Translation Equivalence and Synonymy: Preserving the Synsets in Cross-lingual Wordnets

Oi Yee Kwong
Department of Translation
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
oykwong@arts.cuhk.edu.hk

Abstract

The Princeton WordNet for English was founded on the synonymy relation, and multilingual wordnets are primarily developed by creating equivalent synsets in the respective languages. The process would often rely on translation equivalents obtained from existing bilingual dictionaries. This paper discusses some observations from the Chinese Open Wordnet, especially from the adjective subnet, to illuminate potential blind spots of the approach which may lead to the formation of non-synsets in the new wordnet. With cross-linguistic differences duly taken into account, alternative representations of cross-lingual lexical relations are proposed to better capture the language-specific properties. It is also suggested that such cross-lingual representation encompassing the cognitive as well as linguistic aspects of meaning is beneficial for a lexical resource to be used by both humans and computers.

1 Introduction

The development of multilingual wordnets has been accomplished mostly by starting with the Princeton WordNet for English (Fellbaum, 1998b) and supplying translation equivalents from another language to individual concepts represented by the synsets. When conceptual gaps are identified, they may be handled by the addition or omission of synsets in the new wordnet. While the approach has the merit of good coverage, reliance on translation equivalents may be at the expense of forming non-synsets in the target language wordnet, for which great caution has to be exerted. Past experience from building multilingual wordnets has observed various difficulties, mostly arising from cross-linguistic differences in lexicalisation, conceptual space and sense distinction (e.g. Vossen, 1998). This paper discusses further observations from the Chinese Open Wordnet (Wang and Bond, 2013), which added new translations from authoritative bilingual dictionaries as a means to increase coverage, to show that translation equivalents need to be very carefully screened to avoid some potential and easily overlooked pitfalls. While a good coverage is appreciated, especially with a view to use the wordnets in a variety of computational and human language applications, it is suggested that alternative representations including additional relational pointers be used to accommodate cross-linguistic differences without disturbing the basic infrastructure of WordNet, in particular its basic definition of synsets.

The rest of this paper is organised as follows: Section 2 reviews the theoretical basis of the Princeton WordNet (PWN) and the construction of the Chinese Open Wordnet (COW). Attention will be focused on adjectives. Section 3 presents some observations from COW in terms of its resulting synsets in the adjective subnet. Section 4 discusses the cross-lingual aspects and proposes alternative ways for representing the lexical semantic relations, followed by a conclusion in Section 5.

2 WordNet Infrastructure

2.1 Synsets as the Building Blocks

The original PWN started as a psycholinguistic project for testing the scalability of relational lexical semantics, where concepts are supposed to be linked by specific relations. Its resulting large lexical database turned out to be well received and popularly used by computational linguists. Concepts are expressed or lexically represented by sets of synonyms (synsets) within individual word classes, and are connected by a variety of relational pointers. This essentially results in four sub-

nets, for nouns, verbs, adjectives and adverbs, respectively (Fellbaum, 1998a).

It is therefore well-known that the basic building blocks of the original PWN are the "synsets", which are unordered sets of words that "denote the same concept and are interchangeable in many contexts", and the main relation in WordNet is synonymy[1]. PWN defines word senses by means of synsets. Given the mutual substitutability that holds among members in a synset, membership of a lexical item in a certain synset indicates a particular sense of the word.

2.2 The Adjective Database

Although PWN has four subnets, it is obvious that the noun database and verb database have been the most discussed and utilised (for PWN and multilingual wordnets alike), not only because they contain a larger number of synsets, but perhaps also for the more clearly defined relations established in them. For example, the hypernymy/hyponymy relation for nouns and the troponymy relation for verbs are typical. The adjective database, on the other hand, appears to receive far less attention.

