corpus. The authors hypothesized that semantic information is predominantly preserved during translation process. Based on this hypothesis, annotations can be transferred from the source language to the target language using word alignment as a bridge. The researchers found that the main bulk of incorrect transfer of annotation was due to translation shifts.

Cyrus (2006) worked on FuSe, an English-German parallel corpus with its text extracted from the EUROPARL corpus (Koehn, 2002) that covers proceedings of the European parliament. The researcher based the approach on predicate-argument structures as they capture the main piece of information in the sentence and are mostly likely to be represented in both source and target sentences. The predicates were marked monolingually for part-of-speech (POS) and lemma. The arguments were given short intuitive role names based on the predicates. After annotation, the predicates and arguments were then aligned to the corresponding words in the target language. Translation shifts in the study were categorized on two levels: grammar and semantics, which were in turn classified into six types respectively. For grammatical shift, these include

1. Category change – a change in POS
2. Passivisation – an active predicate is translated as a passive predicate
3. Depassivisation – a passive predicate translated as an active predicate
4. Pronominalisation – common noun or proper noun translated to pronoun
5. Depronominalisation – pronoun translated to common or proper noun
6. Number change – change in plurality

The other six types of shifts in the level of semantics are

7. Semantic modification – words are not exact correspondence

8. Explicitation – subcategory of semantic modification. This was assigned when the target word contained more information than the source word.
9. Generalization – source word contained more information than target word
10. Addition – a word in the target language that had been added in the translation
11. Deletion – a word in the source language that was not translated in the target language
12. Mutation – words were translation equivalents but were different in meaning

While the annotation work seemed rather comprehensive, it is also idiosyncratic. Also, the drawback with this study is that the words were not supported by any semantic resource such as WordNet. Furthermore, the resource has also not been released and development has ceased.

LinES is an English-Swedish Parallel Treebank with 2,400 sentences taken from a user manual and a novel (Ahrenberg, 2007). It is different from FuSE as it focuses on complete alignments of segment pairs and (semi-)automatic derivation of shifts. Most importantly, LinES was created with the intention of studying translation shifts in terms of syntactic structures. Therefore, annotation was mainly syntactically oriented rather than semantics.

There is also a German-English parallel and comparable corpus of twelve texts from eight different genres (Culo et al., 2008). The CroCo corpus contains about 1,000,000 words and each sentence is annotated with phrase structures and grammatical functions. Words and phrases are also aligned across parallel sentences. However, like the FuSe, it is also not systematically backed by any semantic resources like WordNet.

Another parallel treebank, SMULTRON, contains parallel texts mostly in German, English and Swedish (Volk et al., 2010). It consists of 2,500 sentences from a variety of genres such as novels, economy texts, a DVD manual and mountaineering reports. Mountaineering reports were also available in French and a Spanish version of the DVD manual was also accessible. The sentences were tagged for POS and annotated with phrase structure trees. The phrase structure trees were also aligned on sentence, phrase and word level. Lemma information was also contained in the German and Swedish monolingual treebanks. An initial study has been done on 50 sentences in the

English-Swedish parallel subcorpus where the sentences had been annotated with semantic frames. However, similar to LiNES, the research development is mainly focused on alignment of syntactic structures.

Padó and Erk (2010) conducted a study of translation shifts on a German-English parallel corpus of 1,000 sentences from EUROPARL (Koehn, 2002). The sentences were aligned at word level and annotated with semantic frames from FrameNet (Baker et al., 1998). The researchers aimed to measure the practicability of frame annotation projection across languages. The results showed a significant but small portion of cross-lingual semantic mismatches which equates to translation shifts. This is different from the present study, which uses WordNet as the base semantic resource, in which the word senses have a smoother granularity than frames.

Corpus-based studies on translation shifts as illustrated above focused on European languages. To our knowledge, this present study will be the first corpus-based study of translation shifts that involves Asian languages – mainly Japanese and Chinese.

CHAPTER 3

METHODOLOGY

3.1 Resources

3.1.1 WordNets

One of the most challenging problems researchers face in NLP is semantic information analysis such as WSD as mentioned above in 1.2 (Xu, Gao, Qu & Huang, 2008). Thus, a large and computable semantic resource is essential so that machines can be trained to understand information present in natural language. Machine-readable dictionaries make up a significant portion of these semantic resources and WordNet is an example of machine-readable dictionaries. In this study, we made use of three WordNets, one for each of the languages.

The Princeton WordNet (PWN) of English is developed and maintained by Fellbaum (1998). It is free and publicly available for download. Since its publication, the WordNet has become the primary source of referent for tasks comprising WSD (Bentivogli, Forner & Pianta, 2004). Words are arranged in terms of synonyms and are referred to as synsets. The synsets are further linked to one another through various lexical and semantic associations. Relationships among words in the WordNet are mainly built on synonymy but due to its structure, it also allows users to see the conceptual relationship between words such as hyponymy, holonymy, taxonomy and so on (Fellbaum, 2005). Figure 1 below illustrates some of the concepts associated with the word *table.*

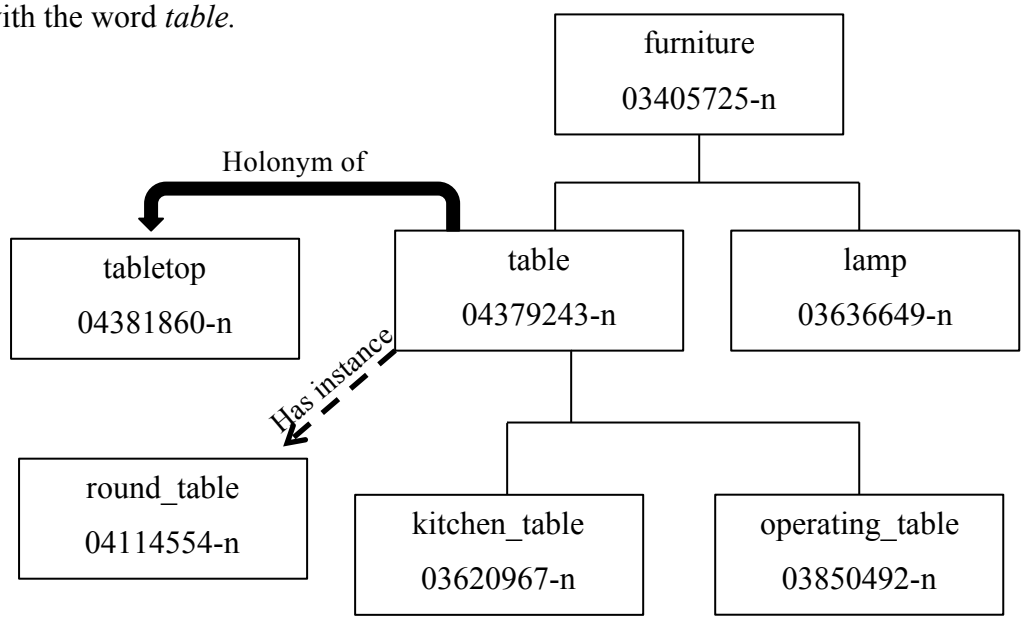

Figure 1. WordNet representation of *table*

The word *table* tagged with 04379243-n is a direct hyponym of *furniture* and direct hypernym of *kitchen table, operating table* and others not shown here. It is a holonym of *tabletop* and has an instance *round table*. A holonym is the name of the whole of which contains a part Y. Meronym is the term for the constituent that makes up the whole. Thus, if X is part of Y, Y is a holonym of X and X is a meronym of Y. An instance is a proper noun that refers to a singular referent and is a specific form of hyponym.

