Shogakukan Corpus Query System in Collaboration with the American National Corpus Project

TONO, Yukio
Department of Foreign Languages and Cultures, Meikai University
8 Akemi, Urayasu, Chiba 279-8550, JAPAN
y.tono@meikai.ac.jp
IWASAKI, Hirosada
Foreign Language Center, University of Tsukuba, Ibaraki, Japan
iwasakiH@sakura.cc.tsukuba.ac.jp
NAKAMURA, Takahiro, SUZUKI, Masanori, EGAWA Eiji
Shogakukan Inc. 2-3-1 Hitotsubashi Chiyoda-ku, Tokyo 101-0081, Japan
{takahiro,masonori,eegawa84}@mail.shogakukan.co.jp

Abstract

This is a progress report on the development of a corpus query language (CQL) interpreter, a web-based concordancer and a corpus query engine using OpenText. This system has been developed as a part of the process of reforming the lexicographic infrastructure at Shogakukan as we join the American National Corpus Consortium. The system will eventually be used for corpus-based dictionary-making for languages other than English. The implementation of a sophisticated query system for large corpora of the kind reported here aims at a more rigorous use of natural language data for dictionary-making and will hopefully contribute to the automatization and streamlining of the entire dictionary-publishing process.

1. Introduction

Until quite recently, the compilation of general-purpose or pedagogical dictionaries in Japan depended primarily on the analysis and synthesis of the information available in existing dictionaries. Especially in the case of bilingual learners' dictionaries, we have a long tradition of importing and modifying the features of monolingual dictionaries published in English-speaking countries. The use of language corpora started to attract the attention of Japanese lexicographers when the first genuinely corpus-based dictionary, the *COBUILD English Dictionary*, was published in 1987, but it was not until 1995, when major monolingual learner's dictionaries, the so-called the Big Four (*OALD, LDOCE, COBUILD* and *CIDE*), were all revised using large corpora that publishers and lexicographers in Japan began to realise how big an impact corpus-based techniques were having on the dictionary–making process. Since then, Japanese publishers have been making attempts at exploiting corpus resources for dictionary making although they all found the first hurdle rather difficult to cross: the systematic compilation of a large, balanced corpus. One of the problems they faced was that most of the corpus resources available were intended for research use only, which made it necessary for publishers to compile their own resources in-house. The problem of copyright, however, proved to be a major stumbling block. Another problem is that while Japanese publishers primarily focus on American English as a target norm, a standard corpus of American English comparable to the British National Corpus (100 million words) did not yet exist. Thus, the judgements about language features based on corpora have tended to be Euro-centric.

This situation is now changing. A group of American linguists and NLP researchers, along with some publishers, have launched a project to compile an American counterpart to the British National

Corpus. Shogakukan, among other Japanese publishers, joined the consortium of the American National Corpus and started developing their own corpus query system in late 1999. Tono and Iwasaki joined Shogakukan's project team as advisory staff. This paper is the first interim report on the progress of Shogakukan's corpus query system.

2. The American National Corpus project

The American National Corpus (ANC) project is fostering the development of a corpus comparable to the British National Corpus (BNC), covering contemporary American English. The organizing committee was established by Charles Fillmore (University of California, Berkeley), Catherine Macleod (New York University), Mark Liberman (Linguistic Data Consortium), Daniel Jurafsky (University of Colorado), Nancy Ide (Vassar University) and Ralph Grishman (New York University). A consortium of publishers of American English dictionaries and companies with interests in language processing has been formed. The founding members of the consortium are providing both materials for inclusion in the corpus and initial financial support for the project. The Linguistic Data Consortium is providing staff time to perform the initial clean-up and base-level encoding of the data and will manage the distribution of the corpus.

At the time of writing (April 2001), there are 15 founding members in the consortium. Seven of them are Japanese dictionary publishers (Shogakukan, ALC Press, Obunsha, Taishukan, Kenkyusha, Benesse Corporation and Sanseido). In Japan, the norm at school tends to be towards American rather than British English. Thus a plan for compiling a large representative corpus of American English and making use of the data for pedagogical lexicography is a quite attractive idea for Japanese dictionary publishers.

The project has three phases. Phase 1 involves the design of the overall architecture of the corpus and the production of the "level 0" corpus. A level 0 corpus has "clean" texts with minimal headers for each document and automatic mark-ups for logical structure down to the level of paragraphs. During Phase 2, a "level 1" corpus will be produced, with mark-up at the level of sentence and word boundaries, including the handling of problematic "word"-tokenisation issues involving categories such as proper names, dates, and numbers, among others. The aim is also to produce a grammatically tagged corpus at this stage although a POS tagger for this project has not been chosen yet. At Phase 3, further validation of existing mark-up and annotation will be done while additional annotations for such linguistic phenomena as discourse or phonetic information will be provided independently by different research groups. The project aims to establish a base corpus (100 million words) initially and add 10% of new text every five years so that they can keep the corpus updated while the original corpus remains always accessible.