According to Fellbaum et al. (1993), WordNet contains descriptive adjectives and relational adjectives. Descriptive adjectives ascribe a value of an attribute to a noun, such as "heavy" as a value for "weight", indicated in the database by the attribute pointer. The descriptive adjective synsets are not hierarchically ordered as nouns, and apart from the basic semantic relation, antonymy, the semantics of adjectives is more naturally perceived as an N-dimensional space. Adjectives similar in meaning may not all have antonyms, and the similarity pointer is used to mark this phenomenon. Not all gradable attributes have most gradation lexicalised. As remarked by Fellbaum et al. (1993), "It would not be difficult to represent ordered relations by labeled pointers between synsets, but it was estimated that not more than 2% of the more than 2,500 adjective clusters could be organized in that way. Since the conceptually important relation of gradation does not play a central role in the organization of adjectives, it has not been coded in WordNet." In fact, adjectives are considered very polysemous and of limited usefulness in conveying information, and they are not even included in EuroWordNet (Fellbaum, 1998b). But whether this phenomenon is equally insignificant for other languages and its exclusion will not affect the construction of wordnets in those languages may require further thought, and will be discussed in the following sections. It is also noted that "adjectives expressing evaluations (good/bad, desirable/undesirable) can modify almost any noun; those expressing activity (active/passive, fast/slow) or potency (strong/weak, brave/cowardly) also have wide ranges of applicability", which is also a key point to consider when multilingual wordnets are built.

[1] http://wordnet.princeton.edu/

2.3 Wordnets with Translation Equivalents

Since the inception of the EuroWordNet project (Vossen, 1998), which aimed at building a multilingual lexical database for several European languages in the form of PWN, subsequent development of wordnets in other languages has often similarly followed one of the two approaches: the Merge Model or the Expand Model. With the Merge Model, vocabulary selection and synsets are developed separately and locally, followed by generating equivalence relations to PWN. The Expand Model, on the other hand, starts with PWN vocabulary and synsets, and translates the synsets using bilingual dictionaries into equivalent synsets in the other languages.

There have been various attempts for Chinese wordnet (e.g. Huang et al., 2004; Huang et al., 2010; Wang and Bond, 2013; Xu et al., 2008). They primarily relied on some ways to identify translation equivalents, including automatic means and human verification (e.g. Huang et al., 2004). Some limited the number of translation equivalents to be included for a synset (e.g. Huang et al., 2004), while others (e.g. Wang and Bond, 2013) intentionally added more entries.

The Chinese Open Wordnet (COW), in particular, followed the Expand Model and started with the core synsets in PWN (Boyd-Graber et al., 2006), and formulated detailed guidelines to build a better Chinese wordnet. According to Wang and Bond (2013), among the 4,960 core synsets, adjectives occupy only 13.8% of the total. In building the COW, Chinese translations for the core synsets were first obtained by merging existing data from the Southeast University Chinese Wordnet (Xu et al., 2008) and the Open Multilingual Wordnet linked with lemmas extracted from the English Wiktionary (Bond and Foster, 2013). The resulting translations were checked manually, with dele-

tions and amendments as necessary, while new translations found from authoritative bilingual dictionaries were added. The lexical semantic relations were also checked with a random sample from the database (Wang and Bond, 2013).

The manual checking was intended to ensure that the Chinese translations match the English synsets in terms of meanings and parts of speech. Cross-linguistic differences have been recognised all along, especially with respect to lexicalisation, where a specific lexicalised concept in English may not find an equivalent lexicalised form in Chinese, and in such cases a phrase or definition will be used for representing the concept in the Chinese wordnet. Wang and Bond (2013) have also identified a range of situations for which discrepancy within synsets may be found. Where conceptual meaning is concerned, there are cases where two languages may have similar basic conceptual meanings that differ in severity and usage scope. Where affiliated meaning is concerned, words may differ in their affection, genre, and time. Strictly speaking, such cases should be ruled out from the synsets, although a looser standard was adopted for COW, which keeps them to ensure higher coverage but admittedly lower accuracy.

