Corpus-Driven Study of Translation Units in an English-Chinese Parallel Corpus

Weiqun Wang[1]

Abstract

It is widely acknowledged that texts are not translated word by word, but unit by unit. Single words are polysemous and therefore ambiguous in translation. Corpus linguistics, in monolingual context, has replaced the traditional basic notion of meaning (words) with the extended unit of meaning. Accordingly, this paper argues that in bilingual context, the translation unit, as the counterpart concept of the unit of meaning, replaces single words as the basic unit in translation. This paper aims at turning the study focus of parallel corpora from single words to larger units—translation units. It shows how to extend a selected sample of thirty Adjective+Noun (A+N) phrases into complete translation units by looking at their translation equivalence in an English-Chinese parallel corpus.

1. Introduction

The prime achievement of corpus linguistics is to look at words embedded in context. Represented by Sinclair (1991, 1996, 2004), corpus linguists argue that words will be disambiguated if they are looked at together with their collocates. They thus extend the basic unit of language from single words to the extended unit of meaning – the lexical item (Sinclair, 1996, 2004; Teubert, 2005). Teubert (1996, 2001, 2002) proposes that in a bilingual context, larger units, rather than the single words, should be analysed to identify their translation equivalence. In bilingual or multilingual contexts, the concept of a unit of meaning should be replaced by the concept of a translation unit. He asserts that parallel corpora could provide a solution to the problem of disambiguation because parallel corpora consist of authentic translation texts and they are repositories of translation units and translation equivalents. In practice, he has led the *TranslationBase* project at the University of Birmingham which aims to provide ready-to-use translation equivalents in order to facilitate the translation (see Chang et al, 2005). The aim is to turn the focus of bilingual or multilingual parallel corpora study from single words to multiword units - the translation units. Chang et al (2005) focused on the automatic extraction of English-Chinese translation equivalence through statistical approaches. However, their successes are limited because the statistical approach alone does not work, and many of their extracted units are not expected translation units.

The unsuccessful attempt of the extraction by pure statistical methods suggests that we need to describe the linguistics features before we can automatically extract them. What this paper seeks to do is to identify the translation units by looking at their translation equivalents according to the definition of the translation unit as the

[1] Centre for Corpus Research, Department of English, University of Birmingham, Birmingham B15 2TT, UK
e-mail: w.wang.2@bham.ac.uk

smallest monosemous unit in translation. It will also propose linguistic criteria for identifying translation units. Based on the Hong Kong Legal Document Parallel Corpus (henceforth HKLDC), the Adjective +Noun (henceforth A+N) phrases have been extracted from the corpus. From these, thirty samples have been chosen and treated as the node. The translation units and their equivalents have been studied. The hypothesis is that each translation unit should have only one translation equivalent. The paper aims to verify whether the above theoretical assumption is correct or not. This paper will suggest how and when they could be expanded into translation units. In this way, some of the characteristics and properties of translation units will be described which hopefully could benefit other researchers in automatic extraction of all translation units and translation equivalents.

The paper is organised as follows: Section 2 defines the concept of translation units. Section 3 describes thirty A+N phrases and their translation equivalents overall. Section 4 analyses those A+N phrases with a unique translation equivalent; and Section 5 investigates those with more than one translation equivalent. Section 6 concludes the paper.

2. Definition of Translation unit

It is widely known that professional translators do not translate texts word by word. They are normally translating large chunks, for example a collocation, as a whole. These chunks, are called “translation-units” (Teubert 1996, 2001, 2002). In some sense, translation units are centred on lexical words. “Lexical unit and relevant context together form the translational unit” (Teubert, 1996: 256). What is more, the translation unit is unambiguously translated. Ideally, it is “the smallest monosemous unit in translation” (Teubert, 2001, 2002).

The equivalence of a translation unit in the target language is called a translation equivalent (Teubert, 2001). The translation equivalent is regarded as the “paraphrase” of the meaning of a translation unit but in the target language (Teubert, 2001:145). In other words, a translation equivalent is the meaning of a translation unit in the target language.

A key theoretical assumption is that a translation unit is, ideally, monosemous, which means that it will have only one translation equivalent. If it has more than one translation equivalent, these translation equivalents should be synonymous and can replace each other. If there is more than one target language equivalent and they are not synonymous, then the source language expression is not yet a translation unit, therefore has to be extended until it becomes a translation unit. In other words, one or more context words have to be added to it until it becomes, from the target language perspective, unambiguous.

