

# 行政院國家科學委員會專題研究計畫 成果報告

## Sketch Engine 為語言學習的工具 研究成果報告(精簡版)

計畫類別：個別型  
計畫編號：NSC 96-2411-H-004-048-  
執行期間：96年08月01日至97年09月30日  
執行單位：國立政治大學外文中心

計畫主持人：史尚明  
共同主持人：黃居仁

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫可公開查詢

中華民國 97年10月20日

# The Sketch Engine as a language-learning tool

Sketch Engine 為語言學習的工具 (NSC 96-2411-H-004-048)

研究成果報告(精簡版)

## Background

Computer-assisted language learning (CALL) is now of great significance and importance in the acquisition of second languages – especially English – in Taiwan and all over the world. There are entire journals devoted to research on the topic, including for example *Computer Assisted Language Learning*, published by Taylor and Francis. At Ming Chuan University, and indeed at most language teaching institutions, the use of online resources is now commonplace in the language classroom, and listening labs are computerized. Ellis (1995) notes that CALL has a particularly important role to play in the acquisition of vocabulary, because this is the part of language study to which the student can most usefully turn his attention in private. Thus, teacher contact hours can be devoted to more communicative activities that cannot so easily be practiced alone. There is, indeed, a great variety of applications available on the web for students to use in private study, such as the *Advanced English Computer Tutor* (MaxTex International) and *WordPilot* (CompuLang.com), and many others. For English, *WordPilot* offers corpus analysis and concordancing features (showing how a particular vocabulary item is used in context) as does Camsoft's *Monoconc*, which is available for other languages too, including Chinese. This system also lists the collocations in which keywords participate the most frequently.

Analysis of text and spoken language, for the purposes of second language teaching, as well as for dictionary making and other linguistic applications, used to be based entirely on the intuitions of linguists and lexicographers. The compilation of dictionaries and thesauri, for example, required that the compiler read very widely, and record the results of his efforts – the definitions and different senses of words – on thousands, or millions of index cards. Dictionary entries which seemed intuitively similar were placed together in boxes or piles, according to Speelman (1997), for later analysis. Thus, the distribution of items among sets preceded the lexical analysis, whereas under a computer-age model the analysis would come first, guiding the distribution: a distribution which could be based on masses of data, rather than the intuitions of the compiler.

Today's approach to linguistic analysis generally involves the use of linguistic corpora: large databases of spoken or written language samples, defined by Crystal (1991) as "A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language". Numerous large corpora have been assembled for English, including the British National Corpus (BNC) and the Bank of English. Dictionaries published by the Longman Group are based on the 100 million word BNC, and corpora are routinely used by computational linguists in tasks such as machine translation and speech recognition.

The BNC is an example of a balanced corpus, in that it attempts to represent a broad cross-section of genres and styles, including fiction and non-fiction, books, periodicals and newspapers, and even essays by students. Transcriptions of spoken data are also included; and this is the corpus that is used with the English Sketch Engine.

Central to corpus analysis is the context in which a word occurs: J R Firth pointed out that information about meaning can be derived from surrounding words and sentence patterns: "You shall know a word by the company it keeps", as he famously stated in 1957. A

convenient and straightforward tool for inspecting the context of a given word in a corpus is the KWIC (keyword in context) concordance, where all lines in the corpus containing the desired keyword are listed, with the keyword at the centre.

Various tools are available for exploring word context in corpora, determining by a statistical analysis which words are likely to appear in collocation with which others. Often, the statistic involved is mutual information (MI), first suggested in the linguistic context by Church and Hanks (1989).

Oakes (1998:63) reported that co-occurrence statistics such as MI “are slowly taking a central position in corpus linguistics”. MI provides a measure of the degree of association of a given segment with others. Pointwise MI, calculated by Equation 1, is what is used in lexical processing to return the degree of association of two words  $x$  and  $y$  (a collocation).

$$(1) \quad I(x; y) = \log \frac{P(x|y)}{P(x)}$$

The SARA tool, widely used with the BNC, and the Sinica Corpus user interface both offer an MI analysis of the corpus contents. Such tools, however, suffer from two important constraints: first, when considering the context of a word, an arbitrary number of adjacent words to the left or right is taken into account, ignoring discontinuous collocations, which occur when other words (in particular function words like *the* and *of*) are found between the collocation components. To illustrate the problem, imagine that we wish to determine which of two senses of the English word *bank* (“the bank of a river”, or “financial institution”) is more common. If the strings *river bank* and, say, *investment bank* are frequent, there might be enough evidence on which to make a judgment. But such an analysis would ignore *Bank of Taiwan* and *bank of the river*, where the important collocates are not adjacent to the keyword, even though *Taiwan* and *river* stand in the same grammatical relationship to the keyword as *investment* and *river* in the other example.

