# Analysis and Annotation of English-Chinese Financial Terms for Benchmarking and Language Processing

**Oi Yee Kwong**
Department of Translation
The Chinese University of Hong Kong
oykwong@arts.cuhk.edu.hk

**Abstract**

This paper reports on our ongoing work in annotating bilingual (English-Chinese) terminology in the financial domain. Arising from a larger project to produce a benchmarking dataset for evaluating term extraction tools, the resulting language resource is expected to be of use in financial narratives processing. The study will make available a gold standard to translators and researchers, especially for the former to leverage in the evaluation of commercial term extraction tools. To accommodate the diverse interests of both end users and researchers from multiple perspectives, an analysis of existing terminological resources for translators was done. Based on the linguistic properties observed, a set of term annotation guidelines was formulated for marking up bilingual financial terms in a corpus, for systematic selection of terms according to various criteria and expectation. The resulting dataset will fill the gap for term extractor evaluation for which bilingual data are lacking, and serve as a shared and transparent evaluation standard to help enhance the mutual understanding between computational terminologists and translators. In a wider context of natural language processing, bilingual terms extracted from a variety of financial texts are anticipated to be of help for information mining especially from regularly updated and structurally repetitive documents.

**Keywords:** financial terminology, bilingual term extraction, benchmarking dataset, term annotation

## 1. Introduction

Terminology management is a core function of computer-aided translation, and automatic term extraction is a commonly affiliated component of it. Off-the-shelf commercial term extraction systems available to translators, as products of software development instead of prototypes in academic research, are often packaged with many user-friendly features, but the opaque operation mechanism (or the algorithm) unknown to the users often leaves them with unrealistic expectation and misunderstanding. Such software tools thus do not always invite positive comments when they enter into the market. This is especially evident when a tool claims to be language independent, but when it is applied to a distant language pair, such as English and Chinese, it often turns out to perform much more poorly than its promotional demo may show, for instance, between English and French. User feedback found in classrooms and on the internet often echoes this observation. The situation is therefore not quite in concord with the many encouraging results reported from studies on automatic term extraction all along. Such an undesirable scenario, to a certain extent, is the result of little mutual understanding between computational terminologists and translators, and the lack of a shared and transparent evaluation standard. These issues merit attention, reflection, and resolution.

The project starts with term annotation in the financial domain for various reasons. Despite the availability of English-Chinese bilingual financial glossaries from various sources, like those mentioned in Section 4.1 below, they might not be ideal for benchmarking in term extraction especially for translators, as they are not tailored for specific sets of source and target texts that come with the translators as end users of the term extraction systems. Second, the bilingual terms in existing glossaries may not cover the precise translation equivalents used in the bilingual texts of specific companies and organisations (such as different banks). Third, the "termhood" as considered in existing glossaries is mostly from domain practitioners' perspective, which may or may not be the same as that from terminologists and translators. In addition, financial terminology is particularly crucial in the Hong Kong context given its long established status as one of the world's leading financial centres, and its continuing economic development with increasingly closer relations with Mainland China. New concepts and thus new terms keep emerging, and their timely and accurate mining would be of great help for professional translators.

## 2. Users' Concerns and Expectations

As end users of commercial term extraction systems, most translators could only judge a system by their own first-hand experience. Of most relevance to them is probably the preparation of bilingual input text for the extraction process and the validation of the term candidates subsequently returned by the system.

### 2.1. Input Text

As Blancafort et al. (2013) pointed out, computer-assisted translation suffers from the terminology bottleneck, and bilingual terminologies generated with statistical machine translation toolkits require parallel corpora. Scarcity of domain-specific parallel corpora is the major hindrance. As they noted, commercial tools only handle parallel corpora. Even though most off-the-shelf tools nowadays often allow a variety of file formats such as .rtf, .doc, .pdf and .xml, in addition to the conventional plain text files, it is impractical to expect a professional translator or translation student to supply a very large corpus, not even for monolingual materials, which may only be more readily available to computational linguists. Nevertheless, the extraction performance

would depend on the size of the input text to a certain extent, especially if the extraction is statistically based, although the actual impact is often opaque to the user. For bilingual term extraction, users often have extra work in preparing bilingually aligned text for input. The aligned text will have to be in a specific format such as the Translation Memory eXchange (.tmx) format. For instance, bilingual texts in other formats like .doc will only be treated as monolingual ones by SDL MultiTerm Extract. To this end, the general users might need to make use of other tools (e.g. SDL Trados) to do the alignment and save/export the aligned sentences in the required format beforehand, which is a time-consuming step, while the more automatic alignment toolkits like GIZA++ (Och and Ney, 2003) would simply be deterrent to them. The extra pre-processing steps thus tend to keep users away, especially those who are less comfortable with computer-aided work.

