The Nature of Cross-Lingual Lexical Semantic Relations: A Preliminary Study Based on English-Chinese Translation Equivalents

Chu-Ren Huang
Institute of Linguistics, Academia Sinica
Nankang, Taipei, Taiwan 115
ichuren@sinica.edu.tw

Wan-Ying Lin
Institute of Linguistics, Academia Sinica
Nankang, Taipei, Taiwan 115
iwaiin@gate.sinica.edu.tw

Jia-Fei Hong
Institute of Linguistics, Academia Sinica
Nankang, Taipei, Taiwan 115
ijiafei@gate.sinica.edu.tw

I-Li Su
Institute of Linguistics, Academia Sinica
Nankang, Taipei, Taiwan 115
isu@gate.sinica.edu.tw

Abstract

In this paper, we propose a new approach to comparative lexical semantics. In particular, a wordnet-like framework is adopted to study the nature of cross-lingual lexical semantic relations. The synsets of an existing monolingual wordnet are often aligned with their translation equivalents in a target languages in order to bootstrap a bilingual wordnet. Previous studies adopting this approach include the Spanish WordNet (SpWN, Atserias et al., 1997) and MultiWordNet (MWN, Pianta, et al., 2002). Such studies brought to attention the importance of cross-lingual lexical semantic relations between two translation equivalents. In this paper, we examine and analyze the contrast and the cross-lingual semantic relations between the English WN synsets, and their Chinese translation equivalents. Generalizations are made based on the distribution of the part-of-speech, semantic relations and concepts in terms of SUMO ontology. Our account sheds the first light towards the nature of conceptual basis for non-synonymous translation, as well as for bilingual wordnet-mapping.

1 Introduction

Translating knowledge from language A to language B is a very direct way to analyze the similarities and differences between two languages. The ideal situation for translating lexicons in language A, is to find the synonyms in language B; however, this does not often turn out to be the case. In fact, there exist some problems about the cross-lingual semantic relations between the lexicons and their translation equivalents (TEs) due to the influences from the context, sentence patterns and so on. Nevertheless, finding the TEs for each lexicon is still a start for digging the varieties of languages.

In this paper, the source language (SL) and the target language (TL) are English and Chinese, and this experiment is limited to examine five semantic relations: synonym, hypernymy, hyponymy, holonym and meronym. There are two main approaches proposed in order to examine and analyze the statistics of the cross-lingual semantic relations of lexicons and their TEs. The first approach is to examine those semantic relations based on their syntactic categories. That is, we use the syntactic categories, nouns (N), verbs (V), adjectives (A) and adverbs (R), as the basis for checking the statistics of the distribution for all semantic relations. The second approach is to use the semantic relations as the basis and check the distribution of each syntactic category. Finally, we hope to use the Suggested Upper Merged Ontology (SUMO, Niles and Pease, 2003) to check the regulation to explain the concept discrepancies between English and Chinese.

The rest of this paper is organized as follows. This paper includes the following sections: in section 2, we briefly introduce the existing resources that are required in our experiment. The statistics after adopting these two approaches are evaluated and analyzed in section 3. The conclusion of this experiment is presented in section 4.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 181–190.

2 Required Resources: WN, ECTED and SUMO

We need three required resources to do our work: WordNet (WN), the English-Chinese Translation Equivalents Databases (ECTED), and Suggested Upper Merged Ontology (SUMO). Princeton WordNet (WN) is a famous resource for the applications in Natural Language Processing (NLP). Many multilingual wordnets, such as the EuroWordNet (EWN, Vossen 1998), are constructed inspiring by the alignment with WN. Undoubtedly, WN provides rich lexical knowledge of English language. Synsets (a set of words having the same part-of-speech and sharing the same unique sense/concept) are the basic units used in WN to organize and represent the lexicon conceptually. They are classified into four main syntactic categories: nouns, verbs, adjectives and adverbs. The semantic relations of those synsets are clearly indicated in WN. Therefore, WN can be regard as a Word-Network because it expresses not only the lexical knowledge but also the semantic relations and the conceptual information of words.