2.4 Potential Blind Spots

In addition to the above known facts, translation equivalents have yet to be more cautiously handled to avoid other potential problems, especially with respect to any incompatibility with the basic WordNet structure. For example, consider the following PWN synset with its correspondence in COW:

01586342-a
nice (pleasant or pleasing or agreeable in nature or appearance)
体贴(的)，合意(的)，美好(的)，和蔼(的)，友好(的)，令人愉快(的)，令人快乐(的)，讨人喜欢(的)

The English synset has only one lexical item, which is not really a problem itself. The tricky part is the "generalness" of this concept, as expressed by the word "nice", in terms of its meaning and usage contexts. As hinted by its gloss, this sense of "nice" can mean "pleasant" or "pleasing" or "agreeable", and such good quality can apply to the "nature" or "appearance" of something. In other words, almost anything can be described as "nice", to mean something good in general without specifying any particular attributes and qualifying how good it is. So strictly speaking, and to be as general as it is, the Chinese equivalent 好 *hǎo* would suffice, and all the items listed above are in a certain sense "over-translation", as they are only conceptually equivalent under certain contexts. For example, 和蔼 *hé'ǎi* can only describe a person, and 美好 *méihǎo* for something inanimate and often more abstract. Meanwhile, 和蔼 *hé'ǎi* is also among the set of words in another adjective sense corresponding to a synset for "kind", as follows:

01372049-a
kind (having or showing a tender and considerate and helpful nature; used especially of persons and their behavior)
体谅(的)，体贴(的)，善良(的)，仁慈(的)，和善(的)，宽厚(的)，友善(的)，好心(的)，好心肠(的)，亲切(的)，温和(的)，和蔼(的)，宽宏大量(的)，友好(的)，乐于助人(的)

Similarly, strictly speaking this sense of "kind" is also quite encompassing, and its fuzziness may be more equivalently represented by 仁慈 *réncí* and 好心 *hǎoxīn*, while leaving others like 友善 *yǒushàn* for "friendly", 乐于助人 *lèyúzhùrén* for "helpful", and 体贴 *tǐtiē* for "considerate".

Given the co-existence of the same lexical items like 和蔼 *hé'ǎi* in correspondence to two synsets relating to "nice" and "kind" separately in PWN, whereas the conceptual distinction in PWN has not considered the two senses synonymous[2], and there is no obvious evidence for multiple senses for 和蔼 *hé'ǎi* according to most dictionaries, it is questionable to treat it as a translation equivalent for the two PWN senses. On the other hand, despite the vague definition for synonymy (as defined by substitutability in a given context), it is readily realised that the criterion is not met for the above examples. No dictionary seems to consider 和蔼 *hé'ǎi* and 体贴 *tǐtiē*, for instance, synonymous in any case as they refer to different qualities of a person. In other words, the set of Chinese words can no longer be qualified as a "synset" as originally

[2]The specific sense of "kind" is not linked to the specific sense of "nice" in PWN via the see-also and similar-to connections. The sense distinction is thus different from other resources, such as the Roget's Thesaurus, where "nice" and "kind" co-exist in group 884 for their sense of "amiable".

defined for the WordNet structure. Moreover, to a certain extent, the conceptual meaning is mingled with specific contextual usage. Thus, when we refer to someone being nice (as in "he is very nice"), it is only as much as saying 他这个人很好 *tā zhège rén hěn hǎo*. Only with more specific context or additional information given could one decide on the way in which he is nice, such as being easy to get along with, very helpful, very generous, or others.

Complete equivalents are generally rare (Svensen, 1993), especially for distant language pairs like English and Chinese, except for very domain-specific concepts and terminologies. The difference in lexicalisation of concepts is also an issue. Since other wordnets are centered on PWN, the lexicalisation in English is taken as a default, which may lead to the use of longer expressions in a synset in other languages. This brings up two issues in constructing wordnets in other languages. One is the seriousness of the problem with respect to different parts of speech. Given the references available for nouns and verbs, and the fuzziness and subjectivity involved in adjectives, we expect that the problem is more pronounced among adjectives. Second, when the coverage of the meanings by the translation equivalents is at the expense of violating the requirements for synsets, are there better ways to handle such cases? In the following sections, we analyse the situation with reference to COW, and discuss possible alternatives for representing the lexical semantics therein.