### 2.2. Extraction Algorithm

Term extraction approaches are generally categorised as linguistic (e.g. Bourigault, 1992), statistical (e.g. Daille and Morin, 2005), or hybrid (e.g. Daille, 1996; Drouin, 2003). For bilingual term extraction, parallel corpora would be most preferred, but given the scarcity of parallel corpora, often it might have to make do with comparable corpora (e.g. Laroche and Langlais, 2010). The TTC platform, for instance, provides a whole pipeline of tools for terminology mining from comparable corpora for seven languages, including English and Chinese (Blancafort et al., 2013). Terms are separately extracted from monolingual corpora first and then bilingually aligned based on context of occurrence and compositionality (Daille, 2012).

To most translators as end users, the extraction step in commercial systems is just a click of button. The details of the algorithm used by a particular tool are usually unknown to them, that is, a black box. Although many a time users might be told that a certain tool makes use of a statistical algorithm to come up with the term candidates, which is almost the norm of modern term extractors, the algorithms adopted in individual systems could have different degrees of sophistication, and this is often at least partially disclosed from the results they generate. The more computer-literate users are often able to get a clue from the output to reverse engineer the mechanism by which the tools work. For example, some tools relying primarily on simple n-gram frequencies without paying much attention to linguistic validity (e.g. phrasal structures) may output incomplete or ungrammatical word strings among the suggested terms. Users' evaluation of the systems is usually impressionistic, based on their overall experience with the user interface, functionalities, effort needed for validation, and compatibility with their working translation environment, amongst other criteria. For example, Xu and Sharoff (2014) evaluated various term extraction tools working on comparable corpora, and although their performance was at most mediocre, especially on Chinese, student interpreters still found the low precision tolerable.

## 3. Modes of Evaluation

It is this last point regarding evaluation measures that we find mostly responsible for the gap between users and researchers. For computational terminologists, apart from qualitative comparisons among term extraction systems (e.g. Cabré et al., 2001), evaluation may also rely on human judges to go through the system-generated term candidate list (e.g. Fulford, 2001) or compare the term list against an existing term bank (e.g. Drouin, 2003). Nowadays it is often preferred to have system performance to be objectively measured with reference to some benchmarking data, by precision and recall, as is popularly done for many other natural language processing tasks. However, the reliability and validity of such quantitative measures are based on the assumption that a clear task definition exists. As noted by Bernier-Colborne and Drouin (2014): "Whereas other natural language processing tasks have well-defined evaluation schemes and benchmarks, the question of how to evaluate TEs [term extractors] remains unresolved. Evaluations are regularly reported in work on term extraction, yet the methodology varies from one work to the next, such that comparisons are hard to establish." (p.51) For the term extraction task to be well-defined, one must state precisely what counts as a term and thus which expressions should or should not be extracted. In fact, given the different backgrounds and expectations, evaluation criteria also vary, and many reference standards in different studies may only be ad hoc term lists drawn up without adequate systematic control. It is apparent that a translator's evaluation of a term extraction tool usually has no reference to benchmarking data. For instance, in a preliminary study on terms extracted by SDL MultiTerm Extract from a small amount of financial texts (from the annual reports of a bank), we compared the so-called unmatched items against a translation student's so-called "gold standard". It was observed that among the "noise", many are grammatical linguistic expressions (e.g. noun phrases, verb phrases, prepositional phrases) and some are obviously genuine financial terms; and among the expected but unmatched items, some are actually partially extracted already while others might be considered semi-technical terms (Kwong, in press). While translators' concerns are multifarious, including but not limited to system performance (e.g. accuracy of candidates), software design (e.g. user-friendliness), and very importantly compatibility with their own expectation, which lead to an overall perception of a system, access to a gold standard is nevertheless desirable for end users. After all, the validation process is often the decisive factor for whether a translator will find the term extractor a help or a nuisance.