The second constraint is that a list of collocates of some keyword could include, undistinguished, items of any part of speech (POS: noun, verb and so on) and of any syntactic role (such as subject or object). This sort of grammatical information can provide useful clues for sense discrimination, which standard corpus analyses are unable to take advantage of. Consider again the word *bank*, which has at least two verbal senses, illustrated by *The plane banked sharply* and *John banked the money*. The first of these is an intransitive verb – it cannot take an object. Thus, if an object is observed in the sentence featuring the keyword, the chances are that forms of the verb *bank* properly belong to the second sense. One corpus query tool which overcomes these limitations is the Sketch Engine.

The Sketch Engine is embedded in a corpus query tool called Manatee, and offers a number of modules. There is a standard concordance tool, whose output is very similar to that shown at Figure 1. It allows the user to select, as a keyword, either a lemma (in which case the keyword *bank* would yield results for all of *bank*, *banks* and *banking* for example), or a simple word-form match. The user may also specify the size of the window (the numbers of words to the left and right of the keyword) that he wishes to view. Word frequency counts are also available, and the user may define a subcorpus (in the case of the BNC, on which the English version of Sketch Engine is based, one can choose different parts of the corpus such as fiction or non-fiction).

### **Goals and significance of the work**

Sketch Engine (SkE) is a corpus query tool which accesses large linguistic corpora in a number of languages. It has already been used successfully in lexicographical applications,

but not extensively in second language acquisition. In earlier work, reported in Smith et al (2007) we attempted to evaluate the utility of SkE as a Chinese learning and teaching tool, presenting experiments conducted using Mandarin Chinese second language learners. Informants were interviewed, and pre- and post-testing carried out, to ascertain to what extent they had benefited from the availability of SkE. This was one of a limited number of studies on corpus linguistics in Chinese learning. The results of that work were not conclusive, though, because not enough participants responded to the post-test call.

Many teachers have tried using corpora for class preparation, or even encouraged students to refer to them in their private study. Some teachers, however, have found concordances too unwieldy to be of use; and corpus query tools may pull up word partnerships that are not real collocations, purely because they happen to be adjacent in the text. Sketch Engine collocation information, however, is based on the grammatical relations that obtain between words, not merely the fact that they are neighbors. Thus, given *the police were quick to arrest the five suspects*, the “arrest” word sketch shows “police” as a very salient subject collocate, and the lemma “suspect” as an object collocate, while “quick” and “five” would appear only as very low-ranking collocates. What this means for users is that word sketches give more reliable information about usage, and that because they are quite short, they can be conveniently used in any classroom with computer and projector facilities. In the same way as a teacher can use Google Images to flash up a picture of an object he wants to describe, one can show a word sketch to give students an immediate feel for appropriate usage.

We therefore made the software available to students, encouraging them to use it for vocabulary work, and while reading and writing. Students were encouraged to use SkE to figure out word meanings from context, for example, rather than resorting immediately to the dictionary. Also, where memorization of vocabulary is required, SkE helps the student to see how the words really pattern.

### **Current research status**

We have begun to use Sketch Engine to help develop two pedagogical tools, one for the generation of vocabulary lists, the other for the automatic creation of cloze exercises. Progress so far is promising: we have created corpora on topics of interest to learners from the web, performed some analysis of them, and used the output to create vocabulary lists on certain topics. There is room for refinement before the lists can actually be adopted in teaching practice, and they will need to be integrated into other teaching materials. The PI will be making more and more use of the lists over the year, as well as influencing colleagues to do the same.

### **Results**

The second tool, for generating cloze exercises, requires more algorithmic input from the PI. He has already demonstrated in published work how the components of the Sketch Engine can be harnessed to complete the task, but the important and time-consuming part remains: encoding this into a formal algorithm, and writing an implementation (in a mixture of Python and Java). This is clearly a non-trivial task, but once it is done, it will be made available, free of charge, to the Sketch Engine user community, and other communities of language teachers and learners.

In the longer term, it is hoped that the Sketch Engine could form part of a Chinese CALL (Computer assisted language learning) platform, for the benefit of foreign learners. It could also be adapted for native Chinese elementary school students, who are beginning to learn writing skills. We already have a demonstration walkthrough and pre-test questions: these could be extended to form the basis of a workbook and quizzes.

Over the year, four conference presentations were made, and articles written: Smith et al (2008a, 2008b, 2008c, 2008d). Two were domestic, and two international; the two international papers are appended to the present report. There is now enough material for a more lengthy account of the work, and this year at least two journal articles will be presented. One will be to a mainstream TESOL journal such as *TESOL Quarterly*, the other to a technical journal, probably *Language Learning & Technology*.