The Japanese WordNet (JWN) was first published in 2009 (Isahara, Bond, Uchimoto, Utiyama and Kanzaki, 2008). Synsets in the JWN are based on the same lexical arrangement as the PWN. This implies that lexical units in the JWN are also arranged according to their hierarchical relations among words. Ongoing work includes adding synsets that are not in the PWN and also modifying the structures in the hierarchy so that they make a better representation of the Japanese language.

Like the JWN, the semantic hierarchy in the Chinese WordNet (CWN) also derives from PWN (Xu, Gao, Pan, Qu and Huang, 2008). The researchers translated the original English WordNet into Chinese using automatic translation approaches. Several other linguistic resources were used to build the WordNet, such as American Heritage Dictionary and X-Dict Dictionary.

While the JWN and CWN are built with their structures in accordance to the PWN, there are still gaps between the three WordNets. One reason is that the JWN and CWN are less developed compared to the PWN. Table 2 shows a summary of each WordNet.

	Princeton WordNet	**Japanese WordNet**	**Chinese WordNet**
Synsets	117,659	57,238	111,045
Words	155,287	93,834	115,136
Senses	206,941	158,058	168,824

Table 2. Summary of number of synsets, words and senses in each WordNet

As Japanese and Chinese lemmas are simply added to existing synsets in the PWN, there are some synsets in the JWN and CWN that are not yet represented in the PWN and vice versa. This is often due to the uniqueness of the languages (Bond et al., 2009).

In addition to the lack of representation of certain concepts in the WordNets, we would also like to call to attention the minor errors that are present in the WordNets. For example, in the CWN, the POS tag for some words may be incorrect, such as *睡眠 shuìmián* "sleep" is tagged as a verb instead of noun. There are also instances where a wrong lemma is included in the synset. In view of this, we hope to report and correct these errors as we spot them so as to help improve the WordNets.

3.1.2 Corpus

We conducted this study with parallel and aligned versions of "The Adventure of the Dancing Men", part of the Sherlock Holmes short stories written by Sir Arthur Conan Doyle. As the copyright for the original text has expired, redistributable translations in Japanese and Chinese were readily available for use.

This parallel tritext is part of the NTU Multilingual Corpus that is currently being developed (Tan & Bond, 2011). Using original English text as the source language, the Japanese and Chinese texts were first aligned at the sentence level. The texts were tokenized and automatically tagged for POS and lemma. They were then manually sense-tagged with reference to the respective WordNets. Table 3 below shows the composition of each text.

	English	Japanese	Chinese
Sentences	599	698	680
Words	6,842	5,246	5,148
Concepts	11,198	13,483	11,325

Table 3. Summary of number of sentences, words and concepts in each text

3.2 Annotation of relationship between words

In this study, we worked on the available tagged data. The initial word alignment was done using a program that checks each word in a sentence pair. The program links the word if

(i) they are the same synset e.g. *say* and *言う iu* "to say", both tagged with the synset ID 00979870-v, defined as "utter aloud"

(ii) they are direct hypernyms or hyponyms e.g. *wash,* 01535246-v, defined as "cleanse with a cleaning agent, such as soap, and water" is a hypernym of *洗い落とす araiotosu* "to wash out", tagged to 01535742-v and defined as "wash free from unwanted substances, such as dirt".

Mok (2012) tagged the relationship between words in the English and Chinese data. Different symbols were used to show the different relationships between source and target words. We used the tagging guidelines in Mok (2012) as a base to establish the relationship between the words in the English and Japanese data.

For words which share the same synset, "=" is used to show the link between them. For example, *notebook* is tagged with the synset 06415419-n, defined as "a book with blank pages for recording notes or memoranda". In the Japanese version, this concept is translated to *備忘録 bibouroku* "notebook" with the same synset 06415419-n. Hence, we linked the words with the "=" symbol.

When the target word is the direct hyponym of the source lemma (English), ">" is used to show the link. For example, *finger*, tagged with 05566504-n, defined as "any of the terminal members of the hand (sometimes excepting the thumb)", is a direct hypernym of *人差し指 hitosashiyubi* "index finger", tagged with 05567381-n, defined as "the finger next to the thumb". This relationship is tagged with ">".

Likewise, if it is the direct hypernym, "<" is used to show the link. An example is *sunlight,* tagged with 11485367-n, defined as "the rays of the sun", is a direct hyponym of *光 hikari* "visible radiation", tagged with 11473954-n, defined as "electromagnetic radiation that can produce a visual sensation". This relationship is marked with "<".

For words with related meaning that are not identical, the symbol "~" is used to show the link. This includes instances like meronymys, antonyms and similar relationships other than direct hyponymy. One example would be notebook, tagged with 06415419-n, and *メモ帳 memochou* "notepad" tagged with the sense 15021085-n, defined as "a pad of paper for keeping notes". In WordNet, these two synsets are not related in any hyponymy nor meronymy relationship. However, they do share a similar meaning as both refer to an object for recording notes. Therefore, we linked the words together with "~".

An additional instance would be words with POS gaps such as *dull,* an adjective tagged with 00393992-a, defined as "(of color) very low in saturation". In the Japanese text, this is translated to *くすんだ kusunda* "darken", a verb tagged with 00312380-v, defined as "become dark or darker". In this case, they were also linked with "~".

In addition to the four symbols, we introduced a new symbol ":" to indicate combinations of words or phrases that are translation equivalent of the original source but are not lexicalized enough to be linked. One example is shown in the phrase below

(1) English: *be* ***content*** *with my word*
Japanese: *わたくし の 言葉 を* ***信じ*** *-て*
watakushi no kotoba wo ***shinji*** *-te*
1SG POSS word ACC to believe-te
"believe in my words"

In the example, we can perceive the link between individual concepts. However, while we comprehend that the Japanese version is a translation equivalent for the original English text, we would not want to say that *content* is of similar or related meaning to the lemma *信じる shinjiru* "to believe". Generally speaking, these two words are not clearly linked lexically. However, to show that in this context they are somewhat related, we used the symbol ":" to show the connection.

Another instance would be in the following phrase

(2) English: *several fresh dancing men **pictures***

Japanese:	*何*	***枚***	*か*	*新しい*	*踊る*	*人形*
	nan	***mai***	*ka*	*atarashii*	*odoru*	*ningyou*
	DET	CL	Q	new	to dance	figure

"several pieces of new dancing men (pictures)"

In classifier languages, a noun can sometimes be deleted and referred to by the classifier when the reference to it is repeated (Denny, 1986). In example (2) above, the classifier *枚 mai*, used for thin, flat objects is used to represent the concept of *pictures*. While the two words are related in a classifier-noun connection, we do not want to say that they are exact correspondents of each other. Thus, we used ":" to denote it.