## 4. Toward a Gold Standard for Bilingual Term Extraction

Defining the gold standard for term extraction can sometimes be tricky. The main problem has to do with spelling out the criteria for the selection and annotation of terms systematically. As Vivaldi and Rodríguez (2007) remarked, "there is low agreement between terminologists and domain experts on what term candidates should be treated as

terms" (p.244). Estopà (2001) also reported great difference in the type and number of terms manually selected by terminologists, domain experts, translators and information scientists. Hence one important element that any gold standard for term extraction should consider is the diverse expectations from different stakeholders.

To reconcile the considerably varied interests and concerns, a better understanding and thus consensus of the distinction between terms and non-terms in any given domain is most important. Terminology and phraseology should be sufficiently distinguished as far as practicable. Measures should be established to enable us to better define and differentiate along the gradation from common expressions to core technical terms for a certain domain. This calls for a more thorough analysis of a whole range of expressions deemed important by translators for a specialised domain, based on well-defined criteria, linguistic or otherwise. Secondly, instead of just black-box testing according to their own subjective judgement, translators should also be entitled to more objective testing on their side, as much as computational terminologists. A gold standard based on a term-annotated corpus, which is obviously lacking for bilingual English-Chinese terminology, will therefore be necessary. Such a benchmarking dataset should annotate a full range of expressions deemed relevant by translators, as well as terminologists and computational linguists, with the type and domain specificity indicated. Moreover, researchers need to re-consider the corresponding linguistic and statistical criteria in automatic term extraction, based on the linguistic description of domain-specific terms, to accommodate a more comprehensive set of concepts and their expressions. The relevance and applicability of compositional approaches need to be studied in more details, and new approaches need to be devised for the non-compositional cases. Testing and evaluation of newly developed tools, of course, should refer to the benchmarking data available.

### 4.1. Bilingual Term Analysis

Linguistic insights are often helpful for extracting terms from one language as well as identifying translation equivalents in another language. Syntactic structures, variant forms and compositionality are particularly relevant considerations (e.g. Baldwin and Tanaka, 2004; Hippisley et al., 2005; Daille, 2005; Bartels and Speelman, 2014). Sometimes regional variation may also be an issue.

We first collected various existing bilingual financial glossaries or term lists available in Hong Kong. These resources include: terms listed in two textbooks on financial translation[1] and the glossary from the Education Bureau for secondary school education[2], as well as glossaries from the Hong Kong Exchanges and Clearing Ltd[3], the Securities and Futures Commission[4], and the Hong Kong Monetory Authority[5]. The data sizes range from a few hundred to over 10,000 term pairs. While the more official glossaries are obviously more comprehensive than the lists given in textbooks, a collection like this can reveal a broader spectrum of what different groups of people, including translators, educators and domain experts, view the nature and scope of terminology in the financial domain.

Samples of English-Chinese term pairs were selected from the various sources and analysed with respect to the following aspects[6]:

- Word classes (e.g. nominal or verbal): As expected, the majority of the terms are nominal. In general, there is less than 1% of the terms in our samples which are not nominal. It was nevertheless observed that the Chinese equivalents are not necessarily in the same word class as the English terms (e.g. "dilution" is a noun while its Chinese equivalent 攤薄 is apparently verbal), although the vast majority of the pairs do have compatible word classes (e.g. "brokerage" 券商).
- Constituent compatibility: The lexicalisation of concepts could be quite different between English and Chinese, although most term pairs are in fact multi-word English and Chinese expressions, as illustrated in Table 1.

| #words (E:C) | % | Examples |
|---|---|---|
| 1:1 | 7.73 | underwriter 包銷商<br>volatility 波幅 |
| 1:N | 12.84 | exclusions 不受保項目<br>jumbomize 將股票化零為整 |
| N:1 | 8.84 | resumption of trading 復牌<br>bad and doubtful debts 呆壞帳 |
| N:N | 70.59 | backdoor listing 借殼上市<br>rateable value 應課差餉租值 |

Table 1: Lexicalisation among E-C Bilingual Term Pairs

- Syntactic structures (for multi-word English terms): As also suggested by Table 1, about 80% of the selected English terms are multi-word expressions. Considering the nominal expressions, the majority take a modifier-head structure, while some have post-modifiers and others have both pre- and post-modifiers, as shown in Table 2.
- Compositionality (for multi-word Chinese equivalents): For the multi-word English terms, it is important to see whether the corresponding Chinese equivalents are also formed compositionally. For example, the Chinese equivalent for "interim dividend" can be compositionally formed as 中期/interim

[1]周兆祥、范志偉 (2004) 《財經翻譯精要》香港：商務印書館 and 李德鳳 (2007) 《財經金融翻譯：闡釋與實踐》香港大學出版社 (Both books are written in Chinese.)