Based on the above definition of translation units, this paper will propose three linguistic principles as the criteria of extracting complete translation units: *complete principle, monosemous principle* and *minimal principle. Complete principle* means that a translation unit should be a complete unit of meaning in the source text, and its translation equivalent is a complete unit of meaning in the target text. All the words and even domains which help disambiguation should be included in the translation units. *Monosemous principle* refers to a translation unit which should only have one meaning -- represented in the target language as one semantic translation equivalent, although sometimes the equivalents may have synonymous variations. The third principle *minimal principle* means that a translation unit should be kept as small as possible. Like the principle of Occam's Razor, the simplest or smallest translation unit is the best.

3. Extraction of Translation Candidates and Their Equivalents

The test bed for the above assumption is the HKLDC. The HKLDC is an English-Chinese parallel corpus compiled at the University of Birmingham. It contains the statutory laws issued by the Department of Justice of the Hong Kong S.A.R. Government (http://www.justice.gov.hk). The whole corpus has more than 10 million words (approximately 5.6M English and 4.6M Chinese characters). No matter whether they are in English or Chinese text, Hong Kong bilingual laws have equal status in legislation. However, linguistically, English is the source. The HKLDC has been POS-tagged, and sentence-level aligned and the Chinese text is segmented. For details please see Chang et al (2005).

A Perl programme is used to extract all the A+N bigram English phrases in the HKLDC. This yields more than 9,000 A+N phrases with three occurrences and above. Among them, thirty English A+N phrases have been selected to extract their Chinese translation equivalents, listed in Table 1. The first column in Table 1 gives the frequency of each phrase in the whole corpus. These thirty phrases were chosen because they appeared to be promising candidates for translation units, and because they occurred around 100 times (the highest frequency was 105 times and the lowest was 88)[2], which means they were not the most frequent ones but sufficiently frequent to permit reliable conclusions.

Table 1: 30 A+N phrases

Frequency	A+N Phrase	Frequency	A+N Phrase
105	straight line	94	legal adviser
104	legal officer	93	registered dentist
101	residential care	93	postal packet
101	criminal offences	93	good order
100	annual allowance	92	special category
99	long term	92	registered scheme
98	human remains	92	provisional registration
98	conclusive evidence	92	judicial trustee
97	written permission	91	internal combustion
97	public bus	91	final Appeal
97	personal representatives	90	necessary modifications
97	first column	89	rateable value
96	notifiable workplace	88	restricted licence
96	listed company	88	reasonable ground
95	light bus	88	medical officer

[2] These frequency figures are calculated by the Perl program which is used to extract the A+N phrases. However, these figures can only be used to ascertain roughly how frequently the phrases appear in the whole corpus. Different concordancing software may not yield exactly the same figures due to the different design of the query (e.g. some software queries may not include capital letters). For example, both ParaConc and the Perl program yielded 105 occurrences of the phrase *straight line*. However, Concapp, a free concordancing program by the Virtual Language Centre of the Polytechnic University of Hong Kong, yielded 106 instances of this phrase. Still, the results should be and actually are approximately the same. This study will use only the frequency figures yielded by the Perl program unless there is a fundamental difference between the figures in this study and the figures according to other software.

The translation equivalents of the above A+N phrases are manually identified. For each phrase, thirty sentences are extracted where the phrase occur. All the phrases have their lexical equivalents apart from one case of a phrase. This zero correspondence will be ignored in this paper.

All these thirty A+N phrases have been translated into nominal Chinese equivalents except those of *long term* and four equivalents of *good order.* This will be discussed later in Section 5. Among all the thirty A+N phrases, two-thirds have unique translation equivalents. The remaining ten phrases have more than one translation equivalent.

According to the number of their equivalents, the thirty phrases fall into two groups: twenty phrases have one unique Chinese equivalent and ten phrases with more than one. The 20 A+N phrases with a unique translation equivalent further fall into three groups: The first type are complete translation units; they do not need to be used with other words to form a whole unit of meaning. The second type are those which have not been found being used independently; they are only part of complete translation units. This type should be extended to bigger units to form complete translation units because they are always used together with other words. These words are indispensable in forming a complete unit of meaning. The third type are complete translation units only in certain cases but in the remaining cases they need to be extended to form complete translation units (see Section 4). The remaining ten phrases are not translated into the same Chinese equivalents. These ten phrases have more than one translation equivalent. They need to be extended to complete translation units. Again, these ten A+N phrases can be classified into two groups according to whether they have synonymous equivalents or not (See Section 5).