For direct antonyms, we used the symbol "!" to represent the relationship. A simple example would be *hot* translated to *寒くない samukunai* "not cold", whereby the lemma is *寒い samui* "cold".

We used another symbol, "#" to represent weak antonyms, which were observed more often than direct antonyms in the English-Japanese pair. One such example is *propose*, tagged to 00708980-v and defined as "propose or intend" in the sentence below

(3) English: "*you do not **propose** to invest in South African securities?"*

Japanese:	*「君*	*は、*	*南-アフリカ*	*の*	*証券*	*へ*	*の*
	kimi	*wa*	*minami-afurika*	*no*	*shouken*	*he*	*no*
	2SG	NOM	south-africa	POSS	securities	DAT	NMZ

投資	*を*	***思いとどまった。」***
toushi	*wo*	***omoitodomat-ta***
investment	ACC	to hold back-PST

"You held back investment in South African securities?"

In the Japanese text, *propose* was translated to the lemma *思いとどまる omoitodomaru,* tagged to the synset 00613393-v and defined as "abandon idea or claims; stop maintaining or insisting on". This synset represents a meaning opposite that of *propose.* However, as this is unlike a definite antonym such as *hot* and *cold,* we interpreted this as a weak antonym and assigned the symbol "#" to show the relationship.

We gave a confidence level of 95% to each hand-tagged relationship. Table 4 shows a brief summary of the symbols used to annotate the relationships. In this study, a translation shift has occurred when concepts are linked with symbols other than "=".

Symbol	**Relationship**
=	Same synset
>	Direct hyponym of source language
<	Direct hypernym of source language
~	Similar or related meaning, including different POS
:	Translation equivalent but not clearly linked lexically
!	Direct antonym
#	Weak antonym

Table 4. Symbols used to indicate relationships between words

In addition to tagging relationships between content words, we also showed relationships between interjections, pronouns-proper or common nouns and vice versa, as illustrated in the phrase below.

(4) English: *"**So**, Watson," said **he**, suddenly,*

Japanese: 「***だから*** ワトソン一」と ***ホームズ*** が 突然

Dakara *watoson-* *to* ***ho-muzu*** *ga* *totsuzen*

so watson to holmes NOM suddenly

口 を ひらく。

kuchi *wo* *hiraku*

open ACC to open

"'So, Watson,' Holmes opens his mouth suddenly."

In the example above, we linked *so* with *だから dakara* "so" and *he* with *ホームズ ho-muzu* "holmes". Both concepts are linked with "~", with the comments "interjection" and "pronoun-proper noun" respectively. For pronouns and some of the interjections, the WordNets do not have a synset for the concepts. The reason why we are linking these concepts is to show that they are translation equivalents. Also, it helps users of the data to see that pronominalisation occurs during translation and vice versa depending on the translating style.

3.3 Tagging issues

This section documents some of the tagging issues we came across as we were tagging the relationships between words and concepts for the English-Japanese pair. Tagging issues with the English-Mandarin pair have been discussed in Mok (2012) and will not be presented in detail here.

3.3.1 Changes to original tags

As we were going through each sentence pair, we found that there were errors in the previous semantic tagging for some of the words. We corrected these as we progressed with the tagging. For example, *club* in the phrase, *"returned from the club last night"*, was initially given the synset 03054311-n, defined as "a building that is occupied by a social club". The Japanese translation *クラブ kurabu* "club" was given the synset 02931417-n. This synset is defined as "a spot that is open late at night and that provides entertainment (as singers or dancers) as well as dancing and food and drink". Both the semantic tags given to the words are legit, however, there is another sense that fits the context more accurately. This is the synset 08227214-n, with the definition "a formal association of people with similar interests". Therefore, we corrected the original tags for both and linked the words with "=".

We also tried to tag words with "=" whenever possible. Therefore, when the tagged sense for one word is appropriate for the corresponding word, we changed it accordingly. For instance, *key* was originally tagged with the synset 06424869-n, defined as "a generic term for any device whose possession entitles the holder to a means of access". The Japanese equivalent *鍵 kagi* "key" was tagged with synset 03613294-n, defined as "metal device shaped in such a way that when it is inserted

into the appropriate lock the lock's mechanism can be rotated". Both senses are appropriate for either word. However, the synset 06424869-n is not available in the JWN. On the other hand, the synset 03613294-n is available in both PWN and JWN. Thus, we changed the original tagged sense for the English word from 06424869-n to 03613294-n, and linked the words with "=".

We also changed the original tags for some words that were tagged with "s" which denotes a missing sense. This is quite often the case with Japanese words that were represented in *hiragana* only. For example, in the phrase below

(5) English: *to* *steady* *the* *cue*
Japanese: キュー が すべら-ない よう
kyu- *ga* *subera-nai* *you*
cue NOM to slide-NEG in order
"so that the cue does not slide"

The lemma for すべら *subera,* すべる *suberu,* was initially tagged with "s". This was because the lemma in its *hiragana* form is not listed in the JWN. However, 滑る *suberu* "to slide" which is a *kanji* representative for the word, is listed and a suitable synset 01870275-v is also available. This synset is defined as "move obliquely or sideways, usually in an uncontrolled manner". As the synset is appropriate for the lemma, we corrected the original tag from "s" to 01870275-v. Also, as we deemed すべら *subera* "to slide", together with the negative marker ない *nai,* as a translation equivalent for *steady* but not clearly linked lexically, we linked the two words with ":".

In Japanese, it is not uncommon to find a concept represented with a multiword expression, as exemplified by example (6).

(6) English: *said* *he* *suddenly*

Japanese: *ホームズ* *が* *突然* *口 を 開く*

ho-muzu *ga* *totsuzen* *kuchi wo hiraku*

Holmes NOM suddenly mouth ACC to open

"Holmes opens his mouth suddenly"

This concept could be replaced with a simple verb *言う iu* "to say", just like the original English text. However, the translator chose to use a multiword expression. The tokenizer did not recognize this as a multiword concept previously. As this concept is lexicalized, we grouped the words together and tagged the new concept to the synset 00941990-v, where the lemma "speak" is defined as "express in speech". The concepts were then linked together with "~".

3.3.2 Changes to the tagged Chinese data

Although the English-Chinese data was tagged before the English-Japanese data, tagging was not complete and many words that could be linked were not tagged. In addition, we introduced the two symbols ":" and "#" after the English-Chinese data was tagged. Hence, we retagged the English-Chinese corpus and checked through the relationship between words when the English-Japanese data was completed. Unlike with the English-Japanese data, changes were only made to the Chinese tagged sense if they can be linked with "=" to the English correspondent.

We also created new concepts for words that are lexicalized in the Chinese language but not represented in the CWN. This includes words like *桌 zhuō* "table" which is often used in combinations like *桌椅* zhuōyǐ "tables and chairs" and *桌上 zhuōshàng* "on the table". In the CWN, only *桌子* zhuōzi "table" is listed as a lemma for *table,* tagged to 04379243-n, defined as "a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs". As *桌 zhuō* "table" is well-

established, we deemed it appropriate to include the lemma into the synset. Hence, a concept is created for the word and tagged to the synset 04379243-n.