One of the remarkable applications in using WN is to extract the multilingual wordnets by strictly aligning the English synsets with their translation equivalents (TEs) in different languages. The existing wordnets, such as MultiWordNet (Piana, et al., 2002) and Spanish WordNet (Atserias et al., 1997), are all bootstrapped from the WN based on such strategy. It seems that the strategy is really doable, so we attempt to construct a Chinese wordnet through WN and the TEs of the synsets. Thus, we build a database, the English-Chinese Translation Equivalents Databases (ECTED), to provide the Chinese translation equivalents for each WN English synset. The WN version we adopt is WN1.6, because this version is currently used by most applications. The first step of building ECTED is to use several on-line English-Chinese or Chinese-English bilingual resources to search all the possible candidates of Chinese TEs for each WN synset. Afterward, a group of translators choose or create the three best ones and determine the cross-lingual semantic relations between each English synset and its TEs. There are eleven selections of the semantic relations provided in the database for translators to make the choices. The details for the eleven selections are shown in Table 1.

There are 99,642 English synsets in WN 1.6 and those synsets can be expanded to 157,507 English lemma tokens. However, if the statistics are expanded by the cross-lingual semantic relations, the English lemma tokens will be reduced to 117,274 due to the way we count them. For example, if three TEs of a synset are all marked [1-0] correct (synonym), we only counted it once; however, if two of the TEs are marked "[8-0] hypernym" and one is marked "[6-0] not lexicalized," the English lemma token would be counted twice.

In order to have the more precise data, we are still proceeding to the re-examination for all the corss-lingual semantic relations in ECTED. Currently, we have already completed the checking of [1-0] correct (synonym), [8-0] hypernym, [9-0] hyponym, [10-0] meronym and [11-0] holonym. The sum of those semantic relations is about 50 % of all the data.

Marks	Explanation	Examples
[1-0] correct (synonym)	English Synset and the TE are synonyms	00934421A (fat) = 肥胖的;圓胖的
[2-0] incorrect	English Synset and TE are not synonyms and have no relevant semantic relations	00012704N(article) = 一件; 一個
[3-0] others	English Synset and TE are not synonyms but they have certain semantic relations	02609065N(drug) = 麻醉藥
[4-0] debatable	Difficult to determine the semantic relations	00103419N(persuasion) = 信仰; 信念; 信條
[5-0] phrases	English Synset and the TE are synonyms but the TE is a phrase not a word	07110073N(best) = 最佳
[6-0] not	English	07334873N(immune)

lexicalized	Synset and the TE are synonyms but the TE is not lexicalized in Chinese	= 具免疫力的人; 免疫者
[7-0] near-synonym	TE is a near-synonym of English Synset	00072344V(piffle) = 胡扯
[8-0] hypernym	TE is a hypernym of English Synset	03461955N(table) = 桌子; 桌
[9-0] hyponym	TE is the hyponym of English Synset	00133939N(escape) = 逃生出口; 避難裝置
[10-0] meronym	TE is a meronym of English Synset	03225469N(refrigerator) = 冷藏庫; 冷凍庫
[11-0] Holonym	TE is a holonym of English Synset	04291831N(cranium) = 頭顱

Table 1.Th Evaluative Tags of the Semantic Relations in ECTED

Suggested Upper Merged Ontology (SUMO, Niles and Pease, 2003) is an ontology system that is constructed by the IEEE Standard Upper Ontology Working Group.[1] The main task of SUMO is to conceptualize most general concepts or common-senses in order to be used by different domain ontologies. Basically, SUMO contains 11 categories which cover 3912 conceptual nodes. The concepts and the axioms are represented clearly in six different languages.

3 Descriptions and Analyses

3.1 Examine the Data Through POS

In order to clearly examine the generation for non-synonyms, in our experiment, the syntactic categories are restricted to nouns and verbs only and the cross-lingual semantic relations are restricted to the relation of non-synonyms, such as hypernym, hyponym, meronym and holonym. As shown in Table 2, hypernym and hyponym get the higher percentages rather than meronym and holonym no matter in what categories. Such observation indicates that under the situation of non-synonyms, the first instincts that humans would use to decide the semantic relations are hypernym or hyponym. Besides, according to the Table 2, the distribution of nouns appears in all the semantic relations, but the distribution of verbs is only revealed in hypernym and hyponym. This proves that the semantic relations of meronym and holonym do not exist in verbs; and this also proves that verbs do have meaning variations and such variations especially appear in hypernym.