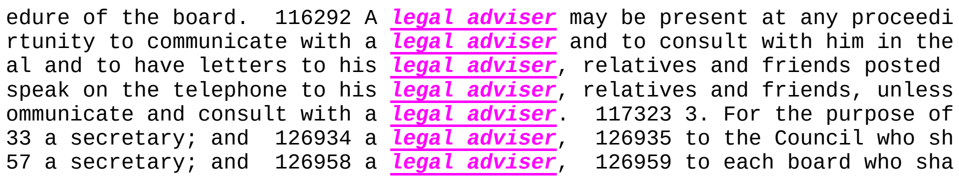
Semantically, the translation equivalents of the thirty A+N phrases fall into three categories: 1) all the translation equivalents of an A+N phrase are the same; 2) the translation equivalents of an A+N phrase are not exactly the same, but they are synonymous; 3) the translation equivalent of an A+N phrase is neither the same nor synonymous. They are different in meaning.

4. Analysis of Phrases with Unique Translation Equivalent

There are thirteen A+N phrases that can be regarded as whole translation units. Each of them occur independently in the sentence and has only one translation equivalent, or unique translation equivalent. For example, with *legal adviser*, the following sample concordance lines show that they occur independently, without semantic interference of other pre-modified lexical words or post-modified lexical words. They do not have a strong collocability with other grammatical words either. All these occurrences of legal adviser have been translated as *法律顾问*. Therefore, *legal adviser* is regarded as a complete translation unit and *法律顾问* as its translation equivalent.

Figure 1: Concordance of *legal adviser*

```
d by him or by his friends or legal adviser, under the same conditions as
ns as apply to a visit by his legal adviser.  92095 Every prisoner awaitin
d by him or by his friends or legal adviser, under the same conditions as
ns as apply to a visit by his legal adviser.  92111 Every appellant may se
ation Committee may appoint a legal adviser to advise it on any points of
edure of the board.  116145 A legal adviser may be present at any proceedi
edure of the board.  116199 A legal adviser may be present at any proceedi
```



```
edure of the board.  116292 A legal adviser may be present at any proceedi
rtunity to communicate with a legal adviser and to consult with him in the
al and to have letters to his legal adviser, relatives and friends posted
speak on the telephone to his legal adviser, relatives and friends, unless
ommunicate and consult with a legal adviser.  117323 3. For the purpose of
33 a secretary; and  126934 a legal adviser,  126935 to the Council who sh
57 a secretary; and  126958 a legal adviser,  126959 to each board who sha
```

There are other twelve phrases fall in this category. These are: *straight line, criminal offences, annual allowance, first column, notifiable workplace, listed company, legal adviser, registered dentist, postal packet, registered scheme, judicial trustee, rateable value*. Each of them has been unanimously translated into one translation equivalent in Chinese.

The second type is those which are always a part of larger translation units. There are four phrases belonging to this type and they have been listed in Table 2. The larger units extended from these phrases, are listed in the middle column in Table 2.

Table 2: A+N phrases as Parts of Larger Translation Units.

A+N Phrase	Complete Translation Unit	Chinese Equivalent
special category	special category space(s)	特种舱
final appeal	(the) court of final appeal	终审法院
restricted licence	restricted licence bank	有限制牌照银行
internal combustion	internal combustion engine/12	内燃机
	internal combustion type machinery/8	内燃式机械
	internal combustion marine machine/2	内燃船机
	internal combustion type propelling machinery/9	内燃式推进机械

All the occurrences of *special category* have been extracted by using the software *Wordsmith*. The results have shown that whenever *special category* occurs, it occurs with the word *space*, either in singular form (*space*) or in plural form (*spaces*). This indicates that the phrase *special category* itself, in this corpus, is not an independent unit but normally requires the company of the third lexical word in order to make a full translation unit. In other words, *special category* collocates with *space(s).* All the instances of *special category space(s)* have been translated as *特种舱* in the Chinese text. Therefore, the complete translation unit should be *special category space(s)* instead of *special category*.

Similarly, *final appeal* does not occur alone but with *(the) court of final appeal*. In the translation, *(the) court of final appeal* has been translated as *终审法院.* This indicates that *final appeal* is only a part of a larger translation unit -- *(the) court of final appeal*.