3.3.3 Concepts that could not be linked directly

There are many cases where words cannot be linked directly. Other than grammatical divergences, there were instances where the words belong to the same POS but due to the subtle, additional meaning in the words, the concept cannot be said to be exact translation of each other. This is illustrated in example (7).

(7) English: *"How on earth do you know that? I* ***asked.***

Japanese: 「いったい、どうして　その　こと　を？」と、　私

ittai,	*doushite*	*sono*	*koto*	*wo*	*to*	*watashi*
on earth	why	that	thing	ACC	QUOT	1SG

は	***聞き返す***。
wa	***kiki-kaesu***
NOM	to ask in return

"'Why on earth, (do you know) that thing?' I ask in return."

In the above, the corresponding word in the Japanese sentence has an additional meaning of "to ask in return" while in the original English sentence it is just "to ask". Although these two words are very similar in meaning, we were apprehensive to say that they are exact translations and therefore did not link them directly with "=".

Another example involves light verbs and their nouns, such as in the example (8) below where the Japanese multiword expression *身を震わせる mi wo furuwaseru* "to tremble convulsively, as from fear or excitement" is a translation equivalent of *gave a start.* However, as *gave* is a light verb with no significant semantic content of its own, we linked the Japanese verb to the noun *start* with the symbol "~" instead of linking the words directly.

(8) English: *I gave a **start** of **astonishment**.*

Japanese: *私 は **驚き** の あまり*

*watashi wa **odoroki** no amari*

1SG NOM astonishment POSS much

身-を-震わせ-た

mi-wo-furuwase-ta

body-ACC-to shake-PST

"I shook my body (due to) much astonishment."

3.3.4 Concepts that could not be linked

One of the most common cases where concepts could not be linked is when the words to be annotated are not present in the aligned corresponding sentence. There are several cases in the data where a sentence in English is translated into two or more sentences in the other language, and vice versa. This is exemplified by the sentence below, where the first clause of the sentence right until the word *suddenly* corresponds to (a) in the Japanese text and the remaining clause corresponds to (b).

(9) English: *"So, Watson," said he, suddenly, "you do not propose to invest in South African securities?"*

Japanese: (a)*「だから ワトソン一」と ホームズ が 突然*

dakara watoson- to ho-muzu ga totsuzen

so watson QUOT holmes NOM suddenly

口 を ひらく。

kuchi wo hiraku

mouth ACC to open

"So, Watson," said Holmes suddenly.

(b)*「君は、 南-アフリカ の 証券 へ の*

kimi wa minami-afurika no shouken he no

2SG NOM south-africa POSS securities DAT NMZ

投資 を 思いとどまった。」

toushi wo omoitodomat-ta

investment ACC to hold back-PST

"You held back investment in South African securities?"

Another instance where linking is impossible is where the translated sentence or phrase does not match the original text at all, as illustrated in the following example.

(10) English: *Holmes had been seated for some hours in silence with his long, thin back curved over* ***a chemical vessel in which he was brewing a particularly malodorous product.***

Japanese: *ホームズ*	*は*	*黙り込ん-だ*	*まま、*	*その*	*細く*	*長い*
ho-muzu	*wa*	*damarikon-da*	*mama*	*sono*	*hosoku*	*nagai*
holmes	NOM	in silence-PST	while	DET	thin	long
身体	*を*	*猫背*	*に*	*して、*		*何*
shintai	*wo*	*nekoze*	*ni*	*shi-te*		*nan*
body	ACC	cat's back	DAT	to make-CONJ		DET
時間	*も*	*化学*	*実験室*	*に*	*向かっていた。*	
jikan	*mo*	*kagaku*	*jikkenshitsu*	*ni*	*mukattei-ta*	
hour	INT	chemical laboratory		LOC	to face-PST	

"While Holmes was in silence, he made his thin long body into a cat's back, and faced the chemical laboratory for several hours."

In (10), the final clause (underlined) in the Japanese sentence was very loosely translated and roughly means "faced the chemical laboratory". However, in the original text, this meaning was not inherent at all. As a result, the Japanese words in the final clause are left without any suitable corresponding words in the original text. Similarly, the underlined phrase in the English sentence above does not correspond to any words in the translated Japanese text.

The following two examples illustrate another occurrence where linking is difficult. This usually occurs where the sentence as a whole is a translation equivalent but due to the choice of words and translation style, it is impossible to link the words at all.

(11) English: *"I can not say that she did not give me every chance of getting out of it if I wished to do so."*

Japanese:	*私*	*が*	*訊ね*	*さえ*	*すれ-ば、*	*包み隠さ-ず*
	watashi	*ga*	*tazune*	*sae*	*sure-ba*	*tsutsumikakusa-zu*
	1SG	NOM	to ask	if	to do-COND	to conceal-NEG
	言って	*くれ-た*	*と*	*思って-い-ます。*		
	itte	*kure-ta*	*to*	*omotte-i-masu*		
	to say	to give-PST	QUOT	to think-PTCP-POL		

"I think (that) if I ask, (she) will not conceal (anything) and tell me."

(12) English: *I am.*

Japanese:	*まったく*	*だ。*
	mattaku	*da*
	absolutely	COP

"Absolutely"

In (11), one can infer the translation equivalence just by looking at the sentences themselves. In example (12), one would have to look at the context in which the sentence appeared in order to understand how they are related. In both examples, although the meaning of the original English sentences is reflected in the Japanese translated sentences, there is no possible way of linking any words together.

In the following example, linking relationship between translation equivalents was difficult, as the concepts are not listed in the WordNets.

(13) English: *"I am sure that I shall say nothing of the kind."*

Japanese:	*「いやいや、そんな*		*こと*	*は*	*言わ-ん*	*よ」*
	iyaiya	*sonna*	*koto*	*wa*	*iwa-n*	*yo*
	by no means	that kind of	thing	SBJ	to say-NEG	yo

"no no, I will not say that kind of things"

In the Japanese text, the word いやいや *iyaiya* "by no means" in this context was interpreted as more colloquial in which a double negation was used, such as in "*no no,*

that is not it." As this sense was not included in the WordNets, we were unable to make any relationship links to the words.

CHAPTER 4

RESULTS

Having linked and tagged the relationships between words, we proceeded to count the number of each type of link. Table 5 and 6 below give a summary of the distinct concepts, synsets and linked synsets for each of the parallel corpus.

	English	Japanese
Distinct Concepts	6,587	5,119
Distinct Synsets	5,125	4,433
Distinct Linked Synsets	2,542	2,535

Table 5. Summary of English-Japanese corpus

In the English-Japanese corpus, 49.60% and 57.18% of the distinct synsets were linked in the English and Japanese texts respectively.

