Corpus linguistics and translation equivalence

Wolfgang Teubert
University of Birmingham
Email: teubertw@bham.ac.uk

The Hong Kong Legal Document Corpus (HKLDC)

- Statutory laws issued before 2001, in English
- Translated consistently into Chinese
- High in terminology
- Ca. 5.5 million words per language
- Ca. 200000 aligned sentences
- Aligned on the sentence level
- Chinese text is segmentised
- Chinese and English subcorpus is POS-tagged

The Hong Kong Legal Document Corpus (HKLDC): Shortcomings and advantages

- Not representative of general language
- Chinese version not mainland standard Chinese
- Much more consistent and uniform than normal translations

But

- Easy to align
- High consistency in translation
- Low in ‘noise’
- Good for testing methodology
- Ideal testbed

People and institutions involved

- Researchers: Wolfgang Teubert, Wang Weiqun (University of Birmingham); Sun Le (Chinese Academy of Sciences) (2003)
- Consultants: Feng Zhiwei (Beida, CUC), Chang Baobao (Beida), Ji Donghong (National University of Singapore)

Our assumptions (I)
What is wrong with bilingual dictionaries?

- Based on the single word in isolation
- Perhaps good for translating into native language
- Insufficient for translating into a non-native language
- Constrained by space
- Not enough instructions for ambiguity resolution
- Polysemy based on monolingual perspective
- Not taking into account the target language perspective
- Normally no entries for 'translation units'

Our assumptions (II)
What is wrong with machine translation?

- Based on single word translation equivalents
- Often working with the interlingua approach
- Language-neutral conceptual ontologies work only for standardised terminology
- No natural language ambiguity resolution

Our assumptions (III)
A look at translation practice

- Ambiguity is a problem of language description
- Readers have no problem with ambiguity
- Translators never translate word by word
- Translators translate text segment by text segment
- The text segments translated as a whole are not ambiguous from the target language perspective

Our assumptions (IV)
Another look at translation practice

- There is no ideal translation.
- Translation equivalence is created.
- It is the community of bilingual speakers who negotiate translation equivalents.
- Translators make mistakes.
- Acceptable translations of texts segments will be repeated; wrong translations won't.
- The community of translators know more than any bilingual dictionary.
- Parallel corpora are the repositories of the combined translation knowledge.

Our goals

- Extracting translation equivalence from parallel corpora
- Describing translation equivalence in such a way that the problem of ambiguity disappears
- Replacing the single word by the ‘translation unit’
- Setting up a database of translation units and their target language equivalents: The *TranslationBase*
- Using the *TranslationBase* for human translation and for machine translation

Defining the translation unit

- The translation unit is a text segment that is translated as a whole.
- We identify translation units in a parallel corpus by recurrence (i.e. as repeated events).
- The translation unit has only one meaning from the target language perspective.
- Therefore there is, for each translation unit, only one translation equivalent, or, if there are more, they are synonymous.
- A translation unit consists of a word plus all the words in its context that make the expression (the text segment) monosemous.

How to identify translation units in a parallel corpus

- We could search for statistically significant n-grams, but that does not tell us about their semantic relevance.
- Most translation units belong to a small list of syntactic patterns, such as adjective+noun, noun+noun, noun+*of*+noun etc.
- Frequency is essential: a minimum of three occurrences.
- Not all of them qualify as 'translation units': there is more than one non-synonymous translation equivalent.

How we extracted translation unit candidates from the HKLDC

- We searched, in the POS-tagged English version, all bigrams identified as adjective+noun.
- The result: 9000 phrases occurring at least three times.
- We selected 30 phrases occurring ca. 100 times each.
- For each phrase, we randomly selected ca. 30 citations (=sentences) for each phrase.
- We then identified the aligned sentences in the Chinese version of our corpus [sentence alignment].
- We then aligned the equivalent Chinese phrases with the English phrases [lexical alignment].

Extracted adjective+noun phrases

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

What makes an adjective+noun phrase a translation unit? Dictionary lookup (I)

- Example: *straight line* 直线 [zhi xian]
- Same translation for all occurrences
- Dictionary lookup (New English-Chinese Dictionary, Centenary Edition)
- Default translation of *straight*:直的 [zhi (de)]
- Default translation of *line*: 线 [xian]
- Default translation of *straight line*: 直线 [zhi xian]
- Is *straight line* a translation unit because it can be translated word by word? (Weiqun: No!)
- Cf.: only 直线 [zhi xian] is a mathematical term!