[2]http://www.edb.gov.hk/attachment/en/curriculum-development/kla/technology-edu/whats-new/bafs_glossary_071130.pdf

[3]http://www.hkex.com.hk/eng/global/documents/glossary_ec.pdf

[4]http://www.sfc.hk/web/doc/EN/inutilbar/glossary/2006/full_list.pdf

[5]http://www.hkma.gov.hk/gdbook/chi/main/index_c.shtml

[6]The quantitative figures reported here are based on some 1,900 term pairs selected from two of the above sources.

| Structure | % | Examples |
|---|---|---|
| Head-Mod | 5.17 | merger by absorption, statement of capital |
| Mod-Head | 88.74 | accrued expenses, bare trustee |
| Mod-Head-Mod | 4.50 | straightline method of depreciation, carrying amount of an asset |
| Others | 1.59 | stores and spares, delivery vs payment |

Table 2: Syntactic Structures of English Multi-word Terms

股息/dividend, while the correspondence is not as straightforward between "initial margin requirement" and 開倉保證金, or between "evening evaluations" and 最後收盤價. Nevertheless, at least two-third of the samples are the compositional type.

## 4.2. Devising Annotation Guidelines

Guidelines are then to be drawn up to explain the what and how for selecting bilingual terms from a corpus for our gold standard. In addition to the setting-based criteria, linguistic criteria, and formal criteria discussed in Bernier-Colborne and Drouin (2014) governing term selection for their test corpus on automotive engineering, we include further considerations. First, more attention will be paid to translators' expectations, which might include not only terms that are likely to be found in specialised glossaries but also other semi-technical expressions. Second, the linguistic criteria will be adjusted with the observations from our own analysis in the project, especially noting the variety of syntactic structures, term variants and compositionality relevant to English-Chinese financial terms.

### 4.2.1. Scope of Terms

Bernier-Colborne and Drouin's (2014) setting-based criteria took on relatively strict terminological considerations. Given the domain of automotive engineering, only expressions denoting tangible and intangible objects or products directly related to the understanding of the subject matter are considered valid terms, while those referring to more generally associated entities are excluded.

*Our consideration*: Given the nature of the financial domain, the concepts are possibly more intangible than tangible (e.g. "insurance" 保險 is a relatively intangible concept), and very often involve processes (e.g. "closing transaction" 平倉交易) other than objects. In addition, from the perspectives of translators and domain experts, the boundary between terms and non-terms is more fuzzy than that as viewed by terminologists. For instance, organisation names are often included in professional glossaries, while common fixed phrases are found in translators' term lists. Hence it is less easy to limit the scope to "pure" terms, and that would not meet most users' expectation. For our purpose, the scope of terms is therefore less restricted, and annotators are asked to distinguish among four types of expressions. Type A will contain the core financial terms, for which expressions (single-word or multi-word) could be selected as long as they carry self-contained domain-specific meanings relevant to the financial context. Examples are "profit before tax" 除稅前利潤, "derivatives" 衍生工具, etc. Type B will be the semi-relevant expressions, such as "Board of Directors" 董事會, which is not only relevant to banking and finance, but also to a wider business context or even other non-commercial settings. Type C may include frequent phrases and jargon which appeal to translators as warranting specific translations. However, these expressions, despite being found frequently in financial documents, are not necessarily terms. The inclusion of this type inevitably loosens the scope of terms to a large extent, but it also allows us to accommodate broader views and to make better distinction for various situations. For example, the phrase "top and emerging risks" 首要及新浮現風險 has appeared more than 10 times in the annual report of one particular bank, but "top" and "emerging" do not refer to any intrinsic quality of "risks" or indicate a specific kind of risks (unlike "credit risk" 信貸風險, for instance), so the phrase should not be considered a term. But naturally translators are tempted to include it in their glossary to facilitate translation. Annotators are thus not to go simply by frequency, as a less frequent expression like "mitigating action" 緩減行動 may have a more specific meaning (in risk management for this example) and thus merit a different categorisation. The fourth type of expressions, which does not need to be annotated, contains general words and phrases like "location" 所在地, "information" 資料, "charitable donations" 慈善捐款, etc. They are not considered for the current purpose, whether or not a translator finds them difficult or special.

