# From Business Corpus to Business Lexicon*

Li Lan *(eglilan@polyu.edu.hk) and* Grahame T. Bilbow
*(eggrabil@polyu.edu.hk)*, *Department of English, Hong Kong Polytechnic University, Hong Kong*

**Abstract:** Language corpora are now indispensable to dictionary compilation. They help broaden the role of the dictionary from standardizing the vocabulary to recording a language. The trilingual corpus generated by the Hong Kong Polytechnic University gives a record of business languages used in Hong Kong. It differs from other corpora in that (1) it includes English, Chinese and Japanese; (2) it shows local characteristics; and (3) it focuses on a specific area (financial services, including banking, accounting, auditing, insurance and investment). The paper discusses various issues of setting up a tricorpus, and how to make full use of the data to generate a trilingual lexicon.

**Keywords**: MULTILINGUAL, SPECIAL PURPOSE, CORPUS, LEXICON

**Opsomming: Van sakekorpus tot sakeleksikon.** Taalkorpora is tans onontbeerlik vir die samestelling van woordeboeke. Hulle help om die rol van die woordeboek uit te brei vanaf die standaardisering van die woordeskat tot die optekening van 'n taal. Die drietalige korpus wat deur die Hongkongse Politegniese Universiteit ontwikkel is, verskaf 'n opgawe van die saketale wat in Hongkong gebruik word. Dit verskil van ander korpora deurdat (1) dit Engels, Chinees and Japanees insluit; (2) dit plaaslike eienskappe vertoon; en (3) dit op 'n spesifieke gebied (finansiële dienste, insluitende bankwese, rekeningkunde, ouditering, versekering en belegging) fokus. Die artikel bespreek verskillende aspekte van die totstandbrenging van 'n drietalige korpus, en hoe om volle gebruik te maak van die data om 'n drietalige leksikon te genereer.

**Sleutelwoorde**: MEERTALIG, SPESIALE DOEL, KORPUS, LEKSIKON

## 1.    Introduction

There is a large variety of dictionaries to satisfy different users worldwide. Among them, English monolingual learner's dictionaries undoubtedly occupy one of the most competitive markets in the publishing world. In Hong Kong, as in many other ESL/EFL countries, the "Big Four" of British publishers have taken a large market share, not only with monolingual but also with bilingual versions. However, room still exists for culture specific dictionaries for local users. Tom McArthur (1998: 206) predicted that in the new millennium there

---

will be four trends in compiling dictionaries: globalization, localization, bilingualization and semi-bilingualization.

Hong Kong is one of the most important international financial centres in the world. According to government statistics, there are 172 internationally or locally licensed banks, 209 insurers, and 930 CPA (Certified Public Accountant) firms conducting business in Hong Kong. It is generally accepted that good communication is central to any successful business. Investigating business languages in Hong Kong enables one to form a clear picture of how languages function in business. The current research project, started in 1999, has the following objectives:

— to establish a trilingual corpus of business languages used in Hong Kong (English, Standard Written Chinese and Japanese),
— to compile a trilingual content-based lexicon, and
— to carry out linguistic research into and between languages in business in Hong Kong.

## 2. Generating the business corpus

The Business Lexicon started with the setting up of a business corpus. This procedure is not new, of course. Since the 1970s the field of dictionary making has been influenced by empirical and corpus-based methods. "All the major UK dictionary-publishers now have access to large and diverse corpus resources which provide the raw materials for a far more reliable description of English" (Rundell 1998: 315). Large corpora of general English, such as the British National Corpus (BNC) and the Bank of English, are regarded as standard corpora in that they include both spoken and written texts with a balanced mixture of different types. The texts are basically monolingual, produced by native English speakers only. These corpora are normally the result of a collaboration between linguists, lexicographers and computer scientists.
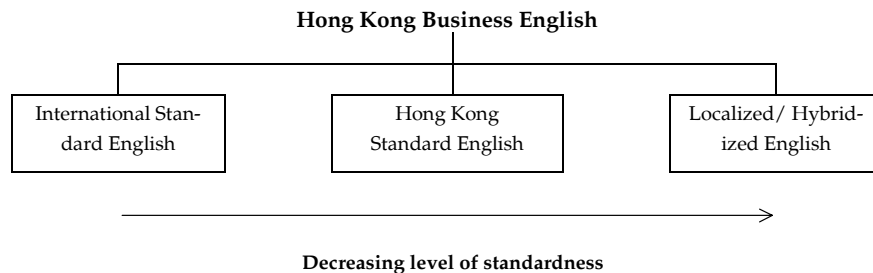
In addition to Standard English corpora, modern technology has also enabled the creation of many free selection corpora — involving guided text collection from the World Wide Web, newspapers and book CDs, and other machine-readable sources. Each corpus is unique in its own right and for its own purposes. Many free corpora are also multilingual and serve the purpose of keeping a record of languages in everyday use. The Hong Kong PolyU Business Corpus (PUBC) is a corpus which is compiled for special purposes in a specific region, and bears vocational, geographical, national and regional features.