3 Synsets in COW

The Chinese Open Wordnet (COW)[3] consists of 42,312 synsets (Nouns 65.9%, Verbs 12.2%, Adjectives 20.2%, Adverbs 1.7%) with 80,009 lexical items (Nouns 57.9%, Verbs 16.7%, Adjectives 22.9%, Adverbs 2.5%). The following discussion covers the three major word classes, namely nouns, verbs and adjectives, with focus on adjectives, and adverbs are excluded.

3.1 Synset Size and Polysemy

In terms of synset sizes, as measured by the number of items in a synset, the largest range was observed for nouns, from 1 to 39 items in a synset, followed by adjectives and verbs, from 1 to 15 and from 1 to 13 respectively. As shown in Figure 1, noun synsets tend to be of smaller sizes than adjective synsets, and there are relatively even more larger synsets for verbs. Many of the extreme examples in the noun database have to do with biological nomenclature, as when a certain plant species is known by many formal and informal names in Chinese, as well as culture-specific items which lack one-to-one correspondences, such as:

12896307-n
black nightshade, common nightshade, poisonberry, poisonberry, Solanum nigrum (Eurasian herb naturalized in America having white flowers and poisonous hairy foliage and bearing black berries that are sometimes poisonous but sometimes edible)
老鸦酸浆草, 乌归菜, 野葡萄, 酸浆草, 救儿草, 黑姑娘, 天泡果, 地戎草, 七粒扣, 山海椒, 黑茄, 野茄子, 天泡草, 地泡子, 天天茄, 天茄子, 野辣角, 野海椒, 后红子, 天茄苗儿, 老鸦眼睛草, 水茄, 水苦菜, 野伞子, 天茄菜, 山辣椒, 狗钮子, 苦葵, 苦菜, 野茄菜, 飞天龙, 龙葵, 耳坠菜, 乌疔草, 野辣椒

09823502-n
aunt, auntie, aunty (the sister of your father or mother; the wife of your uncle)
妗, 姑母, 伯母, 姑姑, 老大妈, 阿姨, 妗母, 叔母, 姑妈, 舅母, 姑, 姨妈, 姨, 舅妈, 婶子, 婶婶, 姨母, 婶母

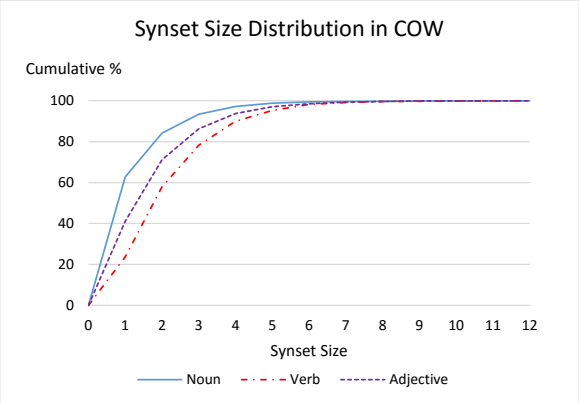

Figure 1: Synset Size Distribution for Various Word Classes in COW

The above two examples actually reveal two very different scenarios. Although we probably need a biologist or an expert in herbal medicine

[3]Downloaded from http://compling.hss.ntu.edu.sg/omw/

to verify the many renditions for the very same plant species, as far as they are valid names, they can certainly be considered synonymous. But the second case corresponds to an obvious difference in sense distinction as a consequence of cultural difference. While "aunt" can refer to one of the many female relatives as indicated in the gloss, the Chinese words are not all interchangeable because each of them only refers to one type of the female relatives, e.g. 姑母 *gūmǔ* and 姑姑 *gūgu* for "the sister of one's father" (further distinguished as the elder and younger sister respectively in some dialects), 舅母 *jiùmǔ* and 舅妈 *jiùmā* for "the wife of the brother of one's mother", etc. In other words, although they can be considered translation equivalents for "aunt" in a given context, they are definitely not synonyms.