POS	N	V	Total
[8-0] Hypernym	906	132	1038
	66.2%	80.0%	67.7%
[9-0] Hyponym	356	33	389
	26.0%	20.0%	25.4%
[10-0] Meronym	61	0	61
	4.5%	0%	4.0%
[11-0] Holonym	46	0	46
	3.4%	0%	3.0%
Total	1369	165	1534
	100%	100%	100%

Table 2.The Statistics of the English Lemma Tokens

3.2 Examine the Data Through SUMO Concepts

After the above-mentioned statistics of POS is shown in section 3.1, we would like to examine those statistics again via the distribution of concepts, so we used SUMO to complete this examination. WN 1.6 is mapped to SUMO and the WN / SUMO mapping can function as natural language index to the concepts in the ontology (Niles and Pease, 2003). We incorporated WN synsets with SUMO concepts in order to classify the related nodes to groups. Nouns and verbs would be the key categories in the discussion here. Through comparing with English synsets and Chinese translation equivalents, we have different semantic relations. We try to find out how these semantic relations are combined with SUMO concepts. Also, we can see which concepts are mainly distributed in each semantic relation. After connecting the English synsets to SUMO upper node concepts, we try to figure out the percentages of the semantic relations and SUMO concepts in our examination. When we understand more about the WN / SUMO

[1] SUMO now is maintained by the IEEE Technical Editor, Mr. Adam Pease. Please refer to SUMO's website: http://www.ontologyportal.org/

mappings, we are able to predict the cross-lingual lexical semantic relations between the synsets and their corresponding translation equivalents.

In SUMO, we noticed that some WN synsets are mapped to more than one SUMO concept. Such mappings are considered as the multiple inheritances. However, such mappings will cause the difficulty in choosing the proper translation equivalents in Chinese due to the language discrepancies. The total numbers of such SUMO concepts are 291 (for 140 synsets) in [1-0-V] correct (synonym), 52 (for 26 synsets) in [8-0-N] hypernym, and 18 (for 9 synsets) in [9-0-N] hyponym. In order to simplify such difficulty, we decided to ask some translators to recheck and choose the main concept from those SUMO concepts for each synset. Thus, the totals of the one-to-one mapping relation are 6988 in [1-0-V] correct (synonym), 906 in [8-0-N] hypernym, and 356 in [9-0-N] hyponym. Please notice that there are two verbs in [1-0-V] correct (synonym) do not map to any SUMO concepts, as shown in Table 3.[2] We do not count these two verbs in, so the total of verbs in [1-0-V] correct (synonym) is now revised to 6986.

ID	01503219V	00767910V
Mean	make bitter	force or drive back
Synset	bitter	repel, repulse, fight_off, drive_off, rebuff, drive_back,
Translation	[1-0]苦澀	[1-0]擊退 [1-0]逐退
SUMO Concept	0	0

Table 3. [1-0-V] Examples without SUMO Concepts

Firstly, we would like to observe the term frequencies and the percentages of English synsets in different semantic relations with the distribution of concepts in SUMO. The purpose for doing this will help us to get the idea that what sorts of terms are easier and clear for humans to decide the semantic relations between the English synsets and translation equivalents. According to our observation, we notice that most of the English nouns can get their synonyms in Chinese and the percentage for this relation is over 55 %. However, this phenomenon does not apply to the distribution of verbs. The percentages of synonyms and non-synonyms in verbs distribution are almost the same. Such phenomenon is very interesting and we cannot help but wonder why such phenomenon appears in the distribution of verbs. According to Table 4, the concept, *Communication*, gets the highest percentage in the relation of synonyms. The distribution reveals the accuracy for finding the translation equivalents in Chinese of the speech act verbs in English that is fairly high. As for *Process*, *Motion*, *Intentional Psychological Process* and *Intentional Process* are also important concepts in synonymous relation, and the percentages of these concepts are shown in Table4. In addition, *Subjective Assessment Attribute* and *Social Interaction* concepts are usually connected with synonyms.

Table 5 indicates the distribution of verbs in hypernym. The SUMO concepts, *Motion and Removing*, both get the highest percentage in hypernym. However, in hyponym, as shown in the Table 6, there are three concepts found in hyponym and *Communication* is the concept that gets the highest percentage. Besides, *Motion* and *Body Motion* can be regarded as a kind of motion. After comparing Table 5 and 6, we notice that the verbs in *Communication* and *Motion* are more easily marked by both the semantic relations, hypernym and hyponym. The verbs in *Intentional Process* occupy as the fifth main concept in synonyms, while it is also important SUMO concept in hypernym. However, such SUMO concept is not one of the major conceptual distributions shown in hyponym. Interestingly, if the synset connecting with *Body Motion* concepts are in non-synonyms, its semantic relation is often hyponym.