	English	Chinese
Distinct Concepts	6,587	5,143
Distinct Synsets	5,125	4,194
Distinct Linked Synsets	2,607	2,608

Table 6. Summary of English-Chinese corpus

In the English-Chinese corpus, 50.87% of the English distinct synsets were linked and 62.18% of the Chinese distinct synsets were linked. This is 1.27% and 5% more than the English-Japanese data. This could be due to the fact that the CWN contains more synsets than the JWN, so more synsets can be linked with those in the PWN.

Table 7 below shows the number of each type of link in the corpora. In both corpora, more than half of the linked concepts are tagged with "=", that is to say they are exact correspondents of each other. Relationships marked with "~" also take up a significant percentage in both corpora, 36.07% in the English-Japanese and 30.25% in the English-Chinese. Compared to the English-Chinese corpus, there are more instances where the target Japanese word is a direct hypernym of the English source

word. In both corpora, direct antonyms were used very sparingly. In total, there are 2745 linked concepts in the English-Japanese corpus and 2850 linked concepts in the English-Chinese corpus.

Link types / Relationship		English-Japanese	%	English-Chinese	%
=	Same synset	1,416	51.58	1,712	60.07
>	Direct hyponym	75	2.73	94	3.30
<	Direct hypernym	63	2.30	39	1.37
~	Similar/related	990	36.07	862	30.25
:	Translation equivalent	186	6.78	128	4.49
!	Direct antonym	1	0.04	2	0.07
#	Weak antonym	14	0.51	13	0.46
Total		2,745	100	2,850	100

Table 7. Number of each link types in the corpora

Among the concepts that are linked with "~", a number of them can be identified into certain categories such as pronominalisation, derivations etc. The table below shows the number of these identified types of relationships tagged with "~".

	English-Japanese	%	English-Chinese	%
Pronominalisation	0	0.00	7	0.81
Depronominalisation	86	8.69	22	2.55
Holonymy relationship	12	1.12	0	0.00
Derivation	56	5.66	30	3.48
2nd level hyponym of source	8	0.81	13	1.51
2nd level hypernym of source	10	1.01	18	2.09

Table 8. Identified types of relationships tagged with "~"

We found no instances of pronominalisation in the English-Japanese corpus. In the English-Chinese corpus, 0.81% of the concepts linked with "~" were found to be of pronominalisation. On the contrary, depronominalisation occurred more often in the English-Japanese corpus than the English-Chinese corpus. Derivations that could be

identified make up 5.66% and 3.48% of the concepts linked with "~" in the English-Japanese and English-Chinese corpus correspondingly. For example, the verbs *答える kotaeru* "to answer" in Japanese and *回答 huídá* "to answer" in Chinese are the derived forms of the noun *answer* in English.

Table 9 gives a summary of the linked concepts and their POS tags. We included all linked concepts except those linked with "=", since they are exact correspondents of the target words. In addition, we only count for those concepts where both source and target words had WordNet synsets.

POS Gap	**English-Japanese**	**%**	**English-Chinese**	**%**
Noun-Noun	386	32.80	347	31.98
Noun-Adjective	25	2.12	42	3.87
Noun-Verb	52	4.42	79	7.28
Noun-Adverb	11	0.93	24	2.21
Adjective-Adjective	72	6.12	103	9.49
Adjective-Noun	93	7.90	40	3.69
Adjective-Verb	33	2.80	27	2.49
Adjective-Adverb	28	2.38	33	3.04
Verb-Verb	298	25.32	274	25.25
Verb-Noun	87	7.39	16	1.47
Verb-Adjective	2	0.17	7	0.65
Verb-Adverb	7	0.59	10	0.92
Adverb-Adverb	40	3.40	60	5.53
Adverb-Noun	20	1.70	3	0.28
Adverb-Adjective	18	1.53	18	1.66
Adverb-Verb	5	0.42	2	0.18
Total	1177	100	1085	100

Table 9. Summary of linked concepts and their POS tags, excluding concepts linked with "="

From the table, 67.64% of the linked concepts in the English-Japanese corpus where translation shifts had occurred share the same POS tags in both source and target

language. In the English-Chinese corpus, 72.25% of the linked concepts share the same POS tags.

In the English-Japanese corpus, 7.90% syntactic mismatches occurred between English adjectives and Japanese nouns and 7.39% occurred between English verbs and Japanese nouns.

On the other hand, syntactic mismatches in the English-Chinese corpus are mainly observed between English nouns and Chinese verbs (7.28%). This is followed by mismatches between English nouns and Chinese adjectives at 3.87%.

In both corpora, mismatches between verb-adjective, verb-adverb, and adverb-verb do not occur frequently.

CHAPTER 5

GENERAL DISCUSSION

In this section, we will analyze and discuss some of the translation shifts that were found in the data.

In some of the sentences, we found holonym relationships in some of the sentences, such as *eye* and *瞳 hitomi* "pupil" in Japanese. *Eye* bears a two level holonym relationship with *pupil* in the WordNet system (Figure 2). In another example, *back* is a meronym of *torso* which is a meronym of *身体 shintai* "body" in Japanese. As there are exact correspondents of these in the Japanese language, *目 me* "eye" and *背中 senaka* "back", we attribute the semantic shifts to translating style.

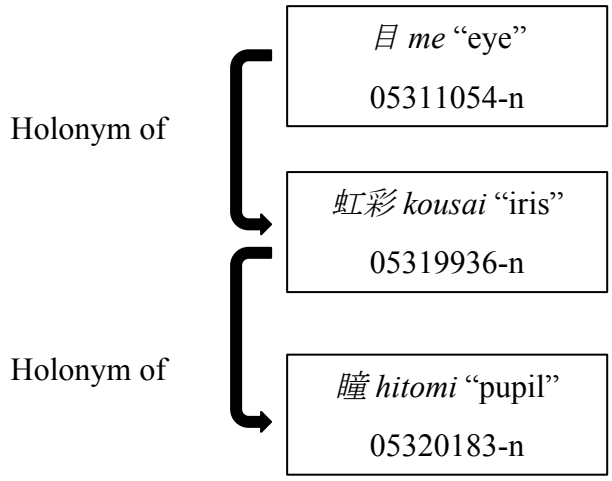

Figure 2. Holonymy relationship between *eye* and *pupil*

Lexical divergence was also found in translation equivalents of nouns derived from names of countries, as in *American* from *America.* In English, the word *American* can be interpreted as a person from America, or of things relating to the language or culture of America. On the other hand, in Japanese, if one were to refer to people, language or culture relating to a country, a suffix is often necessary. Thus, an *American* would be referred to as *アメリカ人 amerikajin* in Japanese, where the suffix *人 jin* indicates a person. Likewise, the concept *American car* would be translated to Japanese as *アメリカの車 amerika no kuruma* where *の no* is a particle

indicating POSSESIVE. These words were tokenized separately, for example and 人 *jin* "person" in Japanese. Therefore, when linking the words, we could not treat them as exact correspondences but linked with "~" instead, and these contributed to the mismatches. In addition, アメリカ *amerika* "America" is a holonym of the word *American* which bears the sense "native or inhabitant of the United States".