What makes an adjective+noun phrase a translation unit? Dictionary lookup (II)

- Example: *long term*:长远 [chang yuan] (36);长期的 [chang qi] (2)
- Same translation for most occurrences
- Dictionary lookup (New English-Chinese Dictionary, Centenary Edition) (NECD)
- Default translation of *long*: 长 [chang]
- Default translation of *term*:期的 [qi]
- Default translation of *long term*:长期的 [chang qi]
- Is *long term* is a translation unit ?

long term revisited: part of a larger unit

- 36 long term interest: always长远 [chang yuan]
- 2 long term business: always [chang qi]

Translation equivalent vs.NECD default translation: phrases not listed

A+N Phrase	Dictionary Default Translation	Corpus Translation
annual allowance	每年的允许/mei nian de yun xu	年积金/ nian ji jin
criminal offences	犯罪的冒犯/fan zui de mao fan	刑事罪行/ xing shi zui xing
final appeal	最后的上诉/zui hou de shang su	终审/ zhong shen
first column	第一的柱/di yi de zhu	第1栏/ di yi lan
public bus	公的公共汽车/gong de gong gong qi che	公共巴士/ gong gong ba shi
rateable value	可估价的价值/keg u jia de jia zhi	应课差饷租值/ ying ke cha xiang zu zhi
reasonable ground	合情合理的地/he qing he li de di	合理的理由/he li de li you
registered dentist	已登记的牙医/yi deng ji de ya yi	注册牙医/ zhu ce ya yi
registered scheme	已登记的计划/yi deng ji de ji hua	注册计划/ zhu ce ji hua

Translation equivalent vs.NECD default translation: phrases listed (subentries)

A+N Phrase	Dictionary Default Translation	Corpus Dominant Translation
long- term	长期的 [chang qi]	长远 [chang yuan] 36
internal combustion (engine)	内燃(机) [nei ran (ji)]	内燃(机) [nei ran (ji)]
medical officer	卫生官员 [wei sheng guan yuan]	公职医生 [gong zhi yi sheng] 18

(*internal combustion engine*, but not *internal combustion*, is a subentry: syntactic structure: adjective +noun+noun)

Translation equivalent vs.NECD default translation: phrases listed (examples)

A+N Phrase	Head-word	Dictionary Default Translation	CorpusDomiant Translation
conclusive evidence	con-clusive	确证 [que zheng]	确证 [que zheng]
listed company	list	上市公司[shang shi gong si]	上市公司 [shang shi gong si]
postal packet	packet	小件邮包[xiao jian you bao]	邮包 [you bao]

What makes an adjective+noun phrase a translation unit? One-to-one relationship

- A phrase is a translation unit when it cannot be translated by the default equivalents of its parts but must be translated as a whole.
- Translation units are unambiguous.
- A phrase is a translation unit if there is only one target language equivalent, or, in case there are more, these equivalents are strictly synonymous.
- If there is more than one equivalent for a phrase, then we have to search for other words in the context that make the phrase monosemous.

Phrases whose equivalents are synonymous: i.e. translation units (I)

- Example: written permission
- Equivalent 1:书面准许 [shu mian zhun xu] (17)
- Equivalent 2:书面许可 [shu mian xu ke] (7)
- Equivalent 3:书面批准 [shu mian pi zhun] (3)
- (Equivalent 4: 准许 [zhun xu] (3))
- The equivalents 1, 2, and 3 can be substituted for each other.
- (Equivalent 4 can be used if 书面 [shu mian] can be derived from the wider context.

Phrases whose equivalents are synonymous: i.e. translation units (II)

- Example: light bus
- Equivalent 1: 小巴 [xiao ba] (31)
- Equivalent 2: 小型巴士 [xiao xing ba shi] (22)
- 小 [xiao] is short form of小型 [xiao xing]
- 巴 [ba] is a short form of 巴士 [ba shi]
- Equivalent 1 is perhaps more colloquial than equivalent 2. But both equivalents are synonymous.

Phrases whose equivalents are synonymous: i.e translation units (III)

- Example: human remains
- Equivalent 1: 人类遗骸 [ren nei yi hai] (41)
- Equivalent 2: 遗骸 [yi hai] (1)
- 遗骸 [yi hai] means remains of plants, animals, people
- 人类 [ren nei] can be omitted if it can be derived from the wider context:

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—

Phrases whose equivalents are not synonymous: i.e no translation units (I)

- Example: conclusive evidence (1)
- Equivalent 1: 确证 [que zheng] (27) (‘factual evidence’)

5608 A certificate of the Official Receiver that a person has been appointed trustee under this Ordinance shall be **conclusive evidence** of his appointment.