Although NSC funding has been refused for the current academic year, the work will continue: in fact, the non-availability of funds, precluding conference submissions, will make the successful submission of journal articles even more probable!

The successful use of corpora could make a real difference to the way language is taught in Taiwan. The laborious grammatical explanations, lists of sentence patterns using invented examples, and lists of vocabulary items could often be supplanted by the use of real language data. Ultimately, this could usher in an adjustment of perceptions about the great difficulty of acquiring a second language; we are confident that it would result in a more enjoyable and more fruitful language learning experience.

### **Bibliography**

黃居仁，主編 (2003) 中文的意義與詞義之一/之二/之三。中央研究院文獻語料庫與詞庫小組技術報告 03-01/03-02/03-03。

中央研究院資訊所、語言所詞庫小組。1995/1998。「中央研究院漢語料庫的內容與說明。」詞庫小組技術報告 95-02/98-04 號。

史尚明,巫宜靜, 2005. 中文搭配詞語搜尋介面之設計與應用, 第四屆全球華文網路教育研討會, Taipei, pp 19-27

吳毓傑、陳振南 (2002)以叢聚式作法進行中文新聞分類，第八屆國際資訊管理研究暨實務研討會，Vol. 1, pp.619-627

Agirre, E. and Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the Coling 2000 Workshop on Semantic Annotation and Intelligent Annotation, Centre Universitaire, Luxembourg.

Bruce, R. and Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994), pages 139-146, Las Cruces.

BTExaCT (2001) OASIS First Utterance corpus, version 2.23 (annotated transcription, audio files and release notes)

Carroll, J. and McCarthy, D. (2000). Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities*, 34(1-2):109-114.

Chen, J. N. (2000) Adaptive Word Sense Disambiguation Using Lexical Knowledge in Machine-readable Dictionary, *Computational Linguistics and Chinese Language Processing*, Vol. 5, No. 2, pp. 1-42.

Chen, Jen-Nan and Sue J. Ker. (2001) Towards a Conceptual Representation of Lexical Meaning in WordNet. In the 15th Pacific Asia Conference on Language, Information and Computation, pp. 97-108, Hong Kong, February 1-3

Chodorow, M., Leacock, C., and Miller, G. (2000). A topical/local classifier for word sense identification. *Computers and the Humanities*, 34(1-2):115-120.

Choueka, Y. and Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the*

- Humanities, 19:147-158.
- Church, K. W. and Hanks, P. (1989) Word association norms, mutual information and lexicography. In *Proc. 27th Annual Meeting of ACL*, Vancouver. 1989: 76-83
- Clear, J.H. (1993) The British National Corpus in Paul Delany & G. P. Landow (eds) *The Digital Word : text-based computing in the humanities* . Cambridge, Mass.: MIT Press, : 163-187.
- Cowie, J., Guthrie, J., and Guthrie, L. (1992). Lexical disambiguation using simulated annealing. In Proceedings of the 15th International Conference on Computational Linguistics (Coling 1992), pages 359-365, Nantes.
- Crystal, D (1991) *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.
- Dagan, I. and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563-596.
- Dagan, I., L. Lee, and F. Pereira. (1999) Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1)
- Dini, L., di Tomaso, V., and Segond, F. (2000). GINGER II: An example-driven word sense disambiguator. *Computers and the Humanities*, 34(1-2):121-126.
- Edmonds, P. and Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering, Special Issue on Word Sense Disambiguation Systems*, 8(4):279-291.
- Ellis, Rod. Modified oral input and the acquisition of word meanings. In *Applied Linguistics* 16/4 (1995), Pp. 409-441.
- Escudero, G., Marquez, L., and Rigau, G. (2000b). A comparison between supervised learning algorithms for word sense disambiguation. In Proceedings of the 4th Conference on Computational Natural Language Learning (CoNLL-2000), pages 31-36, Lisbon.
- Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Firth, J.R. (1957) A synopsis of linguistic theory, 1930-1955. In Palmer, F.R. (ed) (1968) *Selected papers of J.R. Firth 1952-9*. Harlow: Longman
- Gale, B., Church, K., and Yarowsky, D. (1992b). A method for disambiguating word senses in a corpus. *Computers and the Humanities*, 26:415-439.
- Gale, B., Church, K., and Yarowsky, D. (1992d). Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54-60, Cambridge.
- Garner, P (1997) On topic identification and dialogue move recognition. In *Computer Speech and Language* 11(4): 275-306
- Huang, Chu-Ren, Adam Kilgarriff, Yiching Wu, Chih-Min Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-jiann Chen. '2005. Chinese Sketch Engine and the Extraction of Collocations,' Presented at the Fourth SigHan Workshop on Chinese Language Processing, October 14-15, Jeju, Korea.
- Huang, Chu-Ren, I-Ju E. Tseng, Dylan B.S. Tsai and Brian Murphy. 2003. *Cross-lingual Portability of Lexical Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations*. *Language and Linguistics*. 4.3.509-532.