## 2.1    Delimiting the languages of the tricorpus

The languages used in Hong Kong are commonly known as "two languages in three tongues". However, the extent to which they are used and the situations

in which they appear, vary considerably. English and modern standard Chinese are official languages used in formal situations, e.g. in law, in business, in education and by government institutions. Spoken local Chinese (Cantonese) functions as the spoken lingua franca among 97% of the ethnic Chinese population in Hong Kong. In any working environment, oral communication tends to be dominated by Cantonese or a mixed code (English/Cantonese) whereas formal internal and external written communication such as through letters, memoranda or notices tend to be in English. When working on the International Corpus of English in Hong Kong, Bolt and Bolton assumed that the particular range and quality of the Hong Kong data would be affected by a matrix of sociolinguistic relationships. They (Bolt and Bolton 1998: 198) found that "these relationships include those between the ethnic and linguistic background; between the local educational system and the linguistic community; and between the linguistic backgrounds of local English speakers and salient features of the type of English found in Hong Kong". Interestingly, they (Bolt and Bolton 1998: 212) comment that "more people than ever are speaking 'good English', and more people than ever are speaking 'bad English'".

Written English business texts in Hong Kong fall into three categories. The first involves internationally shared materials, such as financial news reports by Reuters, Associated Press, and Bloomberg. Within the branches of large international companies such as HSBC, AXA and AIA, there are also a fairly large number of shared documents (such as regulations and annual reports) from parent companies. The second type of text are those written locally by expatriates and Chinese educated in English-speaking countries, possessing a proficiency in English comparable to that of native speakers. The third type of text are those written by local employees for daily internal and external communication. This trajectory can be presented in the following chart.

**Hong Kong Business English**

| International Standard English | Hong Kong Standard English | Localized/ Hybridized English |
|---|---|---|

→

**Decreasing level of standardness**

The first block can arguably be regarded as an international standard, which provides norms for the local language users. "Hong Kong Standard English" refers to English with local norms accepted in the region. Users of HKSE commonly act as standardisers in companies producing publicly distributed texts in English. "Localized/Hybridized English" refers to a mixture of English and Cantonese. These three types are linked together by certain shared features;

there is no clear division between them. However, few would dispute the striking differences between the two ends of the spectrum. The data in the PUBC reflects the first two types, as all the collected texts are publicly distributed and demonstrate a high level of standardisation.

Chinese is another official language in Hong Kong. Although its written form is based on standard Putonghua from the Mainland, certain localized features can be clearly identified. Since the corpus is limited to languages used in Hong Kong, all the Chinese texts have been collected from local institutions. Apart from geographical and linguistic limits, there also exists a technical barrier. "Big5" is the coding system for complex Chinese characters used in Hong Kong and Taiwan, not compatible with the "GB" system for simplified characters used in Mainland China.

The language contained in the PUBC is Japanese. However, very little written Japanese appears to be used in Hong Kong. The majority of Japanese companies in Hong Kong tend to use English as medium of written communication. Unlike the varieties of English and Chinese, Japanese exhibits few formal differences at home and abroad, and very little variety appears to be acceptable. Therefore, in the PUBC, Japanese texts have been collected from resources in Japan as well as from Hong Kong. Since English has been made the second language in Tokyo, and the websites of big financial institutions in Japan tend to be bilingual, the temptation to include parallel texts (Japanese/English) in the tricorpus was great. However, this temptation was resisted for the following reasons. First, Japan has never been an English-speaking territory and in Japan English is almost entirely a "learnt" language. Second, the principal focus of the project has been defined as "English used in Hong Kong". The inclusion in the corpus of English from Japan might cause problems in its description. It is hoped that with texts from the same domain and of the same type such as business news, company reports and government policy statements, the Japanese texts will be thematically parallel to the English and Chinese texts.