The situation is the same for *restricted licence*. Whenever *restricted licence* occurs, it occurs as *restricted licence bank*. Throughout the corpus, *restricted licence* has to go with the word *bank* in order to make the whole unit of meaning. The Chinese equivalence of *restricted licence bank* is uniformly *有限制牌照银行.*

The phrase *internal combustion* follows the same pattern as the above three phrases except that it can be a part of more than one larger unit in this parallel corpus. Four translation units are formed: *internal combustion engine*, *internal combustion type machinery, internal combustion marine machine* and *internal combustion type* *propelling machinery*. The frequency of each of these units has been listed in Table 2. Each of the units have been translated as respective Chinese phrases.

Although their Chinese counterparts can be identified in the larger translation equivalents, all these four A+N phrases belong to larger translation units since they collocate with the words following them. Together with the words following, they form another different concept. For instance, *special category space* is different from *special category* in meaning. Thus, we argue that in this corpus, the bigger units should be the complete translation units.

The third type is listed in Table 3. These examples can both occur independently to form a unit of meaning, and also form another unit of meaning with the other adjacent words. There are three of this kind of A+N phrases: *personal representative*, *public bus*, and *provisional registration.*

Table 3: A+N Phrases Both as Translation Unit and as parts of Translation Units:

A+N phrase	**Translation Unit/Freq.**	**Chinese Equivalent**
personal representatives	Personal representative/35	遗产代理人
	Legal personal representative/4	合法遗产代理人
public bus	public bus/2	公共巴士
	public bus service/30	公共巴士服务
provisional registration	provisional registration/23	临时注册
	certificate of provisional registration/9	临时注册证明书

Among the thirty-nine occurrences of the phrase *personal representative*, it occurs independently thirty-five times; that is, it occurs without the accompaniment of any other lexical words to form a unit of meaning. All these thirty-five examples of *personal representative* have been translated into *遗产代理人*. There are another four occurrences where *personal representative* occurs with the word *legal* in front to form another unit of meaning, *legal personal representative*. This new unit of meaning has been translated as *合法遗产代理人*. The word *legal* is polysemous in the English monolingual dictionary, and has more than one translation equivalence in HKLDC, e.g., *法律*, *合法*, *法定* and *律政(的)*. In the translation of *legal personal representative*, *legal* has lost the meanings of the other three translation equivalents but all the four uses of *legal* in the phrase *legal personal representative* have been translated as *合法的.* The phrase *legal personal representative* is regarded as a new translation unit although the translation equivalent of *legal* seems only to be added in front of the translation equivalent of *personal representative*.

Public bus and *provisional registration* are similar to *personal representative*. However, their dominant forms vary. For *personal representative* and *provisional registration*, the independent forms *personal representative* and *provisional registration* occur more than their extended forms (*legal personal representative* and *certificate of provisional registration*). However, for *public bus,* the larger translation unit *public bus service* occurs more (occurring twenty-eight times more than the independent form *public bus*). The frequency of the different forms may only be defined by the content of the text. Here we would like to focus on the translation equivalents of these units. No matter whether they occur alone or as parts of larger units, these three A+N phrases have been translated as the same equivalents. In other words, all *personal representatives* have been translated as *遗产代理人*, whether it is used alone or in the larger unit *legal personal representative.* Similarly, all references to *public bus* have been rendered as *公共巴士*, and all citations of *provisional registration* have been rendered as *临时注册.*

It may seem to be a theoretical notion that the larger unit should be regarded as a new translation unit. In practice, the reader may feel sufficiently sure that he/she knows what the translation equivalents of the A+N phrases are. In other words, they may be aware that *personal representatives* is *遗产代理人* in Chinese, but may not be aware that *legal personal representative* is *合法遗产代理人*. They may ignore the fact that these phrases can sometimes form larger translation units. They will only be concerned with the larger translation units when they do not know how to translate the whole (e.g. cluster, phrase, segment etc.).

5. A+N Phrases With More Than One Translation Equivalent

Among all the ten A+N phrases, there are six phrases that have synonymous translation equivalents: *light bus, written permission, necessary modifications, reasonable ground, human remains* and *conclusive evidence*. Their translation equivalents are listed in the order of their frequency.

Table 4: The 6 A+N Phrases Whose Translation Equivalents are Synonymous.