- Huang, Chu-Ren, Keh-jiann Chen, and Lili Chang. 1997. Segmentation Standard for Chinese Natural Language Processing. *Computational Linguistics and Chinese Language Processing*. 2.2.47-62.
- Huang, Chu-Ren, Zhao-ming Gao, Claude C.C. Shen, and Keh-jian Chen. 1998. Quantitative Criteria for Computational Chinese Lexicography: A Study based on a Standard Reference Lexicon for Chinese NLP. *Proceedings of ROCLING XI*. 87-108.
- Huang, Chu-Ren. 1995. Observation, Theory, and Practice: The application of corpus-based linguistic studies in Chinese language teaching. Presented at the First International Conference on New Technologies in Chinese Language Teaching. San Francisco. April 27-30, 1995
- Ide, N., Macleod, C. (2001). The American National Corpus: A Standardized Resource of American English. *Proceedings of Corpus Linguistics 2001*, Lancaster UK.
- Ker, S.J. and Jen-Nan Chen, "Adaptive Word Sense Tagging on Chinese Corpus", In *Proceedings of 18th Pacific Asia Conference on language, Information and Computation*, Tokyo, Japan, pp. 267-274, Dec 2004.
- Kilgarriff, A & Tugwell, D (2002) Sketching Words. In Marie-Hélène Corréard (ed.): *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. Euralex 2002.
- Kilgarriff, A, & D Tugwell (2001) WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In Proc. workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation", pp.32-38. 39th ACL & 10th EACL, Toulouse, July 2001.
- Kilgarriff, A, Rychly, P, Smrz, P & Tugwell, D (2004). The Sketch Engine, in *Proceedings of EURALEX*, Lorient, France, July 2004
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31:97-113.
- Kilgarriff, Adam, and David Tugwell. (2001) "WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation". Proc. of MT Summit VII, Santiago de Compostela, pp.187-190.
- Kilgarriff, Adam, Chu-Ren Huang, Michael Rundell, Pavel Rychly, Simon Smith, David Tugwell, Elaine Uí Dhonnchadha. "Word Sketches for Irish and Chinese," Presented at *Corpus Linguistics 2005*. July 14-17. Birmingham, UK.
- Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, David Tugwell (2005). Chinese Word Sketches. ASIALEX 2005. In *Words in Asian Cultural Context*, June 1-3, Singapore.
- Lam, H.C., K.H. Pun, S.T. Leung, S.K. Tse, and W.W. Ki (1993) Computer Assisted Learning for Learning Chinese Characters. In *Communications of COLIPS: Vol. 3, No. 1*, 31-44.
- Leacock, C., Chodorow, M., and Miller, G. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147-165.
- Lee, Y. K. and Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 41-48, Philadelphia.

- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC Conference*, pages 24-26, Toronto.
- Lin, D. (1998) Using collocation statistics in information extraction. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*.
- Miller, G., Richard Beckwith, Christian Fellbaum, Derek Gross, and Katherine Miller. Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory, Princeton University, August 1993.
- Oakes, M (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press
- Pearce, D. (2001) Synonymy in collocation extraction. In *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, CMU
- Shimohata, S., T. Sugio, and J. Nagata. (1997). Retrieving collocations by co-occurrences and word order constraints. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, pages 476-81, Madrid, Spain.
- Smith, S. (1999) Discontinuous compounds in Mandarin Chinese: A lexicalization algorithm. Unpublished MSc dissertation, UMIST, Manchester.
- Smith, S. (2003) Predicting query types by prosodic analysis. Unpublished PhD dissertation, University of Birmingham.
- Smith, Simon, Chu-Ren Huang, Adam Kilgarriff, Mei-Rong Chen, 2007. "A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine." 2007 Practical Applications In Language And Computers, Łódź , Poland.
- Smith, S, Scott Sommers, Adam Kilgarriff (2008a) "Learning words right with the Sketch Engine: Meaningful lexical acquisition from corpora and the web." 2008 CamTESOL conference, Phnom Penh.
- Smith, S, Adam Kilgarriff, Scott Sommers (2008b) "Making better wordlists for ELT: Harvesting vocabulary lists from the web using WebBootCat." 2008 Conference and Workshop on TEFL and Applied Linguistics, Taoyuan.
- Smith, S, Adam Kilgarriff, Scott Sommers (2008c) "Learning words right with the Sketch Engine and WebBootCat: Automatic cloze generation from corpora and the web." 25th Conference of English Teaching and Learning in R.O.C. National Chung Cheng University, Chiayi (May).
- Smith, S., Scott Sommers, Adam Kilgarriff (2008d) "Automatic cloze generation: getting sentences and distractors from corpora." 8th Teaching and Language Corpora Conference, Lisbon (July).
- Speelman, D. (1997). *Abundantia Verborum. A computer tool for carrying out corpus-based linguistic case studies*. Doctoral dissertation Katholieke Universiteit Leuven.
- Tugwell, David and Adam Kilgarriff. (2000) "Harnessing the Lexicographer in the Quest for Accurate Word Sense Disambiguation" *Proc. 3rd Int. Workshop on Test, Speech, Dialogue (TSD 2000)*, pp.9-14. Brno, Czech Republic Springer Verlag Lecture Notes in Artificial Intelligence
- Vossen P. (ed.). (1998). *EuroWordNet: A multilingual database with lexical semantic networks*. Norwell, MA: Kluwer Academic Publishers.