## 2.2    Delimiting business language

It has been noticed that, in particular domains, the use of language is more predictably structured and subject to less ambiguity. Using this observation, computational linguists are now able to postulate the existence of special languages or "sublanguages". A sublanguage in the definition of Grishman and Kittredge (1986: ix) is a variety of language used in a given science or technology that is "not only much smaller than the whole language, but is also more clearly systematic in structure and meaning:

> ... [It is] a subsystem of language that behaves essentially like the whole language, while being limited in reference to a specific subject domain. In particu-

> lar, each sublanguage has a distinctive grammar, which can profitably be described and used to solve specific language-processing problems.

The concept of business can encompass many sectors. A business dictionary may cover terms concerning all business life, from office to stock exchange, and from international trade fair to classroom. With a limited budget, the project needed to be kept manageable. Since financial services are commonly viewed as the most significant for Hong Kong, business language has been limited to the language used in financial sectors, including banking, auditing, accounting, investment and insurance. During the process of data collection, the boundaries between the sectors were naturally found to be blurred: banks provide services of investment and insurance, accounting companies offer auditing and investment consultancy, and insurance companies cover investment services — therefore differences between individual sectors tend to be de-emphasised.

## 2.3    Defining text types

The design of the corpus focuses on a range of internal and external categories of business writing including business news, corporate annual reports, news releases, newsletters, minutes, posters, notices, leaflets, letters, faxes, memoranda and emails. We aimed at as even a distribution as possible between each of these types of texts. Our resources included the World Wide Web, printed materials from companies, government institutions and library on-line databases. The bilingual situation in Hong Kong has made it possible for us to get many parallel texts and thematically linked news reports.

We had much difficulty in obtaining internal administrative and regulatory writing and correspondence. The issue of confidentiality, as also found by Bolt and Bolton, made it very difficult to obtain texts of business transactions. The ICE (International Corpus of English) in Hong Kong settled on a broad conception of business transactions, using two-thirds business texts from the education sector (Bolt and Bolton 1996: 207). The fact that such texts exist but are unavailable to us, was a matter of some concern in obtaining a "balanced" corpus. However, to produce a business lexicon, it is not necessary to give priority to any particular text type. Oostdijk (1998: 169) argued that while general purpose corpora need to be balanced and represent a wide range of styles and registers, corpora representative of a single variety are already defined a priori on the basis of their specific domain of use and topic, "issues of corpus design only play a very minor role".

## 2.4    Setting up the database

The project started in June 1999 and 1.2 million words in each language have been collected, gleaned and put into three parallel subcorpora: English, Chinese and Japanese. The files are stored in ASCII format and the demographic

details of each file are kept in Microsoft ACCESS to create an index system for future retrieval. The records include thirteen items: Title, Author, Origin of the author, Text type, Source, Date, Language, Sector, Keywords, Length, Filename, Parafile name and Text.

## 3.    Towards the business lexicon

### 3.1    The target user group

In the competitive dictionary market, the key to success rests in shaping the dictionary to meet the needs of its users. We positioned our lexicon as a reference tool for the professional discourse community in Hong Kong, i.e. for those in the business sector and those planning to join this sector.

### 3.2    Vocabulary control

Vocabulary control is always a central concern in dictionary compilation. A context-based lexicon starts from a wordlist generated from a corpus. Of the three subcorpora, the English subcorpus was the base or starting point of the lexicon. Using WordSmith Tools (Oxford University Press), we produced a preliminary wordlist, giving the frequency of each word form. The 1.2 million word texts generated a list of 22 600 types, with a type/token ratio 1.80.