In Chinese, the words that represent the names of countries can directly act as a noun modifier. For example, the word 美国 *měiguó* "America" in 美国人 *měiguó rén* "an American" and 美国车 *měiguó chē* "American car" has the sense "of things relating to the person, language or culture of America". It is also common in Chinese to use the particle 的 *de* to describe the preceding noun, such as in 美国的车 *měiguó de chē* "American car".

Translation shifts may occur when a word has gone through a semantic change in the source language but not the target language. We found an example in the English-Japanese corpus. The word *kill* underwent meiosis, a change from stronger to weaker meaning. In the PWN, there is a sense for *kill* with a synset ID 02198819-v, which is defined as "be the source of great pain for". An example of such a use is *"These new heels are killing me."* However, the translation equivalent for *kill* in the Japanese translated text, 殺す *korosu* "to kill", does not have that synset nor a similar sense. The synset tagged to the Japanese word is 01323958-v, defined as "cause to die". We could only link the words as of having related meaning and therefore, a shift has occurred.

In the corpora, the original English text used *eye* throughout regardless of whether it was with reference to the organ of sight or the expression in one's eye. However, the Chinese translator made a distinction between them and translation shifts occurred as a result. The examples below illustrate three distinctions that were made in the Chinese text.

(14) English: *I could see by his **eyes***

Chinese:	*从*	*他*	***眼神***	*中*	*可以*	*看出*
	cóng	*tā*	***yǎnshén***	*zhōng*	*kěyǐ*	*kànchū*
	from	3SG	eye-expression	in	can	see

"can see from the expression in his eyes"

(15) English: *earnest blue **eyes***

Chinese:	*诚实*	*的*	*蓝*	***眼睛***
	chéngshí	*de*	*lán*	***yǎnjīng***
	honest	de	blue	eyes

"honest blue eyes"

(16) English: *his **eye** rested upon the paper*

Chinese:	***目光***	*落在*	*那*	*张*	*纸条*	*上*
	mùguāng	*luòzài*	*nà*	*zhāng*	*zhǐtiáo*	*shàng*
	gaze	to fall on	that	CL	paper	on

"(his) gaze fall on that piece of paper"

The translation shifts observed here are attributed to the translator's effort to be more specific.

As illustrated in 3.3.2 example (7), divergences also occur when a word in one language contains more information than a word in the other language. Combining two verbs forms many words in Japanese, for example *食べる taberu* "to eat" and *みる miru* "to see" are combined to form *食べてみる tabetemiru* "to try eating and see". In the corpora, compound verbs like these are used rather frequently in the Japanese text, such as in example (7), which reflects the agglutinative nature of Japanese language.

In Chinese, two verbs often concatenate and form a single, lexicalized verb. In the following examples, the verbs are composed of two single verbs.

(17) English: *he will not try to **escape***

Chinese:	*他*	*不*	*会*	***逃跑***	*的*
	tā	*bù*	*huì*	***táopǎo***	*de*
	3SG	NEG	will	to escape-to run	de

"he will not escape (and run)"

(18) English: *I held my gun to **scare** him **off***

Chinese:	*我*	*举起*	*枪*	*想把*	*他*	***吓跑***
	wǒ	*jǔqǐ*	*qiāng*	*xiǎngbǎ*	*tā*	***xiàpǎo***
	1SG	to hold up	gun	to want to	3SG	to scare-to run

"I help up (the) gun (to want to) scare him (and make him run)"

In (17) the verb *逃跑 táopǎo* "to escape and run" consists of two verbs. The first verb *逃 táo* already contains the meaning of "to escape" on its own, and the second verb *跑 pǎo* is there to add the further information of "to run", perhaps to show the urgency of the escape. Both verbs take the subject as arguments. In this context, *escape* was tagged to the synset 02075049-v, defined as "flee". The Chinese counterpart also shares the same synset and thus the words were taken to be exact translations of each other.

On the contrary, in (18), the first verb *吓 xià* "to scare" takes both the subject and object as arguments while the second verb *跑 pǎo* "to run" only takes only the object as an argument. The second verb shows an action that was caused by the action indicated by the first verb. The word *scare* and the multiword expression *scare off* were both tagged to the synset 017857480-v, which is defined as "cause to lose courage". However, the target word *吓跑 xiàpǎo*, which a native speaker might even translate as "to scare off", was not represented in the CWN. Nonetheless, we linked *scare* and *吓跑 xiàpǎo* "to scare someone off" with "~" as the Chinese word evidently contains more information than the English word. We also made a comment that a synset for *吓跑 xiàpǎo* "to scare someone off" should be created in the CWN.

In addition, a verb and an adjective can also combine to form a new Chinese verb as long as the resulting state described by the adjective is possible (Palmer & Wu, 1995). The function of verb-adjective words is usually to express a change-of-state event such as *breaking*. The adjectival component of the verb usually conveys the resulting state more explicitly than it is normally done with English. The productivity of such verbs in both Japanese and Chinese may contribute to translation divergences.

Adjectives that are made up of two separate adjectives are also common in the Chinese language. For example, in the following phrase, the English words *long* and *thin* were translated into a single Chinese word *瘦长 shòucháng* "lanky".

(19) English: *his **long, thin** back curved over*

Japanese:	*他*	*弯*	*着*	***瘦长***	*的*	*身子*
	tā	*wān*	*zhe*	***shòucháng***	*de*	*shēnzi*
	3SG	to curve	PROG	lanky	de	body

"he curved (his) lanky body"

In (19), the word *瘦长 shòucháng* "lanky" corresponds to *long* and *thin* in the original English text. In annotating the relationship, we had to link the words separately. Since *瘦长 shòucháng* "lanky" contains more information than *long* and *thin* individually, we considered it a shift that had occurred during translation. In fact, *瘦长 shòucháng* "lanky" is made up of two words *瘦 shòu* and *长 cháng* which means "thin" and "long" respectively. The translator could have chosen to make an exact translation using a conjunction to join the two concepts. However, as the word *瘦长 shòucháng* "lanky" is lexicalized, this may have influenced the decision to use the lexicalized term instead.

The verb *する suru* "to do" in Japanese can be used to convey different meanings depending on the words they are combined with. One of its main functions is to change a word into a verb. Verbs are closed classes in Japanese and they do not readily add new members (Bloch, 1946). New and borrowed verbs are conjugated periphrastically as nouns + *する suru* "to do", such as *結婚する kekkonsuru* "to

marry" and *約束する yakusokusuru* "to promise". Since Japanese has a large vocabulary of Chinese loanwords (Shibatani, 1990), it can be expected that many verbs in Japanese are formed using noun + *する suru* "to do". This nature of the Japanese verb system contributes to the number of syntactic shifts we found in the data where verbs in English were linked to nouns in Japanese.

There are also cases where an established multiword expression in English could not be translated directly into the target language. Examples (20) and (21) below illustrate the two instances where the multiword expression, *to the bottom* was used in the text.