The issue is also quite different from what can be observed from the adjective database and verb database. The large synsets in them do not really contain multiple renditions for the same conceptual meaning as in the noun examples above, but more often reflect the polysemy contained by the concepts as represented by the English synsets which results in translational differences in Chinese, such as:

01256332-a
hot (extended meanings; especially of psychological heat; marked by intensity or vehemence especially of passion or enthusiasm)
流行(的), 热切(的), 激烈(的), 热门(的), 才发行(的), 急躁(的), 销路好(的), 刚出版(的), 轰动一时(的), 最新(的), 紧缺(的), 激动(的), 狂热(的), 热烈(的), 时新(的)

01215137-v
arrest, pick up, nail, apprehend, nab, collar, cop (take into custody)
捕捉, 捉到, 捕获, 逮捕, 拘留, 拘押, 拘捕, 抓住, 抓获, 当场逮捕, 擒获, 逮住

The adjective example is another typical one, like those mentioned in an earlier section, which apparently violates the requirements for synsets. It is least likely that one would equate 急躁 *jízào* (impatient) with 流行 *liúxíng* (popular), although the examples given in the English synset include a whole lot of extended usage of "hot" as in "a hot temper", "a hot topic", "a hot new book", "a hot love affair", and "a hot argument", while the encompassing "hot" has to be rendered according to its subtle sense difference according to the noun it modifies. Thus the "hotness" associated with "temper" is not the same "hotness" associated with "topic" in Chinese, which are therefore non-synonyms.

As for the verb example, the English synset obviously refers to "arrest by police". Nevertheless, the Chinese expressions like 捕捉 *bǔzhuō* (catch) may be too general while those like 当场逮捕 *dāngchǎng dàibǔ* (arrest on the scene) are seemingly over-specific. Issues with the verb synsets are no less complex than those pertaining to adjectives, and will not be pursued further in the current discussion. However, the verbal synset above can also illustrate a logical issue. It is not appropriate to find 逮捕 *dàibǔ* (arrest) and 当场逮捕 *dāngchǎng dàibǔ* (arrest on the scene) in the same synset, not only because the latter is a more specific meaning than the former, but also the latter is a phrasal expression (with modifier and verb) which cannot logically mean the same thing as the simple lexical verb.

3.2 Adjectives and Non-synsets

We selected 200 top-sized adjective synsets from COW and examined the synonymy therein. It turns out that at most 27 out of the 200 synsets do not contain phrasal members (in addition to lexicalised items)[4]. While this does not necessarily mean that over 85% of the English adjectives in these synsets do not have lexicalised translation equivalents in Chinese, it at least shows that bilingual dictionaries may tend to provide translated definitions or paraphrase instead of or in addition to translation equivalents. Although this is an unavoidable practice in bilingual lexicography (Atkins and Rundell, 2008), its compatibility with WordNet structure is questionable. It is thus worth to reconsider their representation in the resource, adhering to the lexicalisation criterion on the one hand (e.g. Huang et al., 2010) and expanding the overall coverage on the other (e.g. Wang and Bond, 2013).

The lexicalisation issue aside, it was observed from the selected data that non-synsets often result from one or more of the following situations:

[4]Some common and fixed four-character expressions are considered single words, e.g. 无忧无虑 *wúyōuwúlǜ* (carefree), while those containing an obvious combination of two or more words are considered phrasal expressions, e.g. 轻松愉快 *qīngsōng yúkuài* (relaxed and happy).