SUMO Concept	**Term Frequency**	**(%)**
Communication	457	6.54%
Process	397	5.68%
Motion	380	5.44%

[2] In the SUMO+MILO mappings, the first synset in Table3 is probably mapped to *Taste Attribute* concept, while the second synset is mapped to *Contest* concept. We ignored to deal with these two synsets into our analysis. However, we will try to figure out why there are concept gaps in connecting with SUMO after we re-examine the semantic relations in ECTED.

Intentional Psychological Process	376	5.38%
Intentional Process	375	5.37%

Table 4. The Major Conceptual Distribution in [1-0-V] Correct (synonym) (a total corpus frequency of 6986)

SUMO Concept	Term Frequency	(%)
Motion	12	9.09%
Removing	12	9.09%
Intentional Process	5	3.79%
Communication	5	3.79%
Content Development	5	3.79%
Cutting	5	3.79%
Attaching	4	3.03%
Putting	4	3.03%
Decreasing	3	2.27%
Touching	3	2.27%
Impacting	3	2.27%
Walking	3	2.27%

Table 5. The Major Conceptual Distribution in [8-0-V] Hypernym (a total corpus frequency of 132)

SUMO Concept	Term Frequency	(%)
Communication	4	12.12%
Motion	2	6.06%
Body Motion	2	6.06%

Table 6. The Major Conceptual Distribution in [9-0-V] Hyponym (a total corpus frequency of 33)

As we mentioned above, the percentage of marking synonyms in nouns is over 55 % and according to Table 2, the distributions of nouns reveal in hypernym, hyponym, meronym and holonym. As the same situation happening in the distribution of verbs, hypernym, is mainly used by humans in order to indicate the relation between the English synsets of nouns and their translation. From Table 7, it shows that the conceptual type, *Flowering Plant*, gets the highest percentage in hypernym. Such result is expectable because not all of the flowering plants can get their corresponding names in Chinese. Sometimes, the translators use the hypernyms to be the translation for those terms. In Table 8, we examine the distribution of the SUMO concepts in hyponym. It is the same as in hypernym – *Flowering Plant* is still the highest percentage in hyponym. *Device*, *Artifact*, and *Subjective Assessment Attribute* are also important SUMO concepts in hyponym. Overall, the concepts, *Flowering Plant*, *Device* and *Artifact*, are notable both in [8-0-N] hypernym and [9-0-N] hyponym.

SUMO Concept	Term Frequency	(%)
Flowering Plant	106	11.70%
Device	35	3.86%
Artifact	29	3.20%
Food	27	2.98%
Occupational Role	26	2.87%
Fruit Or Vegetable	24	2.65%
Bird	22	2.43%
Human	21	2.32%
Natural Language	20	2.21%
Clothing	19	2.10%

Table 7. The Major Conceptual Distribution in [8-0-N] Hypernym (a total corpus frequency of 906)

SUMO Concept	Term Frequency	(%)
Flowering Plant	31	8.71%
Device	19	5.34%
Artifact	17	4.78%
Subjective Assessment Attribute	14	3.93%
Human	13	3.65%
Social Role	11	3.09%

Table 8. The major conceptual distribution in [9-0-N] Hyponym (a total corpus frequency of 356)

Besides, another interesting phenomenon is that several SUMO concepts are excluded after considering the problematic mappings. In [8-0-N] hypernym, the concepts of *Carnivore*, *Female*, *Male*, *Fresh Water Area*, *Not Fully Formed*, *Removing*, and *Unit Of Measure* are excluded and in [9-0-N] hyponym, *Female*, *Male*, *Dead*,

and *Unit Of Measure* are excluded.

In [10-0-N] meronym and [11-0-N] holonym, they keep one-to-one mapping SUMO concepts. The concept, *Flowering Plant*, still gets the highest percentage both in meronym and holonym. Besides, the data shows that most of the mapping processes occurred randomly. The percentages of these two semantic relations are separately shown in Table 9 and 10.