A+N Phrase	1st TE/Freq.	2nd TE/Freq.	3rd TE/Freq.	4th TE/Freq.
light bus	小巴/31	小型巴士/22		
Written permission	书面准许/17	书面许可/7	书面批准/3	准许/3
Necessary modifications	必要的变通/20	必需的变通/7	需要的变通/2	必需的修改/1
Reasonable ground	合理的理由/16	合理理由/15		
Human remains	人类遗骸/41	遗骸/1		
conclusive evidence	确证/27	不可推翻的证据/5		

The two translation equivalents of *reasonable ground*, *合理的理由* and *合理理由*, are the two most obvious synonyms. The Chinese character *的* in *合理的理由* is used as adjective suffix, which can be and often is omitted to achieve concision. These two translation equivalents are actually one. They can, of course, replace each other.

The same is true for *light bus*. Both *小巴* and *小型巴士* are rendered from *light bus*. *小巴* is an abbreviated form of *小型巴士.*

人类遗骸 and *遗骸* from *human remains* are not synonyms if we consider them as two separate terms. However, this impression will disappear after a careful look at their context. There is only one case of *遗骸*, but the rest are all rendered as *人类遗骸.* The context where this case happens is given in the following sentences:

54740 Where a person who has the right to effect the disposal of the **human remains** *of any person-*
54741 within the period of 48 hours after the **human remains** *are received into any mortuary-*

*54740 如具有处置任何***人类遗骸***的权利的人—*
*54741 在殓房接收该***遗骸***后 48 小时的期限内—*

Sentence 54740 and 54741 belong to the same semantic sentence in the text, but they have been cut into two for the sake of alignment during the corpus processing. If we read them together as one part of a whole sentence, we find that the two *human remains* refer to the same object. The second *human remains* has been translated differently because of the Chinese character *该* before *遗骸* in sentence 54741. *该* means *such* or *this* in Chinese. *该遗骸* means *such/this remains,* which refers to the previously discussed human remains. Therefore, in this case, *遗骸* and *人类遗骸* share the same referential meaning because of the Chinese functional character *该*. In fact, the whole translation equivalent is not *遗骸* but *该遗骸*. *该遗骸* and *人类遗骸* refer to the same thing; they are synonymous in this case.

Although *written permission* has more equivalent variations, its translation equivalents are synonymous as well. The terms *准许*, *许可,* and *批准* are synonyms and they mean *permission*. In the first three translation equivalents, the word *written* has all been rendered as *书面.* The first three equivalents *书面准许*, *书面许可* and *书面批准* are synonymous. The fourth translation equivalent *准许* is actually the abbreviation of *书面准许* in this context, *书面* is omitted for the sake of the concision, but it can be deduced while reading the translated text.

In the four translation equivalents of *necessary modifications, modification* has been translated as the same *变通* except in one sentence as *修改*. The three variations translated from *necessary*, *必要的, 必需的* and *需要的,* are synonymous in Chinese. Therefore, the first three translation equivalents are synonymous. Since *修改* and *变通* are synonymous as well, the fourth translation is synonymous with the previous three.

Although the two translation versions of *conclusive evidence* cannot strictly be called synonyms by linguists, they are to some degree synonymous. The literal translation of *确证* is *factual evidence* while *不可推翻的证据* is *the impossible overthrown evidence*. The similarity of the two translation alternatives is that in both the evidence does exist, or is a fact or provides strong evidence. The difference is that the former Chinese translation focuses on the evidence, while the latter emphasises the impossibility of overthrowing the evidence. They do, in fact, share the same meaning but focus on different elements.

There are four A+N phrases that have non-synonymous translation equivalents: *long term, conclusive evidence, good order, medical officer, residential care*.

The two translation variations of *long term* are due to the different contexts. In fact, *long term* forms part of another two larger translation units –*long term business* and *long term interest.* In *long term interest, long term* is always translated as *长远*, while in *long term business,* as *长期.* The different translations are caused by the different collocations. *Long term* itself is not an independent unit of meaning; it has to accompany *business* or *interest* to form a unit.