1. Different sense distinctions

   The difference in the division of semantic space and granularity of sense distinction is particularly salient with the more "general" adjectives already illustrated above. But even for the less "general" adjectives, the broadened coverage may not always match the sense granularity indicated in PWN, especially as PWN is known for its possibly over-fine-grained senses. For example, "civilised" belongs to two synsets in PWN, and here are their parallel Chinese synsets:

   00411886-a
   civilized, civilised (having a high state of culture and development both social and technological)
   文明化(的), 有礼貌(的), 有教养(的), 开化(的), 文明(的), 文雅(的)

   01947741-a
   cultured, polite, civilized, civilised, cultivated, genteel (marked by refinement in taste and manners)
   文雅(的), 有礼貌(的), 优雅(的), 有教养(的), 有礼(的), 文明(的), 有先进文化(的), 有修养(的)

   The two senses of "civilised" are quite distinct, such that the first refers to a general high state of development in a collective sense and the second specifically relates to more personal and individual behaviour. But the Chinese synsets overlap considerably, especially when 有礼貌 *yǒulǐmào* (polite), 有教养 *yǒujiàoyǎng* (cultivated) and 文雅 *wényǎ* (elegant) are more relevant to the second sense than the first.

2. Over-interpretation of concepts

   In addition to the examples like "hot" and "kind" discussed above, over-interpreting a concept may lead to obscure results as in:

   02328659-a
   docile (willing to be taught or led or supervised or directed)
   易管教(的), 驯服(的), 易教育(的), 易驾驭(的), 可教导(的), 容易教(的), 听话(的), 驯良(的), 愿学习(的), 易训练(的), 温顺(的), 顺从(的), 易控制(的)

   While lexicalised items like 驯服 *xúnfú* and 温顺 *wēnshùn* may already satisfactorily represent the concept in Chinese, the others like 易管教 *yì guǎnjiào* (easy to teach) and 易驾驭 *yì jiàyù* (easy to control) may still be acceptable except that they are phrasal expressions. However, 愿学习 *yuàn xuéxí* (willing to learn) seems to have over-interpreted in the sense that "willing to learn" may not necessarily mean "willing to be taught / well-behaved / easy to control".

3. Multiple facets of concepts

   Relating less to sense granularity but more to individual context of usage, some adjectives may highlight different facets of a certain quality when modifying different things. For example:

   02964782-a
   Chinese (of or pertaining to China or its peoples or cultures)
   中国文化(的), 汉, 华, 中文(的), 中国人(的), 汉语(的), 中国话(的), 中国(的), 中

   As clearly indicated by its gloss, the adjective "Chinese" in this synset pertains to various aspects relating to China, while the Chinese synset, although reflecting these many potential facets, does not really contain synonyms, as 中国人 *zhōngguórén* (Chinese people) and 中国话 *zhōngguóhuà* (Chinese language) are both included.

4. Related but subtly different words

   This situation is not simply a one-to-many correspondence, but there are more subtly defined Chinese lexical items which may only be coarsely represented by the same set of synonymous English words. For example:

   00372111-a
   brown, brownish, dark-brown, chocolate-brown (of a color similar to that of wood or earth)
   咖啡色(的), 呈褐色(的), 黑褐色(的), 茶褐色(的), 棕色(的), 褐色(的)

   Strictly speaking the Chinese words correspond to different hues and intensities of "brownness", which are more specific than the English synset.

5. Contradictory connotation

   Logically, lexical items or expressions with opposite connotations cannot be synonyms as they are not mutually substitutable in all contexts. For example:

   00438909-a
   sharp, shrewd, astute (marked by practical hardheaded intelligence)
   狡黠(的), 锐利(的), 精明(的), 狡猾(的), 机敏(的), 诡计多端(的), 锋利(的)

   The English items are somewhat neutral or even positive, which are more or less equivalently represented by 精明 *jīngmíng* and 机敏 *jīmǐn*, but 狡黠 *jiǎoxiá*, 狡猾 *jiǎohuá* and 诡计多端 *guǐjìduōduān* are obviously derogatory.