SUMO Concept	Term Frequency	(%)
Flowering Plant	17	27.87%
Musical Instrument	3	4.92%
Motion	3	4.92%

Table 9. The Major Conceptual Distribution in [10-0-N] Meronym (a total corpus frequency of 61)

SUMO Concept	Term Frequency	(%)
Flowering Plant	7	15.22%
Fruit Or Vegetable	5	10.87%
Ethnic Group	4	8.70%
Artifact	3	6.52%
Body Part	3	6.52%

Table 10. The Major Conceptual Distribution [11-0-N] Holonym (a total corpus frequency of 46)

A more detailed inspection of the SUMO concepts provides some insights into concept generalization. *Flowering Plant*, *Device*, and *Artifact* are the major SUMO concepts corresponding to [8-0-N] hypernym and [9-0-N] hyponym. Moreover, we investigate that *Flowering Plant* SUMO concept is also the major SUMO concept which corresponds to [10-0-N] meronym and [11-0-N] holonym. In a preliminary step, such data reveals the concepts of "plant", "component" and "group" that play the important roles both in [10-0-N] meronym and [11-0-N] holonym. According to the above result, the concept, such as *Flowering Plant*, can be revealed in the relations of hypernym, hyponym, meronym and holonym, but when the humans will use hypernym and hyponym rather than meronym and holonym becomes an interesting question and we probably need to extract those synsets and deeply examine them in order to find the regulation of this phenomenon.

3.3 The Distribution of Semantic Relations in the SUMO Concepts

In order to see how many percentages of each SUMO concept in different semantic relation with the total synset number of each SUMO concept, we examined the results after the calculation processes of the term frequency with the total synset number in each SUMO concept. That means we could see what kind of SUMO concept tends to have specific semantic relations in the translation process. We focused on two parts and observe the data. One focus is to see verbs of [1-0] correct (synonym) , [8-0] hypernym, and [9-0] hyponym. The other focus is to consider about nouns of [8-0] hypernym, [9-0] hyponym, [10-0] meronym, and [11-0] holonym.

SUMO Concept	[1-0-V] Correct (synonym)	[8-0-V] Hypernym	[9-0-V] Hyponym
Communication	51.2%	0.6%	0.4%(tf:4)
Intentional Process	47.3%	0.6%	0.1%(tf:2)
Touching	40.9%	1.9%	0%
Increasing	36.8%	0%	0%
Cutting	34.1%	3.8%	0.8%(tf:5)
Impacting	32.6%	1.2%	0.4%(tf:1)
Decreasing	31.7%	0.9%	0%
Content Development	31.6%	1.5%	0.5%(tf:1)
Removing	28.7%	1.7%	0%
Motion	27.3%	0.9%	0.1%(tf:2)
Walking	26.8%	1.8%	0%
Attaching	26.4%	1.2%	0.3%(tf:1)
Putting	25.5%	0.6%	0.1%(tf:1)

Table 11. The Distribution of SUMO Concepts in [1-0-V] Correct (synonym), [8-0-V] Hypernym, and [9-0-V]Hyponym

Comparing with the percentages of SUMO concepts in several semantic relations, we can have simple statistical results below. In Table 11 we observe that *Communication*, *Intentional Process*, *Touching*, and other SUMO concepts in [1-0-V] correct (synonym) are higher than those in [8-0-V] hypernym and [9-0-V] hyponym. However, several SUMO concepts have interesting distribution in [8-0-V] hypernym. For examples, *Cutting* SUMO concept has up to

3.8% in [8-0-V] hypernym and *Touching*, *Walking*, and *Removing* SUMO concepts are over 1.5%. Though these SUMO concepts are usually in [1-0-V] correct (synonym), they have exceptions to be [8-0-V] hypernym. We have an assumption that the synsets which connect to these SUMO concepts probably have the common characteristics. Unlike the synsets in [1-0-V] correct (synonym), the synsets in [8-0-V] hypernym usually specify the ways, the instruments, and the participants of events.