### 4.2.2. Form of Terms

The linguistic criteria in Bernier-Colborne and Drouin (2014) stipulate that only nouns and noun phrases should be annotated. Base terms and variants are included but distinguished. Following L'Homme (2004), morphologically related forms of a selected term and expressions which are paradigmatically related to a selected term could also be included. There was no restriction on compositionality, but syntactic variants are not considered.

*Our consideration*: As observed in the analysis above, although most terms listed in translation textbooks and professional glossaries are nominal, a small amount of verbal terms are also found. In fact, it is not unusual for the processes taking place in the financial domain to be expressed by verbs. Hence our annotation is not limited to nouns and noun phrases, but also covers verbs and verb phrases as appropriate. For example, while "underwriter" 包銷商 would be found in many glossaries, the verb "underwrite" 承辦 may not be a less important term especially in real financial corpora. In addition, we do not really distinguish base terms and variants, as our analysis shows that some terms do have post-modifiers and they are not exactly variants of a base form (e.g. "statement of capital" 股本說明 is not really the variant of "capital statement").

### 4.2.3. Span of Terms

According to their formal criteria, no limit was placed by Bernier-Colborne and Drouin (2014) on the length of terms as they appear in the corpus, and only maximum-length

terms were to be annotated as far as they fulfilled the other criteria. Otherwise shorter terms embedded therein and satisfying all term selection criteria could be annotated.
*Our consideration*: We basically also follow the maximum-length principle. We do not ask annotators to consider the shorter embedded terms, for if such shorter terms appear elsewhere in the corpus, they would be selected anyway. For example, where "financial system abuse risks" 金融系統濫用風險 is the longest term found in a context, there is no need to annotate "financial system" and "risks" in the same context. "Financial system" would have been annotated in another context when it stands on its own, such as "... to the stability and effective working of the financial system of Hong Kong", if the annotator chooses to mark it. One important point which is specific to our guidelines is the consideration of the Chinese equivalents. All along, the guidelines imply starting from the English text. Nevertheless, any English term fulfilling all selection criteria should only be annotated if a Chinese equivalent in its full form can be located from the corresponding Chinese text.

### 4.3. Term Annotation in Corpus

We started term annotation, based on the above principles, with the Annual Report and Accounts 2016 of the Hongkong and Shanghai Banking Corporation Limited (香港上海滙豐銀行有限公司2016年報及賬目), available in English and Chinese[7]. The corpus size of the various sections in the annual report is shown in Table 3.

| Section | English (words) | Chinese (chars) |
|---|---|---|
| Report of the Directors<br>董事會報告 | 3,917 | 6,749 |
| Financial Review<br>財務回顧 | 1,511 | 2,693 |
| Risk Report<br>風險報告 | 14,383 | 27,013 |
| Capital<br>資本 | 1,825 | 3,426 |
| Notes on the Financial Statements<br>財務報表附註 | 18,360 | 31,350 |
| Total | 39,996 | 71,231 |

Table 3: Corpus Size (HSBC Annual Report 2016)

#### 4.3.1. Preliminary Comparisons

For training, four annotators, all undergraduate translation students, were instructed with the guidelines and asked to mark up bilingual term pairs from the section on Report of the Directors. The results, at least quantitatively, turn out to be quite varied. As shown in Table 4, the first three annotators (AC, MY and JT) apparently form a more lenient group. The number of expressions they selected almost doubles the number of expressions selected by the fourth annotator (CC). Moreover, the distribution of types (A, B or C) looks very different across annotators. Such discord on number and classification reflects that despite being instructed with the same guidelines, individual annotators could still differ considerably in their perception of termhood as well as their understanding of the task requirements.

| Annotator | Type A | Type B | Type C | Total |
|---|---|---|---|---|
| AC | 86 | 52 | 4 | 142 |
| MY | 29 | 71 | 37 | 137 |
| JT | 37 | 62 | 39 | 138 |
| CC | 49 | 19 | 4 | 72 |

Table 4: Number of term pairs annotated

Ignoring the term classification for the time being, Table 5 gives a sketch of the agreement among the annotators. For a total of 246 distinct expressions selected, over half were actually selected by two or more annotators.

| Selected by | N (%) |
|---|---|
| 4 annotators | 28 (11.38%) |
| 3 annotators | 46 (18.70%) |
| 2 annotators | 67 (27.24%) |
| 1 annotator | 105 (42.68%) |