4 Handling Extra-synset Information

While it is intrinsically more difficult to define the synsets and concepts represented by adjectives due to their polysemy, even in PWN, the adjective database also reveals important conceptual and lexical gaps across languages. Multilingual wordnets, in this regard, would provide useful resources for language learning and translation, by humans and machines alike. It has been shown from the above discussion that apart from paying attention to cultural and linguistic differences across languages, building wordnets in other languages based on translation equivalents from bilingual dictionaries does not necessarily result in equivalent and valid synsets. This issue is a salient one, especially for languages with very different morphological properties and word formation mechanisms from English. For instance, while new words can easily be formed by inflectional and derivational morphology in English, the meaning carried by the additional morphemes may often be straightforwardly rendered with an extra word in Chinese, such as *un-X* to 不X (e.g. unhappy 不快乐 *bù kuàilè*) and *X-able* to 可X (e.g. respectable 可尊敬 *kě zūnjìng*)[5].

Realising the importance and potential use of the multiple forms and renditions of a given meaning in Chinese, or other languages which are similarly distant from English, it would therefore be value-adding to accommodate them in wordnets in some way. But the thesis in the current discussion is that the basic structure of synsets foundational to PWN should be maintained in multilingual wordnets. The following proposals are thus made to ensure that synsets are preserved as much as possible in target language wordnets while enabling language-specific properties and useful information to be captured:

1. An equivalent synset to a PWN synset should preferably contain only lexicalised items in the target language, unless no lexicalised translation equivalent is available. It is easy to get too far and result in over-interpretation with phrasal or clausal expressions. For example, synset **01251128-a** *cold* (having a low or inadequate temperature or feeling a sensation of coldness or having been made cold by e.g. ice or refrigeration) could be represented with 冰 *bīng*, 冻 *dòng*, 冷 *lěng*, 寒 *hán*, and perhaps the near-synonymous disyllabic words 冰冻 *bīngdòng*, 冰冷 *bīnglěng*, and 寒冷 *hánlěng*. The expressions above the lexical level, such as 气温低 *qìwēndī*, 温度不足 *wēndù bùzú* and 温度没有达到要求 *wēndù méiyǒu dádào yāoqiú*, which are actually parallel to the gloss, should better be excluded from the synset.

2. The other non-lexicalised expressions which nevertheless convey the meaning close enough to the sense of the original synset, including but not limited to the examples above, could be stored in a separate class in a language-specific structure, instead of the core wordnet structure or the Inter-Lingual-Index. These separate and language-specific classes can be linked to the base concepts in WordNet with an *extension* pointer.

3. For very general adjectives, or those that are highly polysemous depending on the nouns being modified, similarly general equivalents, if available, should be included in the corresponding synset. The collocation-specific equivalents (that is, possible words actually used in the target language when the adjective is used to modify a particular noun) are different facets or even senses of the general adjective, and should therefore be captured at yet another subsuming level. This could be done in one of the two

[5] Sometimes disyllabic words as a more lexicalised form are available, e.g. 不快 *bùkuài* or 不乐 *bùlè* for "unhappy" and 可敬 *kějìng* for "respectable", although they might be considered leaning toward classical Chinese.

ways. If PWN does not have a synset corresponding to a specific meaning of the general adjective, an extra synset can be introduced in the target language wordnet, with a *sub-level* pointer from the general adjective synset to the relevant senses as distinguished in the target language. Meanwhile, if there are existing adjective synsets corresponding to the specific adjectives in PWN, they could be linked as in PWN by relational pointers like *similar_to*. For example, synset **02569558-a** *sagacious, perspicacious, sapient* (acutely insightful and wise) could correspond to a Chinese synset with 睿智 *ruìzhì* with a pointer to the more general adjective synset like **02569130-a** *wise* (having or prompted by wisdom or discernment), while synset **00438909-a** *sharp, shrewd, astute* (marked by practical hardheaded intelligence) as discussed above, revised as 精明 *jīngmíng*, 机敏 *jīmǐn*, can point to synset **00438707-a** *smart* (showing mental alertness and calculation and resourcefulness). The two more general adjectives (wise and smart) can correspond to the more general Chinese adjectives like 聪明 *cōngmíng* and 聪颖 *cōngyǐng*.