   It is clear that a wordlist from a concordance can in no way be used directly for the entries of a lexicon. Such a wordlist includes different forms of a word: word stem, inflected forms, run-on forms, derivatives, subderivatives and even nonsense data. For instance, the base word *act* produced 20 tokens.

| No. | Word | Freq. | % |
|-----|------|-------|---|
| 250 | act | 168 | 0.01 |
| 251 | acted | 30 | |
| 252 | acting | 133 | 0.01 |
| 253 | action | 213 | 0.02 |
| 254 | actions | 97 | |
| 255 | activate | 2 | |
| 256 | activated | 3 | |
| 257 | active | 169 | 0.01 |
| 258 | actively | 82 | |
| 259 | actives | 1 | |
| 260 | activism | 2 | |
| 261 | activist | 4 | |
| 262 | activists | 6 | |
| 263 | activities | 537 | 0.04 |
| 264 | activitity | 1 | |
| 265 | activity | 156 | 0.01 |

```
266    actor          2
267    actors         1
268    acts          31
268    actvities      1
```

This list needs thorough trimming. The word *act* should be approached carefully. Simple words in English are normally polysemic and problematic. Without tagging, there is no way of knowing how many of the occurrences are nouns, and how many verbs. *Act* can be a noun or a verb; *acts* can be a third person singular form or a plural form of a noun. We also need to classify this group into lexical morphemes and grammatical morphemes.

> Grammatical / inflectional morphemes:  acts, acted, acting/action, actions, …
> Lexical / derivational morphemes:  act, action, active, activate, actively, …
> Nonsense words:  activitity, actvities, …

Lexical morphemes are normally taken as entries in general dictionaries. For a special purpose dictionary, entries should be relevant to the subject area. More factors need to be taken into consideration. The words we finally put into the business lexicon as entries are *act, action, acting, active, activate, activity*.

### 3.3    New words or non-words

The PUBC has recorded several instances of recent language change: borrowing words from other languages; creating new words; using the names of people or places to refer to a related object; making shifts and conversions where meanings of words or their parts of speech change.

The introduction of modern technology in the business sector causes the emergence of many new words. Words with prefixes *cyber-*, *e-*, *elec-* and *i-* are now commonly used as proper nouns such as names of websites or publications, and many also shift to proper lexical words. The following are some new words and their occurrences in PUBC.

| Word | Freq. |
| --- | --- |
| cyberbanking | 10 |
| cyberbase | 8 |
| cybercash | 1 |
| cyberforce | 2 |
| cybermall | 1 |
| cyberport | 25 |
| cyberspace | 28 |
| cybertrading | 2 |
| cyberworks | 193 |

The word *dotcom* which appears six times in the corpus, is widely utilized by Internet companies. There are *dotcom connections, dotcom stores, dotcom consulting, dotcom marketing, dotcom industry* and so on. The following concordance is from the PUBC.

```
  But he said the fast-growing  dotcom  companies helped improve the
hnology stock prices soar and  dotcom  companies rush in to the mark
 in.    "No doubt some of our   dotcom  companies will do better than
   the new economy, we own no   dotcom  companies. What we've been lo
ong    STORY: WITH the raging   dotcom  fever, local investors are fr
 o switch from previously hot " dotcom " issues to quality technology
```

Another new word identified is *bancassurance*, which means an insurance service provided by a bank. It was first used as the name of a project or programme and has become a special term in business English. However, even though it appears eight times in the PUBC, there is little evidence to suggest it is sufficiently popular to be recorded in a dictionary.

As the project started immediately before the new millennium, it bears strong features of this period. The Y2K threat was a major concern then due to the fact that modern business relies heavily on computers. The word *Y2K* has a frequency of 380 in the PUBC, occurring in business news reports, company reports and internal and external documents. In contrast, no occurrence of *Y2K* is found in the BNC. We therefore have to consider carefully whether or not to include *Y2K* in the lexicon, for it may be more a general English word than a genre-specific technical term.

### 3.4    Statistical analysis

Although a frequency list provides basic statistical information from a corpus, we cannot use frequency as the sole criterion for the choice of words. The top 20 most frequent words in the PolyU Business Corpus are mainly grammatical words, very much similar to other English corpora. A comparison with the British National Corpus (BNC) is shown in Table 1.