- Wilks, Y. and Stevenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4(2):135-144.
- Wu, Yiching, Simon Smith & Chu-ren Huang, 2005. How to express 'express'? – Pedagogical reflections from web resources, March 2005 Conference and Workshop on TEFL and Applied Linguistics, Taoyuan
- Wu, Y. and Zhang, J. (2004) A Chinese Language Expert System Using Bayesian Learning. In *Proc of the Eighth World Multiconference on Systemics, Cybernetics and Informatics*, Florida. 90-95.
- Xiao, R and T McEnery, 2006. Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27: 103-129

### **Relevant websites**

- British National Corpus, BNC. <http://info.ox.ac.uk/bnc/>
- Sinica Balanced Corpus <http://www.sinica.edu.tw/SinicaCorpus/>
- Sketch Engine <http://www.sketchengine.co.uk>
- WordNet. <http://www.cogsci.princeton.edu/~wn/>
- Linguistic Data Consortium <http://wave ldc.upenn.edu/>
- Sketch Engine Walkthrough and Pre-Test <http://mcu.edu.tw/~ssmith/walkthrough>
- Appendix 1: Smith et al (2008a)

## **Learning words right with the Sketch Engine and WebBootCat: Meaningful lexical acquisition from corpora and the web**

**Simon Smith<sup>\*</sup>, Scott Sommers<sup>\*</sup> and Adam Kilgarriff<sup>†</sup>**

<sup>\*</sup>English Language Center, Ming Chuan University

<sup>†</sup>Lexical Computing Ltd, UK

ssmith@mcu.edu.tw

### **Abstract**

In Taiwan, and other Asian countries, students of English expect and are expected to memorize a lot of vocabulary: Ming Chuan University, for example, relies fairly heavily on vocabulary acquisition and retention in its teaching and testing resources. Oftentimes, lists of vocabulary items to be learned by students do not really belong to a particular topic, or fit it very loosely, because the items have not been chosen in a principled way.

The present paper reviews the arguments for incidental learning and direct learning of vocabulary in ELT, and shows how a web corpus builder (WebBootCat) can be used to build lists of words that are related to a particular topic in an intuitive and statistically principled way. A small number of seed search terms are used by WebBootCat to generate a corpus of texts on a given topic, and this corpus is searched to find vocabulary items which are salient to the topic.

### **Introduction**

In many Asian nations, including Taiwan and Cambodia, educational and career advancement often turns on test performance. Whether it be entrance tests at schools or companies, or language proficiency tests like TOEFL or IELTS needed for study abroad, performance on tests can play an important and life-changing role.

Tutors, textbooks, and commercial cram schools that prepare students for high stakes tests have traditionally relied on the rote memorization of lists of words as a teaching method. This method in various forms has become a standard preparation technique for students who face language proficiency or ability tests. The words on such lists are frequently selected for reasons unrelated to their usefulness. But it is clearly important which words are chosen to be on these wordlists, and which words are selected to be taught and used in textbooks, if learners are to acquire language that is meaningful and useful.

In this paper, we first review the arguments for incidental learning and direct learning of vocabulary, and consider how they are played out in English teaching in Taiwan. We consider one particular textbook, and find that the vocabulary is not systematically selected, with the vocabulary to be learnt not forming a good match either to the topic of the chapter, or to the reading material, or to corpus frequency. We report experiments with WebBootCat (WBC), a software tool which uses Yahoo! web services to harvest linguistic corpora on user-specified subject areas from the World Wide Web. We use WBC to extract from these corpora key vocabulary which can be used to populate wordlists in textbook-writing.

### **Vocabulary: incidental and direct acquisition**

Studies in the acquisition of vocabulary have identified two principal learning strategies, incidental learning (discussed by Nagy, Anderson & Hermann, 1985; Nation & Coady, 1988; Nation, 2001) and direct learning. Research by Nagy and colleagues claimed that learning from context is one of the most significant aspects of incidental learning. This laid the groundwork for the belief that *authentic* context is a particularly powerful source of incidental language learning (Krashen, 1989; Pitts, White and Krashen, 1989).