3. A new corpus query system: rationale

The new corpus query system presented in this paper has been developed at Shogakukan, one of the biggest general publishers in Japan. It has many innovative design features in terms of its user interface, database architecture, and corpus query language (CQL). The development of the system was motivated by several factors. First, more and more people in electronic dictionary publishing have begun to recognize the potential of large linguistic databases or corpora . As we gained better access to large corpora such as the BNC (and the ANC in the near future), we launched a project to develop a corpus query system which can deal with multilingual corpora. Shogakukan produces learners' dictionaries in English, German, French, Spanish, Korean and Chinese, which makes the use of multilingual corpora even more important. Secondly, the rapid growth of the Internet has brought about new possibilities in terms of corpus query interfaces. Until a decade ago, one had to use a character-based interface via telnet such as *COBUILD Direct*, or client software specially designed for a particular corpus, such as *Sara* for the BNC. With these tools, people had to learn how to operate each tool individually. There are now several concordancers such as *WordSmith*, *MonoConc* or *TXTANA* which are more 'generic', but they also each have to be installed on individual PCs and users have to familiarise themselves on how to use each one. Whilst corpus linguists, experienced teachers and lexicographers in Europe did not see any problem with this situation, in Japan it constituted a tremendous problem for dictionary makers. In Japan, we do not usually hire full-time lexicographers for a project. Instead, a team of part-time lexicographers, usually university and senior-high school teachers under the supervision of a university professor as chief editor, is formed and the members work on their entries separately at home, while continuing with their other jobs. Under these circumstances, it is essential to provide client software that does not need to be installed individually and whose GUI is user-friendly enough not to require extensive training or practice. Web-browser-based software can fulfil this demand. It is exactly for the same reasons that the growth of Internet resources has forced the conversion of many client-server systems on the local area network into web-based alternatives. With the web-browser-based concordance software, a lexicographer does not need to work in-house any more. He or she can work at home under exactly the same environment as all the other members of the team. In terms of cost and time effectiveness, it would be a significant accomplishment to create a "virtual electronic dictionary-publishing environment." This makes it possible, for instance, to ask professional lexicographers overseas to join the project, or to recruit Japanese people working abroad for part-time lexicographic work. It is for the above reasons that we embarked on the design of a web-based environment for lexicographic work, where Internet browsers are all that is needed and where corpus data is provided on-line via a web server.

4. Basic design principles

As we designed the system, we took into account the following aims:

- To define a fundamental corpus data structure in order to enhance the scalability of the system. In other words, the system should not depend on a particular corpus data structure. Any corpus and its mark-up can be transformed into a format which our system can use. This is limited to the data structure only (i.e. the text-structural/formatting aspects). It does not affect the linguistic annotations. Such annotations, e.g. POS, lemma, syntactic and semantic tags, can be freely defined by the user.

- To develop or commercially adapt the engine for searching large corpora.

- To design a corpus query language (CQL) and develop the command interpreter for this.

- To develop the GUI of a web concordancer with several different specifications based on the user's skill level. In order to do this, we divided the system into two layers: the GUI and the command interpreter.

We will elaborate on each of these points below.

4.1. The data structure of tagged corpora

The data structure of a corpus is encoded by the use of a markup language such as SGML or XML. These markup languages have been designed to record structural information about text documents. There are some projects such as TEI (Text Encoding Initiative) or CES (Corpus Encoding Standard) which provide guidelines for standardizing the markup, and DTDs (Document Type Definitions) specially designed for marking up large amounts of highly structured electronic texts. There is another level of encoding information in a corpus: *annotation.* Annotation is a method of enriching corpus data by encoding all sorts of linguistic interpretation (syntactic, semantic, pragmatic, prosodic, and phonetic) in the text. Usually this is done by using a specific set of tags. Tagged corpora are extremely useful because they make it possible to search not only words but also all these linguistic annotations, and combinations of word forms and annotations. For example, with word tokens tagged with lemma and POS information, one can search for what types of adverb follow the lemma "LOOK". Without lemma information, one must search for all the variant inflectional forms of the verb "to look". POS information helps to distinguish the verb "look" from the noun "look". This type of information will significantly reduce the work of lexicographers in their retrieval tasks. Another use of such annotations is in collocation analysis. Richer and more sophisticated collocation tables can be produced if one can base the collocational analysis on lemma or part of speech.

The XML or SGML markup will make it possible to restrict queries to specific texts or parts of texts based upon the header information. By choosing particular categories in the header, one can easily create subcorpora, e.g. texts whose target readers are teenagers, or political texts from newspapers. With large corpora such as the ANC, users would probably often find it useful to select subgenres or create their own subcorpora for specific tasks or queries. Therefore, in designing the corpus query system, it is essential that the system has a sufficient level of power and flexibility in order to handle the markup and annotations mentioned above.

In order to realize this, we assume that tagged corpora have independent information layers: simple raw text and corpus-specific annotations. The search engine is not affected by the specification of annotation levels in the corpus. In other words, the query system can perform any kind of query based on user-defined annotations. This may be better understood if one can imagine a two-dimensional table, in which the columns contain each word in the text and the rows indicate the different types of annotation for each word. The columns can be defined depending on how the minimal unit of the text (e.g. sentence, phrase, word, lemma, morpheme) is defined. The rows can be extended as more annotations are added. A complex "multi-level" query of the 2-dimensional patterns from the 2-dimensional data (text + annotation), can be performed, making for extremely flexible and complex search patterns.

4.2. Indexing

Proper indexing is needed for large text databanks of this kind. One million running words with header information in the XML format will easily consume about 1.5 to 2 Gigabytes of hard disk space. A simple pattern search using utilities such as GREP will not work suitably fast enough on such large amounts of text. Complex queries combining both textual information and annotations would be even more demanding on computational resources and prohibitively time-consuming without software which works on indexed texts.