Both *Good order* and *residential care* are parts of larger translation units and their translation equivalents cannot be identified without considering other words in the context. The following is the concordance of the thirty extracted instances of *good order*:

Figure 2: Concordance of *good order*:

```
1      60466 the maintenance of decency and [good order] in the stadium is prejudice
```

```
2  ner.     44679 maintenance of peace and [good order] in any place licensed under
3  s;       54311 maintenance of peace and [good order] in any place licensed under
4  ered, drained, lighted or maintained in [good order],the Building Authority-
5  sanitary condition and shall be kept in [good order] and repair.     56714 Every
6  g Authority, and shall be maintained in [good order] to his satisfaction, by the
7  nd sanitary condition and to be kept in [good order] and repair.     56977 Every
8  articles have been delivered but not in [good order] and condition, of the damag
9   in a clean condition and maintained in [good order] and repair.     57115 Every
10  in a clean condition and maintained in [good order] and repair.     58655 Every
11 icer, and shall deliver the articles in [good order] and condition, fair wear an
12 tion  or of maintaining such shoring in [good order] or of inspecting the same.
13 keep a public dance hall shall maintain [good order] in the premises and shall n
14 to keep a dancing school shall maintain [good order] in the premises and shall n
15-      58752 The licensee shall maintain [good order] on the licensed premises an
16-      58693 The licensee shall maintain [good order] on the licensed premises an
17  any stadium;     54566 preservation of [good order] and prevention of abuses an
18 he notice:     54111 the maintenance of [good order] in slaughterhouses;       5
19 nuisances;     54733 the maintenance of [good order] in public funeral halls.
20 ts of a detainee or in the interests of [good order] in the Centre that a detain
21 his Part;     54434 the preservation of [good order] and discipline and preventi
22 shall not interfere with the running or [good order] of the centre and is otherw
23 terest on the grounds of public safety, [good order] and security, the cost of t
24 n an offensive trade to be kept in such [good order], repair and condition as to
25 be kept clean and shall be kept in such [good order], repair and condition as to
26 be kept clean and shall be kept in such [good order], repair and condition as to
27 noxious matters, and to be kept in such [good order], repair and condition as to
28  noxious matters and to be kept in such [good order], repair and condition as to
29 ion on any problem which may affect the [good order] or discipline of the centre
30 person to do any act prejudicial to the [good order] and security of the centre.
```

According to the concordance in Figure 2, *good order* can have three different senses according to context:

1) *good order* is used to mean the good discipline of a place or premises. In this sense, if a verb such as *maintain* or *keep* or *affect* is used before it, *good order* is translated as *良好秩序* (1,2,3, 13, 14, 15, 16, 20, 22, 23, 29, and 30). If a noun rather than a verb is found before it, such as *maintenance* or *preservation*, then *good order* is translated as *秩序良好* (17, 18, 19, 21).

2) *good order* is used with *maintain* or *keep* to refer to the status of some object. If the words following it are *repair* or *condition*, *good order*, together with the verb, is translated as *保持完好* (5, 7, 9, 10, 24, 25, 26, 27 and 28). Without the words *repair* or *condition* following it, *good order* is translated as *妥善* (6, 8, and 14).

3). *good order* also means the property and sequence of certain articles. Usually, the preceding verb is *deliver*. It is translated into *性能良好* (10 and 13).

Then we find that there are five translation units, with their respective translation equivalents. All these extended translation units are shown in Table 5. Among their five Chinese translation equivalents, only the first one is nominal phrase. The second is adjectival phrase and others are verb phrases. Using a similar approach, *residential care* can be analysed from the concordance and the result is listed in Table 5 as well. In the first two translation units, *residential care* has been translated into *住宿照顾*. But the third translation unit *residential care home* has been translated as a whole --*安老院*.

Table 5: Translation Equivalents of *good order, residential care* and *long term.*

A+N Phrase	**Whole Translation Unit/Freq.**	**Chinese Equivalents**
good order	(keep/maintain)… good order (in some place)/12	(保持某处)…良好秩序
	(maintenance/preservation of) good order (in some place)/4	(保持某处)…秩序良好
	(something to be kept /maintained… in) good order (repair or condition)/9	(某物被保持)完好
	(maintain) in good order/3	妥善(保养)
	(be delivered in) good order (and condition)/2	(保持)性能(和状况)良好
residential care	residential care/1	住宿照顾
	residential care expenses/8	住宿照顾开支
	residential care home/34	安老院
long term	long term interest/34	长远
	long term business/2	长期

However, *公职医生* and *医生* from *medical officer* are more complicated. They refer to two different kinds of doctors. The translation of *公职医生* is encountered in Chapter 136 2(1), while translation equivalent *医生* is found in 298A 2. They are from different laws. Chapter 136 2(1) is the Interpretation part of the *MENTAL HEALTH ORDINANCE* which was issued on 1 February, 1999. Chapter 298A 2, however, is part of the *PROBATION OF OFFENDERS RULES*, which was issued on 30 June, 1997. The referential meanings are different in these two laws. One explanation is that in the English version of these two different laws, the same term *medical officer* has been used to refer to different concepts. When they are translated into Chinese, the translators purposely chose different Chinese terms to indicate their difference.