SUMO Concept	[8-0-N] Hypernym	[9-0-N] Hyponym	[10-0-N] Meronym	[11-0-N] Holonym
Fruit Or Vegetable	3.9%	0.2%	0%	0.8%
Fabric	3.5%	1.4%	0%	0%
Clothing	2.5%	1.0%	0%	0.1%
Food	2.3%	0.4%	0%	0%
Text	2.2%	0.7%	0.1%	0%
Occupational Role	1.8%	0.5%	0%	0.1%
Flowering Plant	1.7%	0.5%	0.3%	0.1%
Fish	1.6%	0.5%	0.1%	0.1%
Social Role	1.5%	0.7%	0.2%	0%
Artifact	1.4%	0.8%	0.1%	0.1%
Device	1.2%	0.7%	0.03%	0%
Human	1.1%	0.5%	0.1%	0%
Body Part	0.5%	0.7%	0.1%	0.2%
Group	0.9%	0.6%	0.6%	0.6%

Table 12. The Distribution of SUMO Concepts in [8-0-N] Hypernym, [9-0-N] Hyponym, [10-0-N] Meronym, [11-0-N] Holonym

Without comparing with the percentages in [1-0-N] correct (synonym), we focus on other important semantic relations of noun terms. Since the percentage in [1-0-N] correct (synonym) is over 50%, most English synsets with Chinese translation equivalents are synonyms, it is interesting to see what distribution of SUMO concepts in other semantic relations tends to have. In Table 12, we can see that the SUMO concepts have several gaps, which are shown as 0% in [10-0-N] meronym and [11-0-N] holonym. For examples, *Fruit Or Vegetable*, *Fabric*, *Clothing*, *Food* SUMO concepts do not exist in [10-0-N] meronym. These SUMO concepts usually tend to be [8-0-N] hypernym. However, we observe that *Body Part* has 0.7% in [9-0-N] hyponym more than 0.5% in [8-0-N] hypernym.

Here, in order to describe the data, we use simple statistic ways and measures of variability to see what these data can show. First, we collect the data which shows the percentages of synset number of SUMO concepts existed in different semantic relations and total synset number in each SUMO concept. Second, we calculate their standard deviations. If the distribution of the data is normal, the standard deviation can give us a lot of information. That is, the significant SUMO concepts in each semantic relation can be generalized and these SUMO concepts can help translators to predict the possible semantic relation between the English synsets and the Chinese translation equivalents. We have several diagrams about the data after standardization. In each semantic relation case, assuming the data states as normal distribution, we observe what SUMO concepts are significant only when they are over 1 standard deviation or over 2 standard deviations.

Diagram1 shows the distribution of major SUMO concepts in [1-0-V] correct (synonym).[3] The distribution of major SUMO concepts in [8-0-N] hypernym is shown in Diagram2. According to these diagrams, the pink line represents the percentages of synset number of SUMO concepts existed in different semantic relations and total synset number in each SUMO concept. The brown line shows the average of these data. The green line shows 1 standard deviation. As for the blue line, it is represented as 2 standard deviations. Based on the data, we can pick the non-synonymous translations that are conceptually dependent. We define the dependency pairs as those concept-LSR pairs whose distribution percentage is greater than one standard deviation from the average, as given in Table13.

Semantic Relations	Significant SUMO Concepts with the Standard Deviations
VERB	
[1-0-V] Correct (synonym)	*Intentional Psychological Process* (2.47), *Communication* (1.50), *Intentional Process* (1.16)
[8-0-V] hypernym	*Cutting* (2.81)
[9-0-V] hyponym	No significant SUMO concepts
NOUN	
[8-0-N] hypernym	*Fruit Or Vegetable* (2.34), *Fabric* (1.74), *Natural Language* (1.57)
[9-0-N]	*Meat* (2.35), *Weapon* (2.15),

[3] Diagram1 and Diagram2 are shown in the Appendix.

hyponym	*Anatomical Structure* (1.97), *Certificate* (1.33), *Fabric* (1.18)
[10-0-N] meronym	*Musical Instrument* (2.55)
[11-0-N] holonym	*Anatomical Structure* (1.40), *Fruit Or Vegetable* (1.12)

Table 13. Conceptual Dependencies of Cross-lingual LSR

In Table 13, we list significant conceptual dependencies of cross-lingual lexical semantic relations in terms of SUMO concepts. In other words, we want to know if the non-synonymous English-Chinese translations are ontology-dependent and can be predicted. In verbs, we found that *Intentional Psychological Process*, *Communication*, and *Intentional Process* are more likely to be translated from English as a synonym in Chinese. Among the three types, *Intentional Psychological Process* is the most significant since its distribution is more than two standard deviation higher than average. This shows that English and Chinese have remarkably similar lexicalization of these highly abstract concepts. For the concepts under *Cutting*, however, we showed that they are more likely to be translated from English to Chinese hypernym,