OpenText Ver.5 is a full-text search engine which is SGML-aware and works on indexed data. It has been implemented in the electronic version of the *Oxford English Dictionary*. It not only performs simple full-text searches but also "region searches", in which one can limit the search to specific positions in texts delimited by SGML. This function helps users save time by allowing them to restrict demanding, multi-level complex searches to specific texts or sections of texts. Another strength of using OpenText is that this region search makes it possible to combine different parts of the header information for the query. Thus, our CQL specification can deal with the querying of header information (which makes it possible to define subcorpora freely) and combine this with word- and annotation-level queries. Since the DTD to be distributed from the ANC consortium may not necessarily conform to the OpenText format, a conversion program has been developed for this purpose.

4.3. The development of Corpus Query Language (CQL)

Commercial concordancers such as *WordSmith* or *TXTANA* have the design of their text-handling facilities closely integrated with their querying facilities. Whilst it is true to say that the 'GUI' has nothing to do with the limitations of the querying facilities and that it, strictly speaking, simply refers to the front-end embellishments (windows, buttons, use of colours, etc.), some programs do not seem to be designed that way. In other words, there is a tendency in some concordancers that how the programs are designed to handle textual mark-up and annotations imposes limitations on query patterns. The real issue is the design of the program itself: how 'aware' it is of the structure of marked-up texts, but the GUI imposes restrictions instead. For instance, *TXTANA* is designed in such a way that subcorpora are selected based on filenames and directory structure and need to be predefined before the search. WordSmith can create subcorpora based on the header information, but the search area is not controlled by SGML, but by position or text length (the first 10,000 characters). For these concordancers, tagged corpora are also handled as plain text. Thus, pattern matching using regular expressions has serious deficiencies, and complex searches combining both document structure mark-up and linguistic annotations are not possible. This kind of design limitation in a concordancer (a lack of full SGML/XML awareness) represents a failure to exploit the sophisticated, complex structuring of data offered by XML/SGML.

We designed our system architecture as follows. First, we divided the system into two layers: the GUI and the corpus query language (CQL). The CQL layer is a software layer where a command interpreter terminates and stays resident. The CQL was developed from scratch based on the specifications to be implemented in the ANC. This approach is parallel to the relationship between the relational database and its query language, Structured Query Language (SQL).

CQLs have been developed by several different research groups (e.g. LDC, IMS (Christ 1994), among others). These query languages were developed by research professionals who needed to specify complex queries in a simple user interface (in most cases, just one query window). Our design architecture, on the other hand, is the result of our careful systems planning, in which the software layer is kept completely independent of the GUI layer.

5. Conclusion

By dividing the system architecture into GUI and CQL layers, we avoid being faced with the dilemma of whether to implement a simple GUI for novice users, who will only make queries for a word or phrase, or a complex GUI for power users, who do complex searches using regular expressions as sophisticated as those afforded by Perl-like regular expressions and header-level searches involving XML/SGML elements. With the GUI independent from the CQL, we can provide character user interfaces (CUIs), for example, a shell screen for users to type in the CQL commands directly. We will give a demonstration of our system in our presentation, where we will show how a simple query table intuitively helps the user to type in a 2-dimensional query. This table is very flexible in the sense that one can add as many rows of annotation as are needed for the corpus.

The next stage of our development is to test this system with the actual ANC data. Since we have not obtained the first data set yet, our system is now based primarily on BNC data with minor adjustments. It is important, therefore, to continue to test the validity of the system as data comes in from the ANC and make revisions to the CQL as necessary. As was said earlier, a full exploitation of tagged corpora needs to be further investigated in the area of the corpus query system development. There is a genuine need for corpus linguistics to develop in line with software engineering. The corpus query system presented here represents a collaboration between a dictionary publisher and corpus linguists, and we hope that the entire process of traditional dictionary-making will be critically reviewed, and that our system will facilitate the automatization and streamlining of the entire dictionary-publishing process.

References

Christ, Oliver. (1994) A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research*, Budapest, Hungary, 1994. CMP-LG archive id 9408005.

Challenges in compiling a corpus based trilingual business lexicon

Li Lan Grahame Bilbow Xu Xunfeng
Department of English
The Hong Kong Polytechnic University

Abstract

General purpose dictionaries can benefit users in many ways, bu(their definitions and examples might not satisfy the needs of users in a professional discourse community. What does '*cushion*' mean in business English? What can be *green* in the domain? The paper investigates what contribution a business language corpus can make to a business lexicon. The trilingual domain-specific corpus, in English, standard written Chinese and Japanese, compiled by the authors is described in some detail, followed by the analysis of the generic wordlist produced by WordSmith Tools. Issues addressed include (1) the relation between lexicon and corpus; (2) the selection of entry words for a business dictionary; (3) how entry words should be arranged (alphabetical Vs thematic); (4) how compound words should be treated, first position or second position, including compositions with the prefixes cyber-, quasi-, etc; (5) the presentation of various form classes in the same or a different entry. Related issues, such as the inclusion of proper nouns & abbreviations, lower-/upper-case, grammatical category (e.g., head word case) are also addressed. A model for what should be included in an entry word article is outlined. Special discussions go to new words and metaphorical use in business language. We try to prove that with concordances directly from the business corpus, such a dictionary can provide clear examples of words in context, more relevant than those from general English dictionaries or from a general English corpus. Another advantage is that a corpus based dictionary can technically combine multiple access methods, using hyperlinks to bring languages, concepts and dictionary entries together. We have followed morphological line and semantic line of words and illustrated the relationships both alphabetically and thematically, together with synonyms, antonyms and versions cross languages. Tom McArthur's snowflake principle is used to explain holes in the lexicon.