Table 6: Translation equivalents of *medical officer.*

Phrase	**Chinese Equivalent**	**Context**
medical officer	公职医生/18	In MENTAL HEALTH ORDINANCE
	医生/ 14	In PROBATION OF OFFENDERS RULES

From the above analysis, it can be seen that all these four A+N phrases need to be further expanded to yield complete translation units: to *long term* needs to be added to *business* or *interest*; *medical officer* needs to be added to its domain (the different laws they occur); *good order* and *residential care* are more complicated and have been expanded as listed in Table 5. Once a phrase has been expanded into large enough units, the ambiguity between its several equivalents disappears. The most complicated expansion of a translation unit is *good order*. It is important to note that not only adjacent words to the left or right of that phrase should be counted into the complete translation units, but also words with a little space in front of or behind the phrase will count as well, as in *good order*. Sometimes even the whole domain will be a factor which helps to disambiguate as, for example, with *medical officer*. Therefore, the different domain should be included in the complete translation units as well.

6. Conclusion

In this paper, I have demonstrated how to extract complete translation units based on thirty A+N phrases from HKLDC. The thirty A+N phrases have altogether been expanded into 43 complete translation units. This work has been done based on the following hypothesis. If the translation is consistent, a translation unit has only one translation equivalent; if a translation candidate has more than one translation equivalent, either of these equivalents is synonymous or the contexts are different and the candidates belong to different translation units. The candidate, accordingly, needs to be extended to larger units and into complete translation units, until their translation equivalents are unambiguous. This result can be useful in the automatic extraction of translation units and translation equivalents. This hypothesis has in turn been verified by the extracted translation units and their equivalents.

This is only a preliminary study that was able to analyse thirty typical A+N phrases based on the specialised HKLDC. Since the legal document belongs to LSP (Language for Specialised Purpose), this work may have generated some characters which a more general corpus may not have. The methodology and results should be tested with a larger scale general corpus studies.

Acknowledgement

This is a perversion of my MPhil thesis. I would like to thank Prof. Wolfgang Teubert for provide the corpus and supervision, and thank Dr. Pernilla Danielsson for the data extraction.

References

Chang, B., Danielsson, P. and Teubert, W. 2005 “Chinese-English Translation Database: Extracting Units of Translation from Parallel Texts”. In *Meaningful Texts*, G. Barnbrook et al (eds.), 131–40. London: Continuum.

Sinclair, J. M. 1991 *Corpus, Concordance and Collocation*. Oxford University Press.

Sinclair, J. M. 1996 “The Search for Units of Meaning”. *Textus IX*: 75–106.

Sinclair, J. M. 1998 “The Lexical Item”. In *Contrastive Lexical Semantics*, Weigand, E. (ed.), 1-24. Amsterdam: John Benjamins.

Sinclair, J. M. (Edited with R. Carter). 2004 *Trust the Text: Language, corpus and discourse*. London/New York: Routledge.

Teubert, W. 1996 “Comparable or Parallel Corpora?” *International Journal of Lexicography*. 9(3): 238–64

Teubert, W. 2001 “Corpus Linguistics and Lexicography”. *International Journal of Corpus Linguistics.* 6: 125–53.

Teubert, W. 2002 “The Role of Parallel Corpora in Translation and Multilinguial Lexicography”. In *Lexis In Contrast,* B. Altenberg and S. Granger (eds.), 189 – 214. Amsterdam: Benjamins.

Teubert, W. 2005 “My Version of Corpus Linguistics”. *International Journal of Corpus Linguistics*. 10(1): 1–13.

Wu, Dekai. 1995 “Grammarless Extraction of Phrasal Translation Examples from Parallel Texts”. In *TMI-96, Proceedings of Sixth International Conf. on Theoretical and Methodological Issues in Machina Translation*. Leuven, Belgium.

Zgusta, Ladislav. 1984 “Translational Equivalence in the Bilingual Dictionary”. In *LEXeter’s 83, Proceedings of International Conference on Lexicography at Exeter*, R.R.K.Hartmann (ed.), 147–54. Tübingen : Max Niemeyer.