Next in order are common lexical words which depend strongly on the design of the corpora, the topic and textual sources. Evidences of the register collected in the project can easily be seen from the list: the words *market, company, financial bank* and *insurance* appearing within the top 50 words, and occurring more frequently than grammatical words like *we, he, up* and *if.* There are also proper nouns with high frequencies, such as names of people, places, companies, and countries. The basic decision was to exclude grammar words, general English words and proper names. The names of important financial institutions and foreign currencies were put in appendices.

As many linguists have noticed, "a statistical model may not necessarily represent the use of a particular word in a particular context" (Oakes 1998: 43).

If a comparatively small corpus is used, frequency should not be the only criterion for word selection. For instance, although the word *millennium* appears 132 times, *hotel* 137 times and *university* 122 times, they are not relevant to a lexicon of financial terms. On the other hand, about 45% of the words in the list appear only once. Even with very low frequency, however, they can be highly relevant to the subject area. Although the words *debenture* and *abate* had two instances and *overdraw* had only one in the 1.2 million words, they are all recorded in the lexicon.

| | PUBC | | | | BNC | | |
|---|---|---|---|---|---|---|---|
| No. | Word | Freq. | % | No. | Word | Freq. | % |
| 1 | the | 91,845 | 6.79 | 1 | the | 109,524 | 5.15 |
| 2 | of | 45,984 | 3.40 | 2 | and | 54,545 | 2.56 |
| 3 | to | 39,315 | 2.91 | 3 | of | 50,741 | 2.38 |
| 4 | and | 32,225 | 2.38 | 4 | to | 49,433 | 2.32 |
| 5 | in | 28,924 | 2.14 | 5 | a | 42,256 | 1.99 |
| 6 | a | 23,753 | 1.76 | 6 | I | 38,787 | 1.82 |
| 7 | for | 13,195 | 0.98 | 7 | in | 35,229 | 1.66 |
| 8 | is | 11,651 | 0.86 | 8 | it | 34,243 | 1.61 |
| 9 | on | 10,726 | 0.79 | 9 | you | 30,562 | 1.44 |
| 10 | be | 9,271 | 0.69 | 10 | that | 30,134 | 1.42 |
| 11 | that | 9,171 | 0.68 | 11 | s | 27,977 | 1.31 |
| 12 | as | 9,162 | 0.68 | 12 | is | 20,400 | 0.96 |
| 13 | by | 8,834 | 0.65 | 13 | for | 17,548 | 0.82 |
| 14 | with | 7,575 | 0.56 | 14 | was | 16,138 | 0.76 |
| 15 | or | 7,185 | 0.53 | 15 | on | 15,394 | 0.72 |
| 16 | said | 6,972 | 0.52 | 16 | n't | 14,368 | 0.68 |
| 17 | will | 6,901 | 0.51 | 17 | we | 14,132 | 0.66 |
| 18 | Hong | 6,814 | 0.50 | 18 | they | 13,015 | 0.61 |
| 19 | it | 6,553 | 0.48 | 19 | be | 12,777 | 0.60 |
| 20 | at | 6,387 | 0.47 | 20 | he | 12,446 | 0.58 |

**Table 1  Comparison of the most frequent words in the PUBC and the BNC**

The other method of word selection is to check the keyness of a word. The British National Corpus (sample, 1.2 million words on CD-ROM) was used as a reference corpus. Although it is composed of both spoken and written texts (as opposed to the PolyU written corpus), it is the only standard English corpus currently available. The table of keyness was obtained by Wordsmith Tools. The top 23 words are shown in Table 2.

Of the words obtained from the corpus, some are highly specific, some are semispecific, and others are general English words. Keyness indicates which words recur consistently in texts of a given genre. For example, the word *consolidate* was found to occur in many of a set of business annual reports. It did not occur very often in each of them, but did occur much more consistently in

the business reports than in a mixed set of texts. The table of keyness helped us to make decisions in terms of word selection. The higher the number, the more relevant the word is to the lexicon. Words with high frequency may have very low keyness. For instance, the word *university* appears 122 times in PUBC, but its keyness is very low, –305.5, and it was excluded from the list of the business lexicon. Other words with high frequencies but low keyness are *problem* (f = 309, KN = –29.20), *few* (f = 387, KN = –24.8), and *always* (f = 159, KN = –315.3). By contrast, some words, although with low frequencies, have comparatively high keyness and are therefore retained in the lexicon, for example, *deflation* (f = 67, KN = 99.4), *inflow* (f = 60, KN = 91.9), and *liquidation* (f = 62, KN = 85.1). In addition to morphological analysis, frequency count and keyness study, we will also seek expert advice in order to minimize undue personal judgement in word selection.