I. Introduction

'Dictionary work is long and complicated, needs consistency and system, endurance and precision, time and money' (Svensen 1993:250). The application of computers in lexicography, however, has made considerable changes in the last three decades in this area. The computer can work consistently with high speed, store large quantity of material and allow quick retrieval. The study of lexicology is therefore to use both knowledge-based and statistical approaches to the lexical analysis of texts. In this project, we investigate a particular register, business languages used in Hong Kong, including English, standard written Chinese and Japanese, which leads to a corpus based trilingual business lexicon.

Business transaction was, is and will be one of the main driving forces for English to spread internationally. Good communication is central to any successful business and research has demonstrated a positive relationship between profitability and good language practice in this area (Wright and Wright, 1994). Users of Business English need to know English particularly for the special subject which they are studying or practicing. A lexicon developed from specialized corpus can help to illustrate linguistic characteristics of language varieties including lexical frequencies, collocations and characteristic grammatical structures. The Hong Kong PolyU Business Lexicon is based on the Hong Kong PolyU Business Corpus (PUBC), which covers three languages, English, Standard Written Chinese and Japanese, and includes all types of business communicative writings, namely company reports, press releases, letters, faxes, news stories, booklets etc., making them three comparable subcorpora.

Leech (1992:341) postulates four models concerning the role of corpus processing: A. the Linguistic Information Retrieval Model; B. the Induction Model; C. the Automatic Corpus Processing Model and D. the Self-organizing Model. We adopted Model A to produce such basic lists as word-frequency and keyword-in-context/concordance to aid human analysts in the study of business language. WordSmith Tools by Mike Scott and ConcApp by Chris Greaves were used for data analysis. WordSmith is one of the best currently available software which can provide powerful statistic information by comparing the list generated from the PUBC with other lists such as British National Corpus. Such information is useful when deciding wordlist of the lexicon. The striking feature of ConcApp lies in its editing power. With a menu bar like the one in Microsoft Word, one can ***copy, paste, blacken, italic*** and ***underline*** freely and save files with any texts. It can also be used for concordancing non-romanized languages like Chinese and Japanese.

The PolyU Business Lexicon differs from other business dictionaries in that, with corpus support, it provides key words in context directly from language used in real business practice. The electronic version can free the editors from space problem and many functions and quick retrievals can be realized by hyperlinks between different databases. This paper will concentrate on lexical description rather than technical consideration.

II. Selection of entry words

Vocabulary control is always a central concern in dictionary compilation. Of the three subcorpora, the English subcorpus was used as the base or starting point of the lexicon. The preliminary word list produced by WordSmith has 22, 600 word types. After trimming off reflected forms, nonsense words, proper names, functional words and words which are not specific in business texts, we settled on a list of 4230 base words. (see Li & Grahame 2001).

Language has been changing all the time. The word list of the Lexicon generated from the English corpus has recorded some change with the evidence that business vocabulary has enlarged in several ways: 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.

Many new words merge from the application of modern technology in business sector. The prefixes *cyber-*, *e-*, *elec-* and *i-* are used as proper nouns such as names of websites or publications and also shift to proper lexical words, mostly nouns and verbs. The following words are from the business corpus of 1.2 million words:

Word	Frequency
cyberbanking	10
cyberbase	8
cybercash	1
cyberforce	2
cybermall	1
cyberport	25
cyberspace	28
cybertrading	2
cyberworks	193

One new word included in PBL is the word *dotcom*, which has become a commonplace to modify Internet business. There are *dotcom connections, dotcom stores, dotcom woe, dotcom consulting, dotcom marketing, dotcom industry*, etc.. The concordances of dotcom from PUBC are also provided in the lexicon.