5608 由破产管理署署长发出以证明某人已根据本条例获委任为受托人的证明书，即为该人获该项委任的确证

9768 14. A certificate signed by the Chief Executive of the Corporation that an instrument of the Corporation purporting to be made or issued by or on behalf of the Corporation was so made or issued shall be **conclusive evidence** of that fact.

9768 14. 一份由公司总裁签署的证明书，证明一份看来是由公司或代公司订立或发出的文书是由公司或代公司订立或发出者，即为该事实的确证。

Phrases whose equivalents are not synonymous: i.e no translation units (II)

- Example: conclusive evidence (2)
- Equivalent 2: 不可推翻的证据 [bu ke tui fan de zheng ju] (5) (‘evidence impossible to overthrow’)
- 8375 In an action for libel or slander in which the question whether a person did or did not commit a criminal offence is relevant to an issue arising in the action, proof that, at the time when that issue falls to be determined, that person stands convicted of that offence shall be **conclusive evidence** that he committed that offence; and his conviction thereof shall be admissible in evidence accordingly.
- 8375 在任何永久形式诽谤或短暂形式诽谤的诉讼中，如某人有否犯某刑事罪行的问题与在该诉讼中出现的争论点有关联，而在对该争论点予以裁定的时间，有证明该人仍就该罪行被定罪，则该证明即为他曾犯该罪行的**不可推翻的证据**，而他就该罪行的定罪亦据此可接纳为证据。
- 不可推翻的证据 in the context of: *offence, proceedings, criminal* etc. (criminal justice)

Phrases whose equivalents are not synonymous: i.e no translation units (III)

Example: good order

- Equivalent 1: 良好秩序 [liang hao zhi xu] (12)
- Equivalent 2: (保持)完好 [(bao chi) wan hao] (9)
- Equivalent 3: 秩序良好 [zhi xu liang hao] (5)
- Equivalent 4: 妥善(保养) [tuo shan (bao yang)] (3)
- Equivalent 5: 性能…良好 [xing neng… liang hao] (2)

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
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
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
15- 58752 The licensee shall maintain [good order] on the licensed premises 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
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.

Phrases whose equivalents are not synonymous: i.e no translation units (IV)

good order: 良好秩序 [liang hao zhi xu] (12) ('maintaining the good discipline of a place')

58693 The licensee shall maintain [good order] on the licensed premises and shall not suffer or permit thereon-

58693 持牌人须使领有牌照的处所保持良好秩序，同时不得容受或准许该处所内有—

46306 Where in the opinion of the Superintendent, it is desirable either in the interests of a detainee or in the interests of [good order] in the Centre that a detainee should be separately confined, he may be so confined by order of the Superintendent:

46306 如监督认为，不论是为了被羁留者的利益或是为了中心的良好秩序，该被羁留者均应隔离拘禁，则监督可下令将该人如此拘禁：

Phrases whose equivalents are not synonymous: i.e no translation units (V)

good order :保持完好［bao chi wan hao］(12) ('good repair')

sanitary condition and shall be kept in [good order] and repair. 56714 Every
nd sanitary condition and to be kept in [good order] and repair. 56977 Every
in a clean condition and maintained in [good order] and repair. 58655 Every
n an offensive trade to be kept in such [good order], repair and condition as to
be kept clean and shall be kept in such [good order], repair and condition as to
noxious matters, and to be kept in such [good order], repair and condition as to

Phrases whose equivalents are not synonymous: i.e no translation units (VI)

good order : 妥善 [tuo shan] (3) ('maintain in good order; good order and condition ')

56447 The walls, floors, doors, ceilings, woodwork and all other parts of the structure of every food room shall be kept clean and shall be kept in such [good order], repair and condition as to-

56447 每间食物室的墙壁、地面、门、天花板、木建部分及结构的所有其它部分均须保持清洁，以及保持完好、维修妥善和状况良好，以—

49658 Where any private street or access road is not so surfaced, channelled, sewered, drained, lighted or maintained in [good order],the Building Authority-

49658 凡任何私家街道或通路未有如此铺设路面、敷设渠道、污水渠及排水渠、加以照明或妥善保养—

Phrases whose equivalents are not synonymous: i.e no translation units (VII)