| Keywords (keyness) | | | | | | |
|---|---|---|---|---|---|---|
| No. | Word | Freq. | All.Lst % | Freq. | BNCS.Lst % | Keyness | P |
| 1 | Hong | 6,814 | 0.50 | 92 | | 12,017.5 | 0.000000 |
| 2 | Kong | 6,299 | 0.47 | 94 | | 11,040.1 | 0.000000 |
| 3 | market | 5,493 | 0.41 | 693 | 0.03 | 6,740.0 | 0.000000 |
| 4 | per | 5,266 | 0.39 | 587 | 0.03 | 6,732.2 | 0.000000 |
| 5 | cent | 4,521 | 0.33 | 358 | 0.02 | 6,349.3 | 0.000000 |
| 6 | securities | 2,951 | 0.22 | 45 | | 5,160.6 | 0.000000 |
| 7 | HK | 2,591 | 0.19 | 0 | | 4,902.0 | 0.000000 |
| 8 | financial | 3,362 | 0.25 | 240 | 0.01 | 4,833.7 | 0.000000 |
| 9 | insurance | 2,953 | 0.22 | 146 | | 4,553.6 | 0.000000 |
| 10 | bank | 3,249 | 0.24 | 283 | 0.01 | 4,454.4 | 0.000000 |
| 11 | company | 3,736 | 0.28 | 599 | 0.03 | 4,176.0 | 0.000000 |
| 12 | the | 91,845 | 6.79 | 109,524 | 5.15 | 4,040.3 | 0.000000 |
| 13 | exchange | 2,331 | 0.17 | 135 | | 3,495.6 | 0.000000 |
| 14 | China | 2,192 | 0.16 | 126 | | 3,291.6 | 0.000000 |
| 15 | billion | 1,945 | 0.14 | 60 | | 3,198.9 | 0.000000 |
| 16 | companies | 2,399 | 0.18 | 246 | 0.01 | 3,143.3 | 0.000000 |
| 17 | of | 45,984 | 3.40 | 50,741 | 2.38 | 3,095.6 | 0.000000 |
| 18 | stock | 2,264 | 0.17 | 239 | 0.01 | 2,940.7 | 0.000000 |
| 19 | million | 2,653 | 0.20 | 446 | 0.02 | 2,903.9 | 0.000000 |
| 20 | investors | 1,698 | 0.13 | 47 | | 2,825.5 | 0.000000 |
| 21 | shares | 2,274 | 0.17 | 276 | 0.01 | 2,824.8 | 0.000000 |
| 22 | banks | 1,715 | 0.13 | 75 | | 2,695.0 | 0.000000 |
| 23 | investment | 1,875 | 0.14 | 142 | | 2,658.9 | 0.000000 |

**Table 2 Keyness of PUBC against BNC 1.2m sample**

### 3.4    Concordance in the lexicon

The idea of a corpus-based lexicon is to provide users with words in their context. Since the Hong Kong PolyU Business Lexicon has been positioned as a pedagogical dictionary, the established conventions of dictionary microstruc-

ture will still be used to meet the needs of English learners. An entry in the lexicon includes the following items:

1. Head word/Entry
2. Pronunciation
3. Part of speech
4. Definition
5. Phrases
6. E-concordance
7. Derivatives
8. Concordance of derivatives
9. Synonyms and antonyms
10. Chinese equivalent
11. Chinese concordance
12. Japanese equivalent
13. Japanese concordance

The special feature of this lexicon is the limiting of the definitions of a word and the examples the concordance provides for the business sector. The word *margin* normally has more than six senses in a general English dictionary, but we only included three in our lexicon: "1. a permissible difference; 2. security deposit; 3. gross profit". The definitions are directly related to a financial context and the senses are supported by evidences from the PUBC.