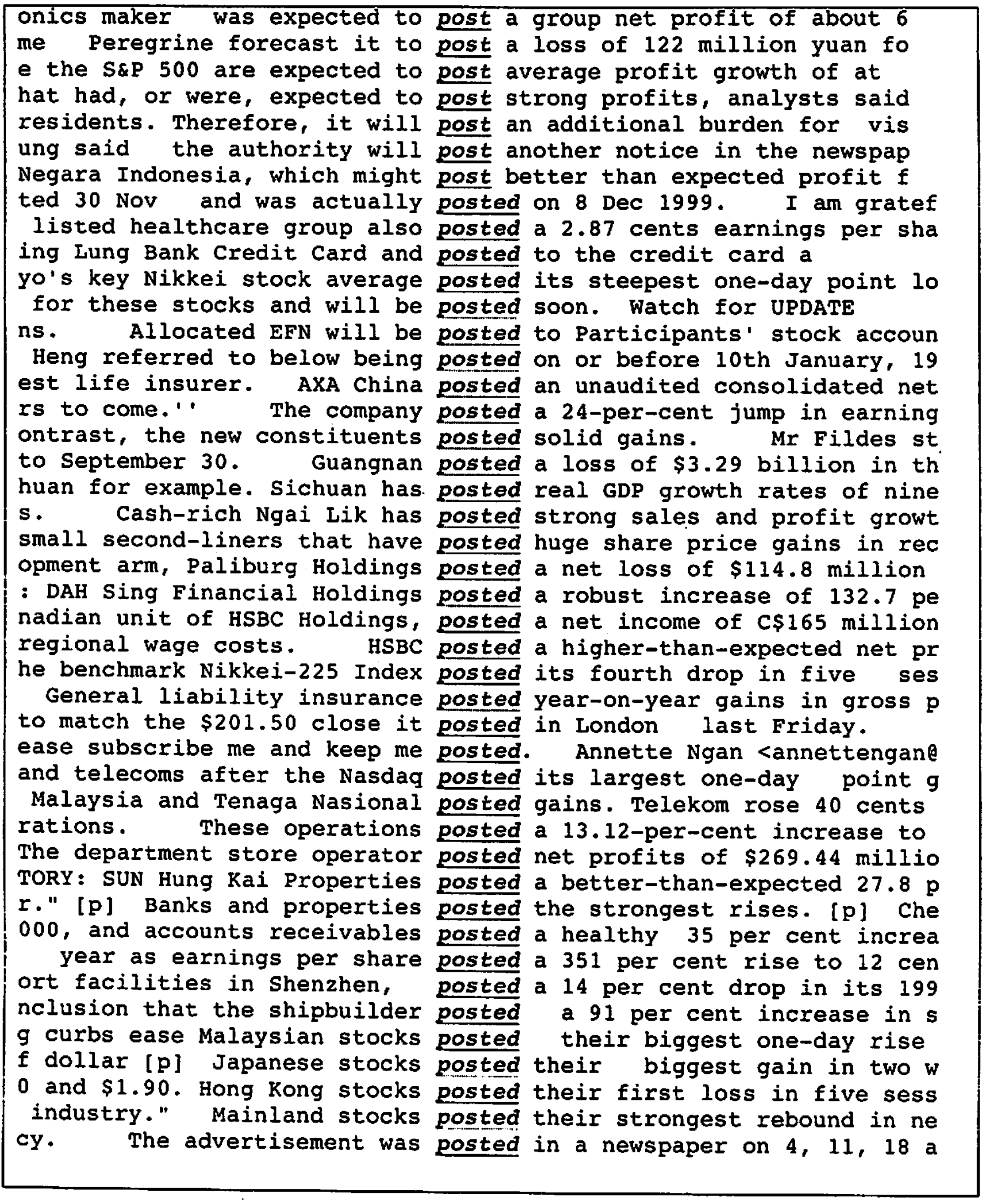
```
But he said the fast-growing   dotcom   companies helped improve the
hnology stock prices soar and  dotcom   companies rush in to the mark
in.     ``No doubt some of our dotcom   companies will do better than
  the new economy, we own no   dotcom   companies. What we've been lo
ong    STORY: WITH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

Another word is *bancssurance*, which means insurance service provided by a bank. It was first used as a name of a project or programme and then become a special term in business English. Some people regarded it as a non-word because it seemed to have been used by only one or two authors. However, we recorded the word in the lexicon, not only because it appeared eight times in the PUBC, but also because it fits into standard English, has got clear orthographical, morphological evidence and stands for a clear communicative concept and has been printed out for the public. We also include words like *interbank* and *interbranch* for the same reason.

The project started right before the new millennium therefore it bears strong features of the period. The Y2K threat was regarded as the main concern due to the fact that without computer, nothing could be done in business today. The word *Y2K* has a frequency of 380 times in PUBC, existing in business news reports, company reports, internal and external documents. As the impact of the event associated with Y2K fades, the use of the name can no longer conjure up the feelings of fear, dismay, or tension that attended its early use, the term becomes more encyclopedic and less lexical, so we did not include it to the lexicon.

III. Sense order

The difficulty in deciding sense order lies in the fact that the PUBC does not have word class tag; neither has it got semantic parameter. Manual work has to be done to discriminate different senses and different parts of speech. The advantage of the concordances from the PUBC is self-explanation. With the concordances displaying words in context from authentic business material, the order of senses of an entry can then be organised accordingly. Take word *post* as an example. In *Cambridge International English Dictionary*, the order of *post* is arranged under six Guide Words: 1. letters, 2. pole, 3. job, 4. stick, 5. place and 6. pay. Oxford Learner's Dictionary of Current English also has similar sense order for the word. However, the examples from the Business Corpus implies that in a business lexicon the sense order should be different from general language dictionaries.