Table 5: Agreement among annotators

Expressions selected by all four annotators include some of the relatively standard financial terms and fixed phrases, although not as comprehensively as one might expect, and the classification is not always uniform. Some examples are shown below:

*Selected by all annotators:*

| | |
|---|---|
| financial statements | 財務報表 |
| material risk takers | 承受重大風險人員 |
| ordinary shares | 普通股 |
| risk appetite | 承受風險水平 |
| share capital | 股本 |
| subsidiary | 附屬公司 |
| terms of reference | 職權範圍 |
| trade corridors | 貿易走廊 |

It happens that there are more cases where the same expressions have not been selected by all annotators but only some of them. This is probably where the translators' (or users') expectation can be shown as a salient factor. Although the annotators have received some basic translation training, they are still students. The more capable ones or those with better language proficiency may choose to ignore the more common expressions and the straightforwardly compositional ones. As they may not really think of including such expressions if they are to keep their own glossaries manually, they may not expect or strongly wish them to be extracted by automatic term extraction systems either. Here are some examples:

[7] https://www.personal.hsbc.com.hk/1/2/hk/regulatory-disclosures

*Selected by 3 annotators:*

| | |
|---|---|
| Banking Ordinance | 銀行業條例 |
| base salary | 基本薪金 |
| consolidated profit | 綜合利潤 |
| debentures | 債券 |
| dividends | 股息 |
| material interest | 重大利益 |
| Nomination Committee | 提名委員會 |
| remuneration policy | 薪酬政策 |

*Selected by 2 annotators:*

| | |
|---|---|
| auditor | 核數師 |
| business strategy | 業務策略 |
| economic capital | 經濟資本 |
| funding structure | 資金架構 |
| priority growth markets | 優先發展市場 |
| risk environment | 風險環境 |
| shareholders | 股東 |
| transactions | 交易 |

Notwithstanding the maximum-length principle discussed in Section 4.2.3, it is also noted that individual annotators may from time to time tend to select phrases longer than necessary. For example, Annotator AC has selected "income tax and social security", while others tend to select "income tax" as one unit, and optionally with "social security" as a separate one.

#### 4.3.2. Implications on Training

The annotation work is in progress, and further analysis of the annotators' work is being done. Feedback is given to the annotators regularly, to gradually bring their understanding of the annotation guidelines closer to one another, especially regarding the selection principles as well as basic linguistic awareness. It is not intended, and in fact not possible, for uniform annotation from everyone, as individual variation is what we are interested in, to see how translators vary in their perception of terms in financial translation, and thus their expectation of automatic term extraction systems. On the other hand, notwithstanding individual differences, the annotators need to follow a similar practice in their selection of terms, and in the end we will take the majority of vote to arrive at the benchmarking data. In addition, the classification of term types may need to be verified with existing professional financial glossaries, instead of relying entirely on the annotators' judgement. After all, they are still novice translators.

Annotation of more bilingual financial texts has been planned, including but not limited to prospectuses, annual reports of listed companies, banking information, as well as government documents relating to economics and finance.

## 5. Potential NLP Applications

Although the primary objectives of the annotation at the outset mainly focus on the evaluation of automatic term extraction from a more translator-oriented perspective, the resulting bilingual terminology dataset is expected to be beneficial to the processing of financial narratives in the NLP context as well.

While the actual usage of the language resource in concrete applications has to await the completion of the resource on the one hand and further in-depth research on the other, a potential example is portrayed here for some preliminary idea of the benefits that the resource might offer in practice. Instead of a general financial glossary, the annotated term dataset is grounded on authentic texts from the financial domain. The annual reports, for instance, are known to follow a standard format and structure every year, with changes mostly on the numerical data and certain qualitative details. Hence, the terms and fixed expressions obtained from the annual report in a certain year could provide a useful resource for language processing systems to anchor at various parts of the reports in other years. Leveraging the structural and linguistic similarities, information mining from reports of different years for comparison should be facilitated. The fact that the resource is document-specific means that the differences in individual series of documents could be taken into account, especially when bilingual processing is concerned. For example, given the same financial terms in English, it has been observed that annual reports of different banks may use different Chinese equivalents. The annotated dataset could thus offer more than existing general glossaries for language processing systems to gauge important information from financial documents of different organisations with individual linguistic conventions.