(20) English: *get **to the bottom** of it*

Japanese: *暴く こと が できます*

abaku koto ga deki-masu

to expose NMLZ NOM to be able to-POL

"able to expose"

Chinese: ***彻底*** *弄* *清楚*

chèdǐ *nòng* *qīngchǔ*

completely to make clear

"to make clear completely"

(21) English: *sift the matter **to the bottom***

Japanese: *最後 まで 調べ-たい*

saigo made shirabe-tai

end until to investigate-to want

"want to investigate until the end"

Chinese: ***彻底*** *弄* *清楚*

chèdǐ *nòng* *qīngchǔ*

completely to make clear

"to make clear completely"

From the examples, the Chinese translator was consistent with the translation of *to the bottom* in both instances, suggesting that the expression *彻底弄清楚 chèdǐ nòng qīngchǔ* "to make clear completely" behaves in the same way as the English

multiword expression and is an established multiword expression. The word *彻底 chèdǐ* "completely" in particular seems to incorporate the meaning of *to the bottom* on its own. In the Japanese text however, the translator used two different approaches to translate the concept of *to the bottom*. The lexical divergence that presents itself here could be due to the absence of a Japanese equivalent to the multiword expression in English.

The translator of the Chinese text also used idiomatic expressions to express certain concepts. These idiomatic expressions transmit the same idea as the original text but they occasionally function as different POS, or even contain much more information that is not reflected in the English counterpart.

One example is *作恶多端 zuòèduōduān* "to have done many kinds of evil". This idiomatic expression was translated from the English word *evil*. In the context, *evil* was used to refer to the acts that had been done. The Chinese idiomatic expression however, was used to describe the agents who did the act itself. Hence, a syntactic shift occurred here from a noun in the English original text to an adjective in the Chinese translated text.

We found a lot more depronominalisation going on between the English-Japanese corpora than the English-Chinese corpora. Although both Japanese and Chinese are pro-drop languages, pronouns are used much less frequently in Japanese than in other languages (Maynard, 1990). In addition, Japanese speakers prefer to refer to another person by title, function or by that person's name. In Chinese however, it is common to use pronouns to refer to another person. The occurrence of depronominalisation and also pronominalisation contributed towards a higher count for relationships tagged with "~".

(22) English: *Oh, that's **your** idea!*

Japanese: *ほう、そう　考える　　か　　ね*

hou　sou　kangaeru　ka　ne

oh　that　to think　QUOT　eh

"Oh, (you) think that eh?"

Chinese: *噢，　那　是　**你　的**　想法*

*o　nà　shì　**nǐ　de**　xiǎngfǎ*

oh　that　is　2SG　POSS　idea

"Oh, that is your idea"

(23) English: ***She** shot **him** and then herself*

Japanese: ***奥-さん**　が　**旦那-さん**　を　撃って、*

***oku-san**　ga　**danna-san**　wo　utte*

wife-HON　NOM　husband-HON　ACC　to shoot at-CONJ

それから　自分　も　撃った

sorekara　jibun　mo　utta

and then　self　too　to shoot at-PST

"(the) wife shot (the) husband and then shot (her)self too"

Chinese: ***她**　拿　枪　先　打　**丈夫，***

tā**　ná　qiāng　xiān　dǎ　**zhàngfū,

3SG　to take gun　first　to shoot　husband

然后　打　自己

ránhòu　dǎ　zìjǐ

and then　to shoot　self

"She took the gun to first shoot (her) husband, and then shot (her)self"

In (22), the pronoun *your* was dropped in the Japanese translation but kept in the Chinese translation with a corresponding pronoun and POSSESIVE marker, *你的 nǐde* "yours". In (23), depronominalisation occurred in the Japanese text as *she* and *him* were translated to the common nouns *奥さん okusan* "wife" and *旦那さん dannasan* "husband" respectively. The suffix *さん san* acts as a honorific marker. In the

Chinese counterpart, depronominalisation only affected the pronoun *him,* where it was translated to 丈夫 *zhàngfū* “husband”.

Due to subtle language differences, word-for-word translation, if possible, may sometimes turn out unnatural. A good example is the use of pronouns in Japanese. The Japanese pronoun for second person singular, あなた *anata* “you” is commonly used by women to address their husbands or lovers. Likewise, the third person singular pronouns 彼 *kare* “he” and 彼女 *kanojo* “she” are commonly used by Japanese speakers to mean “boyfriend” and “girlfriend” respectively. Therefore, if the translator were to translate pronouns exactly as they were, it may appear unnatural to Japanese speakers. In addition, unlike English and Chinese pronouns, Japanese pronouns have many forms for each person, which seem to be correlated with gender, dialectal differences and so on (Shibatani, 1990). In addition, the reference to persons in Japanese is also subjected to sociocultural factors such as status differences. For example, 俺 *ore* “I” is an informal term frequently used by men, and can be seen as rude depending on the situation. It also emphasizes one’s status when used with those who are younger or who have lower social status.

Another instance where word-for-word translation is awkward is the translation of colloquialisms that may appear in dialogues. For example, according to the online Macmillan Dictionary, the word *sick* can be used to mean “very impressive, attractive, enjoyable, etc.” such as in *“The dress is sick! You look amazing!”* A word-for-word translation of this into Japanese and Chinese would give the following

(24) Japanese: *ドレス* *は* *病気* *です*
doresu *wa* *byouki* *desu*
dress NOM sick COP
“(the) dress is sick”

Chinese: *这* *件* *衣服* *是* *生病* *了*
zhè *jiàn* *yīfú* *shì* *shēngbìng* *le*
DET CL dress to be sick le
“The dress is sick”

In (24), both the Japanese and Chinese translated sentences mean *"The dress is sick."* However, the colloquial meaning of *sick* does not exist in the Japanese and Chinese languages. Therefore, the word-for-word translation in (24) would give a sentence that is grammatically correct but makes no sense semantically to a native speaker of Japanese or Chinese.

These language differences prove to be a challenge for machine translation systems. A machine translation system can correctly translate a text word-for-word from English to Japanese or English to Chinese. However, certain connotations may end up being lost in translation. While the translated text may be grammatically sound, the resulting meaning may be unnatural or makes no sense to the user. Likewise, translating words from Japanese or Chinese to English will also result in covert information being lost.

We also observed interesting variations in the English-Chinese corpus where the translator used a different word to represent the same concept in different settings. For example, for the word *doctor*, the translator consistently used 医生 *yīshēng* "doctor" in the narratives and 大夫 *dàifū* "doctor" in dialogues between characters. Such information will be useful for second language learners as they can see how one concept can be represented in various ways depending on the context.

Many of the translation divergences discussed here are due to language differences. In both corpora, most of these translation shifts are semantic mismatches. Nevertheless, translating styles can often be the cause of translation shifts. This data not only allows users to see how a concept gets translated on a word level; it also allows users to see how the underlying message of a sentence can be conveyed in a different way in the target language. This data can also be used to train tagging systems to automatically tag for word-sense and link concepts based on the features in this study. This will allow researchers to save on time and human resources. Furthermore, to improve the efficiency of machine translation, the data can be used to train machine translation systems to identify word environment and context in addition to WSD. With increased feedbacks from such data, machine translation systems may overcome the challenge of lexical divergences and produce more human-like translations.