A+N Phrase	Whole Translation Unit	Chinese Equivalents/Pinyin/Freq.
good order	(keep/maintain)… good order (in some place)/12	(保持某处)…良好秩序 [(bao chi mou chu) liang hao zhi xu] (12)
	(maintenance/preservation of) good order (in some place)/4	(保持某处)…秩序良好 [(bao chi mou chu) zhi xu liang hao] (5)
	(something to be kept /maintained… in) good order (repair or condition)/9	(某物被保持)完好 [(mou wu bei bao chi) wan hao] (9)
	(maintain) in good order/3	妥善(保养)[tuo shan (bao yang)] (3)
	(be delivered in) good order (and condition)/2	(保持)性能(和状况)良好 [(bao chi) xing neng (he zhuang kuang) liang hao] 2

Phrases which are a part of a translation unit

- ‘residential care’ by itself (1):住宿照顾 [zhu su zhao gu]
- ‘residential care expenses’ (*住宿照顾开支* [*zhu su zhao gu kai zhi*]) (8)
- ‘residential care home’: 34 occurrences, translated as: 安老院 [an lao yuan]

English-Chinese Glossary of Legal Terms (ECGLT)

- published by the Law Drafting Division of the Department of Justice in Hong Kong
- web version of the *English-Chinese Glossary of Legal Terms* (ECGLT) is provided by the Bilingual Laws Information System (BLIS)
- updated by the Department of Justice of the HKSARG (The Government of Hong Kong Special Administrative Region of the People's Republic of China)
- www.justice.gov.hk/eng/glossary/homeglos.htm

雙語法例資料系統
Bilingual Laws Information System

The English-Chinese Glossary of Legal Terms

workplace

Find Find headwords only

Search Search all columns for all English and Chinese Expressions

Please click on the words in blue to view chapter and section numbers.

Previous List | Next List | Expand All | Collapse All

English Expression		Chinese Expression	Chapter	Section
workplace		工作地方	313B	2
workplace		工作地点	406	30(6)
workplace	notifiable ~	应呈报工场	59	2(1)
works		工程	265	2
works		工程设施	265	2
works		工厂	51	8(2)(a)(iv)
works	capital ~	基本工程	101	37(3)(a)(i)
works	construction ~	建造工程	203	2
works	~ area	工程范围	374Q	2
World Trade Organization Agreement		世界贸易组织协议	43	13A(7)(a)(iii)
worldwide				
worldwide	~ merger	全球性合并	1138	弁言

How good is the ECGLT?

- provides correct translation equivalents for only 18 out of 30 adjective+noun phrases
- is still considerably better than a general language dictionary
- is linked to the bilingual law database, which greatly improves the convenience of consultation
- but there are still 40% phrases which cannot be found in the ECGLT

How has the ECGLT been produced?

- The ECGLT is not completely corpus-based.
- 27% phrases of the 30 adjective+noun phrases cannot be found in ECGLT at all.
- Some of the collocations are not listed under the relevant headwords.
- The ECGLT sometimes fails to provide the dominant HKLDC equivalent.
- Sometimes the ECGLT provides more equivalents of a translation unit than there are in the corpus.

Conclusions (I)

- It is possible to automatically extract phrases representing syntactic patterns from a parallel corpus, e.g. adjective+noun phrases.
- We can regard these phrases as (unambiguous) translation equivalent candidates.
- Once lexical alignment is carried out, we know if there is only one or if there are more target language equivalents.
- Lexical alignment can be carried out increasingly automatically.
- If there is more than one equivalent: Are these equivalents synonymous or not? (Manual intervention needed.)

Conclusions (II)

- If there is more than one non-synonymous equivalent: Our translation unit candidate has to be expanded (e.g. *internal combustion* ~ *internal combustion engine*; *good order* ~ *good order and repair*).
- Translation unit candidate expansion can be done largely automatically. Minimal frequencies apply.
- Result: List of momosemous source language translation units and their target language equivalents.
- Once there is a one-to-one relationship between translation unit and equivalent, the relationship is reversible.

Conclusions (III)

- A TranslationBase is a database containing unambiguous translation units and their target language equivalents.
- A TranslationBase is reversible.
- A TranslationBase enables translation free of ambiguity errors.
- A TranslationBase can be used for human and for machine translation.
- TranslationBases can be compiled largely automatically.
- TranslationBases are superior to bilingual dictionaries and to MT lexicons based on conceptual ontologies

Conclusions (IV)

- Parallel corpora are the material evidence of translation equivalence.
- The solution to the ambiguity problem in translation is the language knowledge contained in parallel corpora.
- Parallel corpora contain the practice of many experienced translators.
- A TranslationBase is the true expression of translation equivalence.