There is little doubt that incidental learning, particularly that acquired through reading, is key to learning the vocabulary necessary for functioning in an English environment. Some researchers, however, have argued that this form of acquisition has limitations, especially for students taking academic courses delivered in English, who need to develop textbook reading skills, and the ability to follow lectures (see Chaffin, 1997; Zechmeister et al, 1995). These researchers claim that direct instruction of vocabulary and meaning plays a central role. Without this, they believe, long-term retention of new vocabulary is unlikely to follow. The strategy they advocate emphasizes the role of dictionaries and other word reference books; they note, too, that direct instruction is important in fostering an interest in words.

Direct acquisition studies recognize that vocabulary can be learnt using tools that bring the learner's attention into direct contact with the form and meaning of words, such as dictionaries and vocabulary lists. However, the question of how best to use these tools for direct vocabulary acquisition remains unanswered. In Taiwan, and other parts of Asia, the traditional (and intuitively suboptimal) approach has been simply to memorize the vocabulary item along with one or two possible L1 translations.

The memorization of vocabulary items has become the usual method by which students in Taiwan prepare for standardized tests of English proficiency. Ironically, government policies intended to boost the national standard of communicative language skills have actually encouraged this approach to language learning. Previously, lists of words were presented primarily to students in public secondary schools, but nowadays official attempts to promote language proficiency have resulted in the widespread use of proficiency tests such as the GEPT and TOEIC; consequently there has been an explosion of test preparation classes. In almost every case, these classes emphasize vocabulary acquisition through the memorization of lists rather than the use of communicative tasks or the presentation of authentic examples.

Typically, these lists incorporate vocabulary selected by employees and teachers of test preparation schools. In more professional situations, the selections are derived from word

counts of actual standardized tests. In other cases, the lists are created in a fashion that is more or less arbitrary, with only an unclear match between the items on a given list and the topic it is supposed to represent. Furthermore, items are often demonstrated to students using contrived examples. With such poor models of usage available to students, it is questionable whether even the highest standard of instruction will result in the desired acquisition.

If students are to learn lists of English words, it would be better if the lists at least contain words that are useful and relevant. It is of course the purpose of lists such as the CEEC list (a glossary of 6480 words used to help people studying for university entrance exams, described and listed in College Entrance Examination Center (2002)) to cover such useful vocabulary. However no systematic strategy for doing so is universally accepted. Instead, a variety of strategies have been adopted for reasons of convention. One strategy involves the identification of the most common words in a general corpus of English. The most common words are then judged as the most useful. This approach has been taken in Japan, and in 2003 the widely-used JACET list of 8000 basic words was revised substantially on the basis of the British National Corpus (Masamichi 2003, Uemura 2005). Su (2006) has explored the relation between (a 2000 word version of) the CEEC list and a range of other lists and corpora. While the opinion of Su is that the list is largely satisfactory, areas are found in which the corpora and the list do not match.

An essential difference between corpus-derived lists and those compiled manually, whether by individual teachers or government bodies, is that data from corpora is authentic. Such measures as personal intuition or experience of the teacher are far too problematic to produce meaningful results, according to Biber & Conrad (2001). Careful statistical examination of corpus data, however, can help us to construct meaningful, topic-related wordlists.

### **English vocabulary acquisition at Ming Chuan University**

Two of the authors, Smith and Sommers, are employed by the English Language Center (ELC) of Ming Chuan University, where the principal task is to teach general English skills to large groups (around 60) of relatively unmotivated university students. English is taught throughout the four years of a typical undergraduate career (in contrast to many Taiwan institutions where one or two years is the norm). There is little evidence to show how much acquisition of English takes place over the four year period, but certainly there is ample time for boredom to set in students who are principally interested in the taught offerings of their home departments.

The ELC's students are assessed twice a semester by centralized achievement tests. Much of the teaching revolves around communicative principles and as such the teaching of grammar is not a central theme in most instruction or in the assessment of students. Instead, the main focuses of these tests are listening comprehension, and familiarity with the unit vocabulary items. Students generally do not prepare for listening comprehension assessment. The most common form of preparation that teachers observe is vocabulary memorization. This is done by memorizing unit vocabulary lists and internalizing each item with its Chinese "equivalent".

The primary teaching material for these courses is an in-house textbook series called *East Meets West*. *EMW* presents some topics relevant to students' lives and potential future careers, and others which are less relevant or useful. There are a number of different types of activity in each unit, but the standard layout involves a reading on a specific topic (written by an ELC teacher), and a collection of about 12-14 vocabulary items selected from the text. These words are chosen for their difficulty and it is assumed that they are new to the students. They are not necessarily related to the unit topic. In addition, units sometimes contain

exercises and activities that do not have an intuitive relationship with the topic.