4. In fact, very similar words like "clever", "wise", "smart", "intelligent", "sharp", "sagacious", "canny", and many others, are not easy to distinguish in a clear manner. Subtle differences are also found among the many similar words in Chinese such as 聪明 *cōngmíng*, 聪颖 *cōngyǐng*, 聪敏 *cōngmǐn*, 机智 *jīzhì*, 睿智 *ruìzhì*, 英明 *yīngmíng*, 精明 *jīngmíng*, 明智 *míngzhì*, etc. It is nevertheless obvious, and perhaps intuitive to the native speakers, that 聪明 *cōngmíng* describes cleverness in a most general sense, and others describe a more specific aspect of cleverness, such as being mentally quick (e.g. 机智 *jīzhì*) or able to make wise decisions (e.g. 英明 *yīngmíng*). It is thus linguistically unsatisfactory to merge all these items into a particular synset. On the one hand, they may not be equally synonymous with one another as they tend to be used for a particular aspect of intelligence, depending on the usage context. On the other hand, the appearance of the same item in too many synsets may defeat the purpose of defining senses as such, giving a distorted picture of sense distinction and polysemy. In this regard, the *pertainym* relation in PWN could be utilised in a target language wordnet for connecting adjective synsets with noun synsets to enhance the cross-POS relations in wordnets in addition to the morphosemantic links, like the synset with 英明 *yīngmíng* can pertain to both "human" and "decision".

5. To ensure logical validity, words with contradictory connotation should be avoided in a synset. Similarly, phrasal expressions should be prudently handled as the same concept should not really correspond to both one lexical item and another form of it qualified by a degree adverb or so. For example, "very drunk" cannot be at the same time 喝醉 *hēzuì* and 烂醉 *lànzuì*, as the former only means "drunk after drinking" while the latter indicates how seriously one is drunk. Similarly, 贫困 *pínkùn* (impoverished) and 极度贫困 *jídù pínkùn* (extremely impoverished) cannot mean the same thing at the same time. The item which most matches the concept represented by the synset will suffice.

5 Conclusion

This paper has thus raised the issue of preserving the synonymy relation holding in synsets as the basic building blocks for wordnets in other languages, while taking advantage of the translation equivalents from other lexical resources as a starting point. Examples from Chinese were highlighted to illustrate how cross-linguistic differences especially in morphology and word formation may result in non-synsets in the process of building wordnet in a target language. It has been shown that the adjective database is particularly prone to the problem, especially for the relatively "general" concepts expressed by adjectives which can be used to describe many different entities and qualify a wide range of properties. To avoid non-synsets, it is thus suggested that partial equivalence be handled in a target wordnet by connecting the context-dependent equivalents to the basic synset with extra relational pointers. Although the alternative representation may not make any significant difference as far as the coverage and actual usage of the resource is concerned, it is nevertheless fundamentally important to keep the theoretical foundation intact.

Acknowledgements

The work described in this paper was partially supported by grants from the Faculty of Arts of the Chinese University of Hong Kong (Project No. 4051094) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 14616317).

References

B.T. Sue Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1352–1362, Sofia, Bulgaria.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted, connections to WordNet. In *Proceedings of the Third Global WordNet Meeting*, Jeju, Korea.

Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Adjectives in WordNet. In George A. Miller, editor, *Five Papers on WordNet*. http://wordnetcode.princeton.edu/5papers.pdf.

Christiane Fellbaum. 1998a. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Christiane Fellbaum. 1998b. A semantic network of English: The mother of all WordNets. *Computers and the Humanities*, 32(2/3):209–220.

Chu-Ren Huang, Ru-Yng Chang, and Shiang-Bin Lee. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1553–1556.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese WordNet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23.

Bo Svensen. 1993. *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford University Press.

Piek Vossen. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, 32(2/3):73–89.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18, Nagoya, Japan.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In John Domingue and Chutiporn Anutariya, editors, *The Semantic Web: 3rd Asian Semantic Web Conference*, volume 5367, pages 302–314. Springer.