During the data tagging process, we came across a lot of concepts without synsets in the WordNets. This lack of representation resulted in a lot of concepts that could not be linked. Many of these concepts include idiomatic expressions and also lexicalized multiword expressions. One of the challenges that WordNets has to face would be how to best represent these expressions in the systems and link them to the existing structures.

Also, one of the issues with the JWN is that certain lemmas are only represented in *kanji* characters in the synsets. As some concepts in the corpus were represented in *hiragana,* the senses for these concepts could not be found automatically. We noted these concepts and commented that the lemmas should be added into the existing synsets. In addition, the CWN also contains several POS errors. As we have noted these errors, we hope to bring it to the attention of the maintainers of the CWN in order to correct these and hence, improve the system.

CHAPTER 6

CONCLUSION

In this study, we have used a quantitative and qualitative approach in studying translation shifts in parallel corpora in English, Japanese and Chinese. Our results show that a significant of divergences has occurred during the translation of the texts. We have also attempted to describe some of these translation divergences. Although it is challenging to establish the exact number of shifts due to language differences or translating style, we gauge that more than half of the translation shifts observed in this study are definitely due to the differences in target language systems. Future studies can work with texts from different genres such as news articles or contemporary novels. Preliminary work has been done on another text – The Cathedral and the Bazaar. We predict that the amount and types of translation shift will differ depending on the genre of the text.

We would also like to point out that the Japanese and Chinese WordNet are not yet as fully developed as the Princeton WordNet, which may have compromised the results, as many things could not be linked. Through this study, we hope to have been able to help improve the WordNets in terms of coverage and also in correcting errors.

This parallel tri-text is the first such data annotated at this level of granularity for English, Japanese and Chinese. The data from this study will be released so as to allow other scholars to work on it. It can also help second language learners of Japanese and Chinese to understand that a concept in English can be represented in various ways depending on the context. In addition, information from the data can also be used to train machine translation systems in order to produce a more human-like translation.

References

Ahrenberg, L. (2007). Lines: An English-Swedish Parallel Treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA '07), pp. 270–274.* Tartu, Estonia.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers, (Ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honor of Juan C. Sager*, pp. 175-186. John Benjamins

Baker, M., Fillmore, C.J., & Lowe, J.B. (1998). The Berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – Volume 1*, pp. 86-90. Montreal, Canada

Bentivogli, L., & Pianta, E. (2000). *Looking for lexical gaps.* In *Proceedings of the 9th EURALEX International Congress,* pp. 8-12. Stuttgart, Germany.

Bentivogli, L., Forner, P. & Pianta, E. (2004). Evaluating cross-language annotation transfer in the MultiSemCor corpus. In *Proceedings of the 20th International Conference on Computational Linguistics,* pp. 364-370. Geneva.

Bloch, B. (1946). Studies in Colloquial Japanese II: Syntax. *Language, 22*(3)*,* pp 200-248.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009). Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources,* pp. 1-8. Singapore.

Brown, V., Mendes, E., & Natali, G. (1995). *False friends and bugs and bugbears*. Zanichelli, Bologna.

Catford, J.C. (1965). *A Linguistic Theory of Translation: An Essay in Applied Linguistics.* (Volume 8). Oxford University Press.

Čulo, O., Hansen-Schirra, S., Neumann, S., & Vela, M. (2008). Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus. In *Proceedings of "Building and Using Comparable Corpora", LREC 2008 Workshop, Volume 31,* pp. 47-51. Marrakesh, Morocco.

Cyrus, L. (2006). Building a resource for studying translation shifts. In *Proceedings of LREC 2006-Second International Conference on Language Resources and Evaluation,* Genoa, Italy.

Denny, J. P. (1986). The semantic role of noun classifiers. In Craig, C. (Ed.), *Noun classes and categorization*, pp. 297–308. Philadelphia: John Benjamins

Dorr, B.J. (1993). *Machine Translation: A View from the Lexicon.* MIT Press, Cambridge, Massachusetts.

Fellbaum, C, editor. (1998). *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusetts.

Fellbaum, C. (2005). WordNet and wordnets. In Brown, K. et al. (eds), *Encyclopedia of Language and Linguistics* (2nd edition.), pp. 665-670. Oxford: Elsevier

Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics, 24*(1), pp. 1-40.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. & Kanzaki, K. (2008). Development of the Japanese WordNet. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008),* Marrakesh, Morocco.

Koehn, P. (2002). Europarl: A multilingual corpus for evaluation of machine translation. In *MT Summit, Volume 5.*

Liddy, E.D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science*, (2nd edition). New York: Marcel Decker, Inc.

Lo Cascio, V., Boraschi, P., & Corda, A. (1995). Correspondence between senses and translation equivalents: automatic reversal of a bilingual dictionary. In *Translation and Meaning. Part 3 of the Proceedings of the 2nd International Maastricht-Lodz Duo Colloquium on 'Translation and Meaning',* pp. 221-231. Maastricht.

Marello, C. (1989). *Dizionari bilingui: con schede sui dizionari italiani per francese, inglese, spagnolo, tedesco* (Volume 6). Zanichelli, Bologna.

Maynard, S.K. (1993). *An Introduction to Japanese Grammar and Communication Strategies,* (4th edition). The Japan Times.

McCarthy, D. (2009). Word sense disambiguation: An overview. *Language and Linguistics Compass, 3*(2), pp. 537-558.

Mok, S.W.H. (2012). Issues and guidelines for tagging a multitext with WordNet. NICT Internship report (Summer 2012).

Padó, S. & Erk, K. (2010). Translation shifts and frame-semantic mismatches: A corpus analysis. *International Journal of Corpus Linguistics. To appear.*

Palmer, M. and Wu, Z. (1995). Verb Semantics for English-Chinese Translation. *Machine Translation*, *9*(4).

Tan, L & Bond, F. (2011). Building and annotating the linguistically diverse NTU-MC (NTU multilingual corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pp. 367–376. Singapore. ISBN 978-4-905166-02-3.

Shibatani, M. (1990). *The Languages of Japan.* Cambridge University Press. ISBN 0-521-36918-5.

Teich, E. (2003). *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts* (Volume 5). Walter de Gryuter.

Vinay, J., & Darbelnet, J. (1995). *Comparative Stylistics of French and English: A Methodology for Translation.* (J.C. Sager & M. Hamel, Trans.). Amsterdam and Philadelphia, PA: John Benjamins. (Original work publish 1958).

Volk, M., Göhring, A., Marek, T., and Samuelsson, Y. (2010). SMULTRON (Version 3.0) – The Stockholm MULtilingual parallel TReebank. *An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments.*

Wilks, Y., & Stevenson, M. (1996). The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?. Technical Report CS-96-05, University of Sheffield, Sheffield, U.K.

Xu, R., Gao, Z., Qu, Y., & Huang Z. (2008). An integrated approach for automatic construction of bilingual Chinese-English WordNet. *The Semantic Web,* pp. 302-314.

Sick. (n.d.). In *Macmillan Dictionary online.* Retrieved from http://www.macmillandictionary.com/dictionary/british/sick