The first unit of *EMW 1* is entitled “Getting started at university”, an apparently appropriate topic for beginning freshmen. There is a short reading on the experience of an imaginary freshman called Patricia Lin, reading comprehension questions, pronunciation exercises, pattern practice and a couple of listening exercises, along with a vocabulary section. There are also, as in other units, some activities specifically related to the topic: maps of the MCU campus, of use to new students; locations of MCU departments; suggested English spellings of Chinese family names etc.

The list of vocabulary items from this chapter is shown in Figure 1. In this case, many items seem to have no relationship to “Getting started at university”, or to “university”, or indeed to getting started at anything at all.

<b>Vocabulary</b>				
<u>Nouns</u>				
attendance	course	facilities	helmet	
initiative	major	vendor		
<u>Verbs</u>				
accomplish	consider	improve	tease	
<u>Adjectives</u>				
challenging	fortunate	impatient	occasional	protective

**Figure 1** *EMW* Unit 1 vocabulary

Only three of the words – all nouns – have an obvious connection to an educational topic. The first verb and the first adjective are also likely to occur more often in educational contexts.

The reason for the irrelevant selection of vocabulary lies in its selection method. First, a topic-related text is commissioned (in this case the story about “Patricia Lin”) with no requirement to incorporate topic-related vocabulary into the text. Next, items are selected (in most cases, not by the text writer, but by another editor) which are deemed unfamiliar to students or that they ought to learn. Many of the apparently on-topic items which occurred in the texts (*student, university* and so on) were ruled out, because the learners would already know them; instead, words from the texts have been chosen seemingly at random. Learners are expected to be familiar with this vocabulary in the midterm and final tests.

This seems an unprincipled approach to vocabulary acquisition. One might argue that a better approach might have been to write a text around a list of pre-determined vocabulary items, related to the unit topic. Creating such a list is not a trivial task, though; it is difficult to determine what sort of vocabulary *should* be included. Textbook writers cannot produce such a list through contemplation and introspection alone. It might be possible to think of a short list of educational terms (*major, sophomore, classmate, campus* and the like), and a reading text featuring that vocabulary could then be commissioned. However, at least two objections could be raised to that approach.

First, the list would only include items that belong to the domain in the most transparent way. If, for example, it can be shown that items such as *excited, challenging* and *friend* occur more often in texts about “Getting started at university” than they do in texts on other topics, they are candidates for inclusion in our lists.

Secondly, it would be less straightforward to compile such a list for Unit 2 (“Family and hometown”) or Unit 3 (“English learning and you”), to give just two examples. In these

domains, only kinship terms and the jargon of TESOL and Applied Linguistics spring to mind, and neither of these would be useful for MCU freshmen.

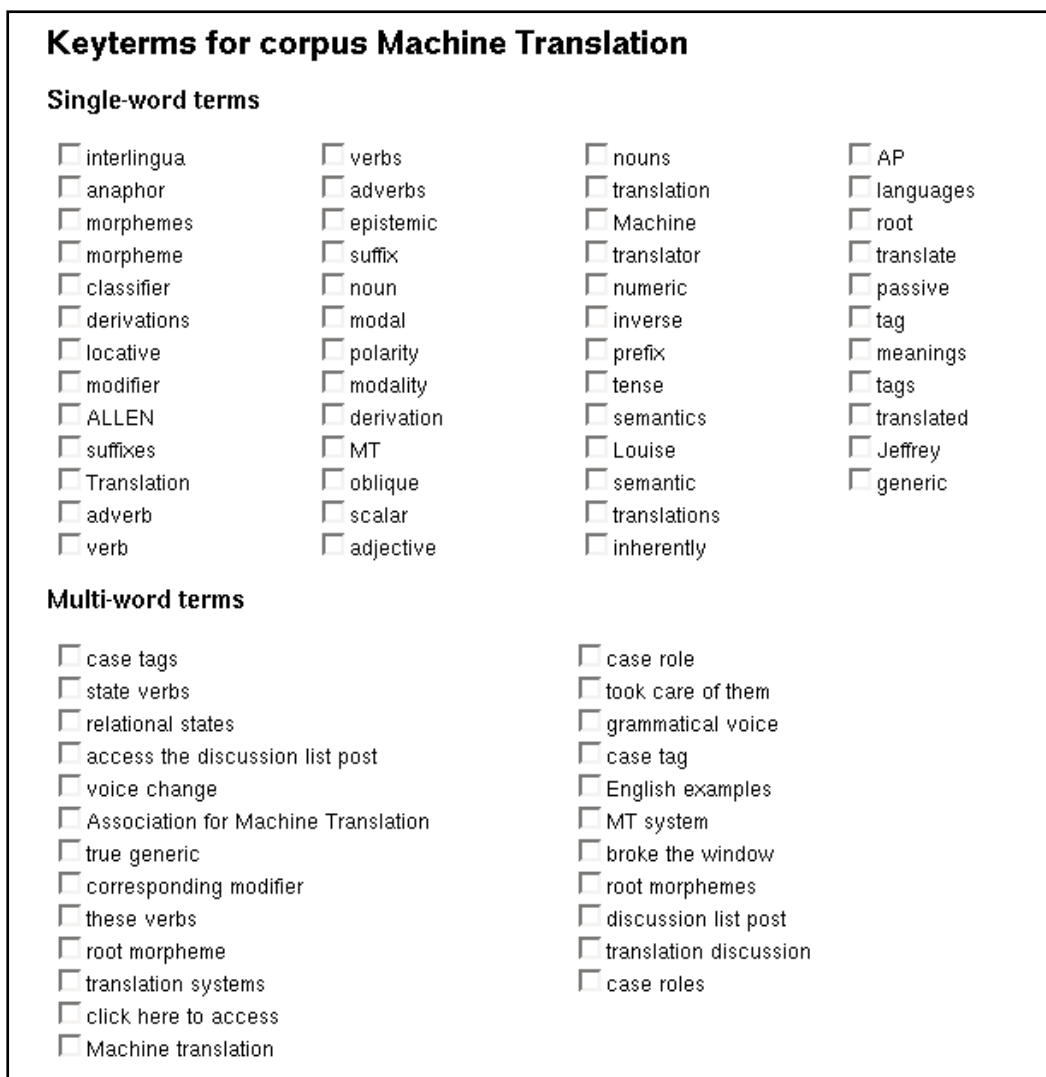
What is needed is a corpus-based vocabulary generation tool.