A closer look at nouns brings even more interesting results. We show that conceptual classes *Fruit Or Vegetable*, *Fabric*, and *Natural Language* are more likely to be translated as hypernyms. This result seem to reflect the fact that the granularity of naming natural objects can be dependent on the experience of the people and they tend to achieve finer granularity with more familiar and important objects. In addition, *Fruit Or Vegetable*, *Meat* and *Weapon* are highly likely to be translated into hyponyms in Chinese. The same principles governing the translational hypernyms should apply here to. But as to why certain conceptual classes are more likely to be translated as hypernyms and other hyponyms will be topic for further future studies. For *Anatomical Structure*, *Certificate*, and *Fabric,* they are likely to be translated as meronyms, And lastly, *Musical Instrument*, *Anatomical Structure* and *Fruit Or Vegetable* are concepts that are likely to be translated as holonyms.

4 Conclusion

In this paper, we present our preliminary study of comparative lexical semantics based on lexical semantic relations. We compared English WordNet (WN) with the English-Chinese Translation Equivalents Databases (ECTED) in order to examine the nature of cross-lingual lexical semantic relations. In addition, adopting the conceptual classification of Suggested Upper Merged Ontology (SUMO), we try to determine if there are any conceptual dependencies for different bilingual lexical semantic relations. We have found preliminary results where some conceptual classes are highly significant for the English-to-Chinese translation lexical semantic relations of synonymy, hypernymy, hyponymy, meronymy and holonymy. In future and more in-depth studies, we will try to give explanatory accounts of the inter-relations between the conceptual classes and the translation lexical semantic relations.

It is interesting to analyze the differences between the English synsets and their corresponding translation equivalents by using their part-of-speech and SUMO concepts. From the above analysis, we notice that most common lexical relation between English and Chinese is still the classical synonymous relation because the total percentage is over 50 %. However, the other half of data of non-synonyms calls for more detailed analysis. We expect our continuing research to shed more lights on the nature of cross-lingual lexical semantic relations.

References

Ahrens, Kathleen. 1999. The Mutability of Noun and Verb Meaning. In Y. Yin et al. (Eds). Chinese Languages and Linguistics V: Interactions in Language. Taipei: Academia Sinica. pp. 335-548.

Asterias J., Climent S., Farreres X., Rigau G. and Rodriguez H. 1997. Combining multiple methods for the automatic construction of multilingual Wordnets. Proceedings of the International Conference "Recent Advances on Natural Language Processing" RANLP'97, Tzigov Chark, Bulgaria.

Fellbaum C. (ed.) 1998. Wordnet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Huang, Chu-Ren, D.B. Tsai, J.Lin, S. Tseng, K.J. Chen and Y. Chuang. 2001. Definition and Test for Lexical Semantic Relation in Chinese. [in Chinese] Paper presented at the Second Chinese Lexical Semantics Workshop. May 2001, Beijing, China.

Huang, Chu-Ren, I-Ju E. Tseng, Dylan B.S. Tsai.

2002. Translating Lexical Semantic Relations: The first step towards Multilingual Wordnets. In Proceedings of the COLONG2002 Workshop "SemaNet:Building and Using Semantic Networks", ed. By Grace Ngai, Pascale Fung, and Kenneth W. Church, 2-8.

Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B.S. Tsai, Brian Murphy. 2003. Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese Wordnet with English WordNet Relations. pp.509-531.

Niles I., and Pease A., 2003. Linking Lexicons and Ontologies: Mapping Wordnet to the Suggested Upper Merged Ontolog. In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003). Las Vegas, Nevada.

Pianta E., Benitivogli L., Girardi C. 2002. MultiWordnet: Developing an aligned nultilingual database. Proceedings of the 1st International WordNet Conference, Maysore, Inda, pp.293-302.

Tsai, D.B.S., Chu-Ren Huang, J.Lin, K.J. Chen and Y. Chuang. 2002. Definition and Test for Lexical Semantic Relation in Chinese. [中文詞義關係的定義與判定原則] Journal of Chinese Information Processing [中文信息學報]. 16.4.21-31.

Vossen P. (ed.). 1998. EuroWordNet: A multilingual database with lexical semantic networks. Norwell, MA: Kluwer Academic Publishers.