```
onics maker   was expected to post a group net profit of about 6
me   Peregrine forecast it to post a loss of 122 million yuan fo
e the S&P 500 are expected to post average profit growth of at
hat had, or were, expected to post strong profits, analysts said
residents. Therefore, it will post an additional burden for  vis
ung said   the authority will post another notice in the newspap
Negara Indonesia, which might post better than expected profit f
ted 30 Nov   and was actually posted on 8 Dec 1999.    I am gratef
 listed healthcare group also posted a 2.87 cents earnings per sha
ing Lung Bank Credit Card and posted to the credit card a
yo's key Nikkei stock average posted its steepest one-day point lo
 for these stocks and will be posted soon.  Watch for UPDATE
ns.     Allocated EFN will be posted to Participants' stock accoun
 Heng referred to below being posted on or before 10th January, 19
est life insurer.   AXA China posted an unaudited consolidated net
rs to come.''     The company posted a 24-per-cent jump in earning
ontrast, the new constituents posted solid gains.      Mr Fildes st
to September 30.     Guangnan posted a loss of $3.29 billion in th
huan for example. Sichuan has posted real GDP growth rates of nine
s.     Cash-rich Ngai Lik has posted strong sales and profit growt
small second-liners that have posted huge share price gains in rec
opment arm, Paliburg Holdings posted a net loss of $114.8 million
: DAH Sing Financial Holdings posted a robust increase of 132.7 pe
nadian unit of HSBC Holdings, posted a net income of C$165 million
regional wage costs.     HSBC posted a higher-than-expected net pr
he benchmark Nikkei-225 Index posted its fourth drop in five   ses
  General liability insurance posted year-on-year gains in gross p
to match the $201.50 close it posted in London   last Friday.
ease subscribe me and keep me posted.   Annette Ngan <annettengan@
and telecoms after the Nasdaq posted its largest one-day   point g
 Malaysia and Tenaga Nasional posted gains. Telekom rose 40 cents
rations.     These operations posted a 13.12-per-cent increase to
The department store operator posted net profits of $269.44 millio
TORY: SUN Hung Kai Properties posted a better-than-expected 27.8 p
r." [p]  Banks and properties posted the strongest rises. [p]  Che
000, and accounts receivables posted a healthy  35 per cent increa
   year as earnings per share posted a 351 per cent rise to 12 cen
ort facilities in Shenzhen,   posted a 14 per cent drop in its 199
nclusion that the shipbuilder posted   a 91 per cent increase in s
g curbs ease Malaysian stocks posted   their biggest one-day rise
f dollar [p]  Japanese stocks posted their   biggest gain in two w
0 and $1.90. Hong Kong stocks posted their first loss in five sess
 industry."   Mainland stocks posted their strongest rebound in ne
cy.     The advertisement was posted in a newspaper on 4, 11, 18 a
```

Collocation forms a very important aspect of the lexicon: indeed the sublanguage may be defined, in part at least by the choice of the collocations used. With *post*, the occurrence of such objects as *profits, loss, and shares* indicates its special register. The meaning of mailing or sending appeared only once, therefore, the sense order of the entry is arranged in this way:

post *vt.* 1. publicize with, announce; 2. display in a public place; 3. mail, send; 4. assign to a post, put into a post.
Sense 4 did not appear in the corpus, but since the meaning is dominant in noun form of *post,* we regarded it necessary to include it.

Another example is the word *concrete.* The collocations show it is used much more often as an adjective than as a noun in business text, and therefore we only include the adjective form in the lexicon.

```
nt.   But he said there is no concrete agreement at the moment.   As
obody expected something this concrete and this good for the gold pr
he share capital in Hong Kong Concrete Company Limited in early 1998
al studies have provided that concrete  evidence of the link, ( alth
cial   architecture, and that concrete mechanisms should be put in p
Hong Kong, but there was   no concrete plan as yet.     Colonial has
ny had not come up with any   concrete plan yet in setting next year
ouldn't we be looking at more concrete plans as to what exactly are
ry union. Whilst there are no concrete plans to extend   this to the
atin America and recommending concrete policy actions. The forum's t
cost.   However, there are no concrete restructuring plans for the t
ient experience to craft more concrete rules, whilst retaining the n
```

IV. Genre-specific words

It is not difficult to define highly technical terms such as *dividend, annuity, insurance, increment,* tax, *stock, share, portfolio, and broker* in business lexicon. There are, however, many so called semi-technical words.

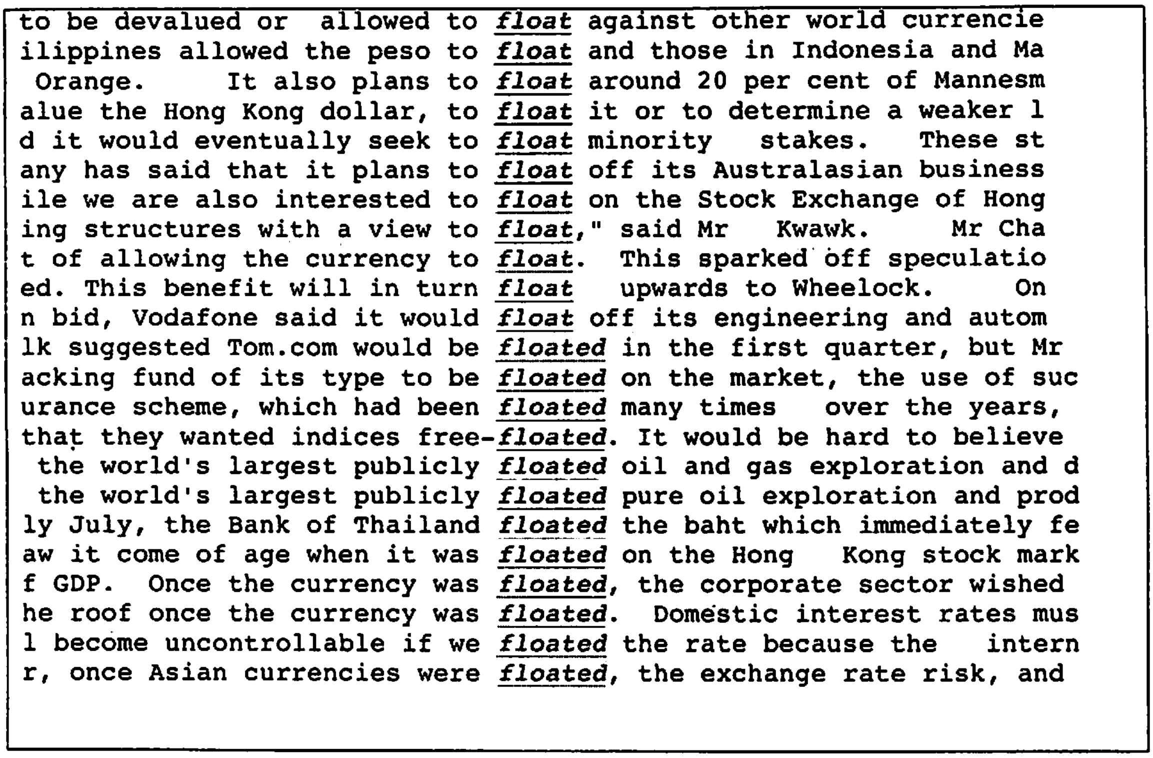
Collocations show clear markers of a register and form a very important aspect of the lexicon. Even with the word *account,* when appearing with such adjectives as *brief* and *full,* it indicates a more general language in use, meaning 'description'. In business English, the clusters are with profit and loss ~, credit ~, current ~, joint ~, trust ~, ~ number. The noun lemma *account* for 'record' also occurs much more frequently than the verb *account* which means 'explain'.

The word *cut* normally has over 30 senses in a general English dictionary. It is another typical example of semi-technical word. Cut is a very common action in business practice and the word has a frequency of 136 in PUBC. However no knives or other tools are needed in such situation. As a verb, it cuts cost, price, interest rate, mortgage, commission and payment. There are also *staff cut, payment cut, budget cut* and *rate cut.* These collocations are particularly helpful to define the word in the business lexicon.

Semi-technical words are often polysemic. The principle we follow is to record the meanings merged from the corpus only. The definition of the word *flirt* has nothing to do with sexual attraction, but 'to experiment with' because of the following concordances.