Semitechnical words are normally polysemic and have different meanings in different registers. Research into dictionary users has revealed that such words cause more problems to foreign language learners than either general or highly technical terms (Li 1998: 72). The concordance from the business corpus can provide clear examples of how they are used and what they mean in a specific domain. The word *cushion* in the corpus does not mean "a soft pillow to make sitting and resting more comfortable or something soft to decrease a collision"; rather, it means "to lessen the adverse effect of".

```
n.
      is ahead and have a firm  cushion  against any new contingencies
 he banks have a large capital  cushion  and a drop in profits (or ev
  content with a weak yen as a  cushion  for its economy, but the US a
   ng also proposes creating a  cushion  for the management of interba
 ese H-shares to gain a bit of  cushion  from US interest rate worries
 se of their relatively strong  cushion  of capital   and liquidity."
 n adequate level to provide a  cushion  of security for   depositors
 AMRO Asia (Holdings) Ltd will  cushion  such provisions which the Boa
 with FRR has provided a sound  cushion  to both investors and market
 iquidity so as to give them a  cushion  to guard against any unforese
 he LAF has already provided a  cushion  to prevent sharp interest rat
v.
  table. But companies can best  cushion  the blow by establishing   an
```

```
t rise for BP Amoco helped to cushion the FTSE's decline.   Defensi
mittee. This   should help to cushion them against the adverse impa
```

The concordance provides evidence of how the word *cushion* is used in financial services rather than in everyday speech. Other examples are *flow, liquid, curb*, and *lobby.* Such metaphors have lost their figurative content and have become genre-specific, essential to be included in a specialised lexicon.

Color words in the business-oriented corpus also have extended meanings, e.g. *blue chips, red chips, gray market, green shares*, and *in the black.* Many examples of this kind can be found with semitechnical words. Showing the range of their use in a certain field in a specialized dictionary can therefore be more convenient for language learners than a general English dictionary.

The trilingual project continues to develop. "Because all humans have the same basic perceptual apparatus and share many other experiences, there would be some strong similarities in the structuring of semantic space across languages" (Hatch and Brown 1995: 116). However, not only do languages differ in the number of terms they use for a concept, but also the range of meaning of each term may cover the concept in different ways.

## Conclusion

A corpus-based lexicon is extremely useful to users in a professional discourse community. However, it is essential to have a clear model before building up a corpus. Decisions need to be made at an early stage regarding language(s), text type, database structure and potential end users. With the help of a self-selected corpus, it is possible to compile a dictionary or a lexicon targeting a special user group, which benefits from a high level of specialty and currency.

## Acknowledgment

## References

**Biber, D., S. Conrad and R. Reppe.** 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

**Bolt, P. and K. Bolton.** 1996. The International Corpus of English in Hong Kong. Greenbaum, S. (Ed.). 1996. *Comparing English Worldwide*: 197-214. Oxford: Clarendon Press.

**Grishman, R. and R. Kittredge.** 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing.* Hillsdale, N.J.: Earlbaum.

**Hatch, E. and C. Brown.** 1995. *Vocabulary, Semantics, and Language Education*. Cambridge: Cambridge University Press.

**Li, L.** 1998. Dictionaries and their Users at Chinese Universities: With Special Reference to ESP Learners. McArthur, T. and I. Kernerman (Eds.). *Lexicography in Asia*: 61-79. Tel Aviv: Password Publishers.

**McArthur, T.** 1998. *Living Words: Language, Lexicography and the Knowledge Revolution*. Exeter: Exeter University Press.

**McArthur, T.** 1999. *The English Languages*. Cambridge: Cambridge University Press.

**Oakes, M.** 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

**Ooi, V.B.Y.** 1998. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

**Oostdijk, N.** 1998. Language Use in a Restricted Domain. Renouf, A. (Ed.). *Explorations in Corpus Linguistics:* 170-179. Amsterdam: Atlanta.

**Rundell, M.** 1998. Recent Trends in English Pedagogical Lexicography. *International Journal of Lexicography* 11(4): 314-342.

**Svensén, B.** 1993. *Practical Lexicography — Principles and Methods of Dictionary Making*. Oxford/New York: Oxford University Press.