Appendix: Conceptual Distribution of Bilingual Synonyms [1-0-V] Correct and [8-0-N] Hypernym

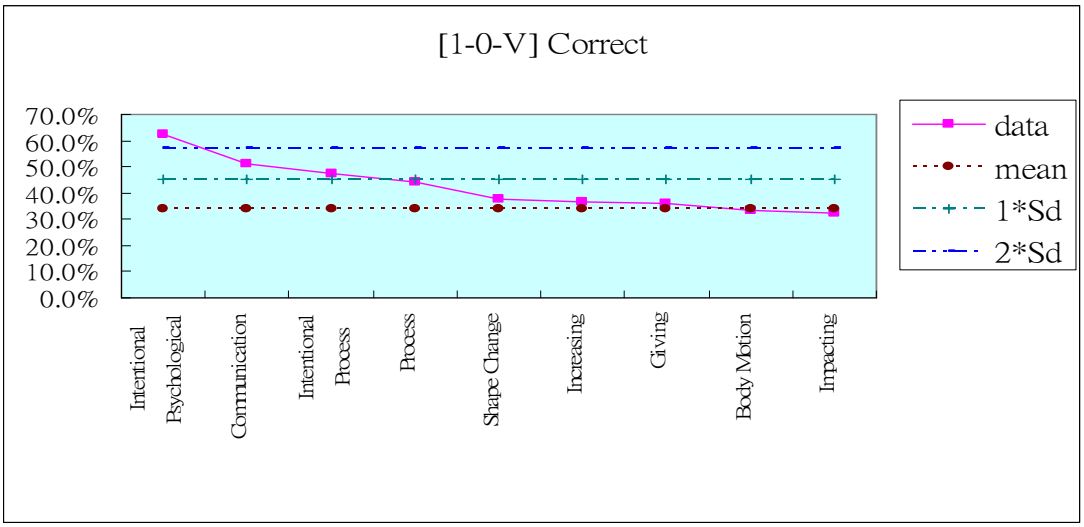

Diagram 1. Conceptual Distribution of Bilingual Synonyms [1-0-V] Correct

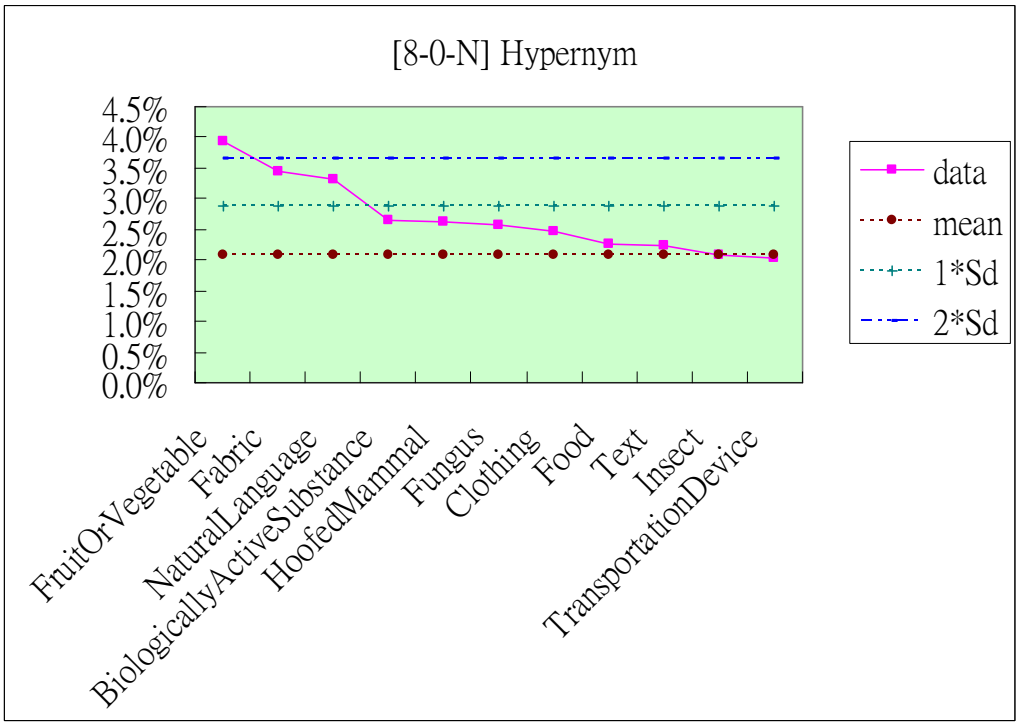

Diagram 2. Conceptual Distribution of Bilingual [8-0-N] Hypernym