```
S trade, enabling the FTSE to flirt with the 4,300-point level fo
-Mart Stores are beginning to flirt with the concept. Merrill cit
```

Another good example is *float.*



```
to be devalued or  allowed to float against other world currencie
ilippines allowed the peso to float and those in Indonesia and Ma
 Orange.     It also plans to float around 20 per cent of Mannesm
alue the Hong Kong dollar, to float it or to determine a weaker l
d it would eventually seek to float minority   stakes.   These st
any has said that it plans to float off its Australasian business
ile we are also interested to float on the Stock Exchange of Hong
ing structures with a view to float," said Mr   Kwawk.     Mr Cha
t of allowing the currency to float.  This sparked off speculatio
ed. This benefit will in turn float   upwards to Wheelock.     On
n bid, Vodafone said it would float off its engineering and autom
lk suggested Tom.com would be floated in the first quarter, but Mr
acking fund of its type to be floated on the market, the use of suc
urance scheme, which had been floated many times   over the years,
that they wanted indices free-floated. It would be hard to believe
 the world's largest publicly floated oil and gas exploration and d
 the world's largest publicly floated pure oil exploration and prod
ly July, the Bank of Thailand floated the baht which immediately fe
aw it come of age when it was floated on the Hong   Kong stock mark
f GDP.  Once the currency was floated, the corporate sector wished
he roof once the currency was floated.  Domestic interest rates mus
l become uncontrollable if we floated the rate because the   intern
r, once Asian currencies were floated, the exchange rate risk, and
```

V. Gaps between languages

As claimed before, the entries of PBL starts from a wordlist generated from the PUBC English subercorpus, which means, every English word has at least one occurrence. The equivalents of Chinese and Japanese, on the other hand, do not necessarily appear in their subcorpora, although the three are thematically parallel. This is due to the fundamental difficulty of co-ordination of lexical units in different languages, which 'is caused by the anisomorphism of languages, i.e. by the differences in the organization of designate in the individual languages and by other differences between languages' (Zgusta 1971:140). There is obviously no ready equivalent for every English word to have its peer in Chinese and Japanese within a limited corpus. We expected gaps between the languages, but cherish the hope to keep them to the minimum. Since little work has been done for the Japanese lexicon, the discussion will concentrate on gaps between Chinese and English.

The equivalents in an interlingual dictionary are usually of two types: translation equivalents and explanatory equivalents. A translation equivalent is a lexical unit which can be immediately inserted into a sentence in a target language. For instance, bank = 銀行, regress = 衰退, share = 股票. An explanatory or descriptive equivalent is one which cannot always be inserted into a sentence in the target language, e.g. endorse = 在 (支票, 匯票) 后面 簽名。 If the translation takes 簽名, it only covers partial meaning of the word. Al-Kasimi summaries that a translation equivalent is either absolute or partial due to several factors:

1. The conceptual systems are not identical in different languages

2. Semantic fields of presumed equivalents in different languages are not always similar
3. The culture-specific words which denote objects particiuliar to the culture of the SL might not have corresponding equivalents in the TL
4. The scientific and technical terminology does not exist in one of the languages, e.g. the vernacular language
5. The meaning of words has a fluid and inconstant nature

(Al-Kasimi 1983:160)

More gaps occur in metaphors. Henderson found that 'the ECONOMY is a PERSON, a MACHINE or a PLANT' (Henderson 1982: 109). However, the ones function in English might not be the case in Chinese and Japanese.
The following table shows semantic differences in business metaphors of the three languages.

English	Chinese	Japanese
fever	4	4
ailing	✕	✕
headache	4	4
infant	✕	✕
cushion	✕	4
lobby	✕	4
ignite	✕	4
iceberg	✕	✕
flow	✕	4
parent	partial	partial

The similarities between Chinese and English metaphors are found in sentences 1and 2.
1. The eventual application of the new standards on all enterprises presented a ***headache.***
2。官股基金定價的確是個　問題□

3. The GEM launch in the latter part of next month would also help cool speculative fever towards technology stocks.

4。他提醒□本港科技股的　情況仍未獲減退。

Partial coverage appears in sentence 5. The Chinese term for parent company is 母公司(mugongsi), meaning mother company, while parent refer to both father and mother, or can be regarded gender free in this sense.

5.To complement the finance products offered by the ***parent*** *bank and to meet the needs of the customers and the local business community, the company has expanded its service to offer 3 finance products.*

6。要求聯交所監管其控股　並　不恰當。

Metaphors in Sentences 7,8 and 9 do not have direct equivalents.
7. Fears for further interest rate rises were ***ignited*** *on Thursday when slightly higher-than-expected US gross domestic product data for the first quarter indicated rising inflation.*

8. Conversely, venture capital can play a different role in funding risky infant companies.

9. Why firms lobby, they argue, is due to government intervention which has an effect on their cash flows.

In Chinese, a newly set-up company can be called a new company, not a baby company;
There are 人流 (flow of people) and 物流 (flow of materials), but not 錢流 (cash flow, capital flow, and money flow)。However, the verb flow 流 is used in business text.

From the concordances of the word *cushion,* it is not difficult to find that in business text its meaning is, for a verb, 'something providing protection against impact'; or for a noun 'to mitigate the adverse effect of'. We have *cushion the blow, decline and Hang Seng Index slide;* However, in Chinese, a bag stuffed with soft material as a comfortable support when sitting or sleeping is always a 靠墊 (kaodian), never has metaphoric meaning of 'softing the effect of'. Therefore, the Chinese equivalent we provide for *cushion* is only an explanation of its figurative meaning.