## 6. Conclusion

This paper has presented our ongoing work which aims to develop a gold standard based on a term-annotated corpus, to offer a resource currently lacking for the development and evaluation of bilingual English-Chinese terminology extraction systems.

The production of the intended benchmarking dataset consists of two major tasks. The first is a thorough linguistic classification and analysis of single-word and multi-word terms (and term pairs). The second task is, based on the linguistic analysis, to devise a set of term annotation guidelines and build up an annotated parallel corpus. Third, and more importantly, we are annotating bilingual terms in the corpus. While previous attempts have been for monolingual term annotation only, our term selection criteria cover different scenarios of English-Chinese correspondence.

Tools in translation technology are intended to assist translators, and the whole purpose will be defeated if translators fail to fully utilise and appreciate them. With a carefully cultivated benchmarking resource, we hope to enable a more translator-oriented perspective for the evaluation of automatic term extraction systems so that they can fulfill their roles in translation technology better and embrace more appreciation from their target users. As Agirre et al. (2000) stated, "... tools for translation cannot be satisfactorily designed without the cooperation of human translators" (p.296). Although our current work deals primarily with English-Chinese financial terms, the rationale and significance underlying a gold standard to accommodate multiple perspectives, including users and researchers, apply to terminology work on other languages and domains alike.

## 7. 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).

## 8. Bibliographical References

Agirre, E., Arregi, X., Artola, X., de Illarraza, A. D., Sarasola, K., and Soroa, A. (2000). A methodology for building translator-oriented dictionary systems. *Machine Translation*, 15:295–310.

Baldwin, T. and Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In *Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 24–31, Barcelona, Spain.

Bartels, A. and Speelman, D. (2014). Clustering for semantic purposes: Exploration of semantic similarity in a technical corpus. *Terminology*, 20(2):279–303.

Bernier-Colborne, G. and Drouin, P. (2014). Creating a test corpus for term extractors through term annotation. *Terminology*, 20(1):50–73.

Blancafort, H., Bouvier, F., Daille, B., Heid, U., and Ramm, A. (2013). TTC web platform: from corpus compilation to bilingual terminologies for MT and CAT tools. In *Proceedings of TRALOGY II*, Paris.

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING '92)*, pages 977–981, Nantes, France.

Cabré Castellví, M. T., Bagot, R. E., and Palatresi, J. V. (2001). Automatic term detection: A review of current systems. In Didier Bourigault, et al., editors, *Recent Advances in Computational Terminology*, pages 53–87. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Daille, B. and Morin, E. (2005). French-English terminology extraction from comparable corpora. In Robert Dale, et al., editors, *Natural Language Processing – IJCNLP 2005. Lecture Notes in Artificial Intelligence, Volume 3651*, pages 707–718. Springer-Verlag.

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans et al., editors, *The Balancing Act: Combining symbolic and statistical approaches to language*, pages 49–66. MIT Press, Cambridge, MA.

Daille, B. (2005). Variations and application-oriented terminology engineering. *Terminology*, 11(1):181–197.

Daille, B. (2012). Building bilingual terminologies from comparable corpora: TheTTC termsuite. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, pages 29–32, Istanbul, Turkey.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Estopà, R. (2001). Les unités de signification spécialisées: élargissant l'objet du travail en terminologie [units of specialised meaning: Broadening the scope of terminology work]. *Terminology*, 7(2):217–237.

Fulford, H. (2001). Exploring terms and their linguistic environment in text: A domain-independent approach to automated term extraction. *Terminology*, 7(2):259–279.

Hippisley, A. R., Cheng, D., and Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. *Natural Language Engineering*, 11(2):129–157.

Kwong, O. Y. (in press). Evaluating term extraction tools: System performance vs user perception. In Sin-Wai Chan, editor, *The Human Factor in Machine Translation*. Routledge.

Laroche, A. and Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 617–625, Beijing, China.

L'Homme, M.-C. (2004). *La terminologie: principes et techniques [Terminology: Principles and Techniques]*. Presses de l'Université de Montréal, Montréal.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Vivaldi, J. and Rodríguez, H. (2007). Evaluation of terms and term extraction systems: A practical approach. *Terminology*, 13(2):225–248.

Xu, R. and Sharoff, S. (2014). Evaluating term extraction methods for interpreters. In *Proceedings of the 4th International Workshop on Computational Terminology*, pages 86–93, Dublin, Ireland.