```
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
se of their relatively strong cushion of capital and liquidity."
n adequate level to provide a cushion of security for  depositors
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

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
AMRO Asia (Holdings) Ltd will cushion such provisions which the Boa
```

VI. Snowflake principle

When we first started the lexicon, we wanted it to be an all-in-one reference tool to uses. The features include definitions and usage specimens, index with lexical clues and built-in pronounced words, alphabetical lexical sets, semantic fields, style labels, synmym and antonym sets, phrases, demonstrations. In real practice we found many of the fields in the database are hard to fill. There is little traditional inventory of devices available to the creators of reference materials' (McArthur 1998:182).

The phenomenon can be explained by Tom McArthur's snowflake principle. The wordstock of a work can be logically expanded alone two lines: (1) a morphological line, through formational process that include derivation, compounding, conversion and abbreviation; (2) a semantic line, through sense relations like synonym, antonym and intelligibility to others. The systematic application of this task produces for each base word a range of derivatives and senses, not all of which are necessarily relevant to the task in hand. 'Each word is a snowflake, its morphosemantic pattern and potential distinct from all other words, because no two words open up morphologically and semantically in precisely the same way' (ibid.)

The words in the PBL are mainly nouns, verbs, adjectives and adverbs. Each word stretched in different way. Nouns may have compounds; verbs in many cases have derivatives, adjectives can have synonyms or close synonyms. It is impossible for an entry work to have all categories, and a presentation with too many holes might damage the lexicon's image, so we leave the choice to the use to decide what kind of information s/he needs to look up.

VII. Conclusions

A specialised corpus-driven lexicon has the following advantages: 1) It can record and demonstrate most up-to-date language use; 2) It can provide language evidence of words used in a special register better than a general language dictionary does; 3) The sense order in such a specialised lexicon would be more useful to second language learners for a special purpose. The limitation of the lexicon is that the entry list generated from a corpus restricts word availability. If a word does not appear in the corpus, the lexicon will not have it. This is especially difficult for a multilingual dictionary. There is complete correspondence between words and expressions in two languages as regards content and register, but some words only have partial equivalent and some have none in the respective languages. When the starting list is restricted, the word lists of target languages could, on the one hand, be smaller than the source one; and on the other, miss out some words in its own corpus. One solution would be to create three preliminary lists from the three subcorpora, but more incomplete 'snowflakes' are bound to emerge, leaving us in a vicious circle. The best way for us, at present, is to stick to the English list and use some hyperlinks to provide users with more resources.

References

Al-Kasimi, A. M. (1983). 'The interlingual/translation dictionary', in Hartmann (ed.) *Lexicography: Principles and Practice*. London: Academic Press, 153-162.

Biber, D., S. Conrad & 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', in S. Greenbaum (ed) *Comparing English Worldwide*. Oxford: Clarendon Press, 197-214.

Grishman, R. & R. Kittredge (1986). *Analyzing language in restricted domains: sublanguage description and processing*. Hillsdale, N.J.: Earlbaum.

Henderson W. (1986). 'Metaphor in Economics', in *Talking about Text*, M. Coulthard (ed.). University of Birmingham: english language research, 109-127.

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

Landau, S. I. (1984). *Dictionaries: the Art and Craft of Lexicography*, New York: Scribner Press.

Leech, G. (1982).

Li, L. (1998). 'Dictionaries and their users at Chinese universities: with special reference to ESP learners', in T. McArthur and I. Kernerman (eds.) *Lexicography in Asia*, Password Publisher, 61-79.

Li, L. & Bilbow, G. T. (2001), "From Business Corpus to Business Lexicon". *Lexikos*, 11(1)

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

Noor, N. (1998). *Word combinations for business English: a study based on a commerce and finance corpus for ESP/ESL applications*. Unpublised PhD Thesis at Lancaster University, UK.

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

Oostdijk, N (1998). 'Language use in a restricted domain', in *Explorations in Corpus Linguistics*, Renouf, A (ed.), Amsterdam: Atlanta, 170-179.

Svensén, B. (1993), *Practical Lexicography -- Principles and Methods of Dictionary Making*, Oxford & New York: Oxford University Press.

Zgusta, L. (1971). *Mannual of Lexicography*. The Hague: Mouton.