### **WebBootCat, a tool for corpus and wordlist generation**

Baroni et al (2006), in a paper which introduces WBC, focused on the tool's utility as an aid to technical translators. Most translators, Baroni et al note, make regular use of the web as a source of information about technical terms and usages; however, search engine design is not optimized for their use.

The task described in the paper consists of creating a corpus associated with a particular domain, and generating a list of the terms most salient to the domain. All of this information is extracted from the web. The resulting corpus can be expected to be both up to date (the terminology is current), and to be firmly focused on the domain in question (in contrast to offline corpora, such as the BNC, intended for general use).

The basic algorithm is conceptually simple. First, a search is seeded with one or more words selected by the user. These **seed words** are sent to Yahoo! (formerly Google was used, as mentioned in Baroni et al's paper), and all the lexical items are extracted from the returned web pages. A substantial amount of filtering is done to exclude web pages which do not mostly contain running text of the language in question. Measures include rejecting pages containing too many words held on a stop list, and very short and excessively large web pages: a user interface provides control over these filters. The resulting corpus may be used in a number of ways. It can be explored in the Sketch Engine, a leading corpus query tool (Kilgarriff et al 2004). The user can also generate keyword lists from it: to do this, all words in the corpus are counted and their frequencies are compared with their frequencies in a general web corpus (the **reference corpus**). A list of the words whose frequencies are most significantly higher in the reference corpus is created. Baroni et al used WBC to generate the list of **keyterms** related to Machine Translation shown in Figure 2. Most, but not all, of the terms are indeed related to that domain in some way. Similar lists of vocabulary could also be generated on topics of interest to language learners.



**Figure 2** WBC output (from Baroni et al 2006)

### Generating vocabulary lists with WBC

The reader probably will already have compared Figure 2 (the list of keywords related to Machine Translation, generated by WBC) with the vocabulary list (Figure 1) on “Getting started at university”, developed by ELC curriculum writers, and drawn the conclusion that the former contains many relevant items, the latter precious few. Figure 3 shows the keywords extracted for a query to WBC, using the seed words *freshman* and *university*, and searching 100 websites which feature those words more prominently than other sites

A glance at the figure shows that almost all of the words extracted are salient for the domain. Many terms such as *graduation*, *SAT*, and *transcripts* are part of the specialized vocabulary of tertiary education; *courses* and *results* probably are not, but are more frequent in that domain than elsewhere.

<input checked="" type="checkbox"/> admission (47)	<input checked="" type="checkbox"/> SAT (19)	<input checked="" type="checkbox"/> enrollment (7)	<input type="checkbox"/> website (7)
<input checked="" type="checkbox"/> University (44)	<input checked="" type="checkbox"/> academic (16)	<input checked="" type="checkbox"/> graduation (6)	<input checked="" type="checkbox"/> Tests (5)
<input checked="" type="checkbox"/> school (40)	<input type="checkbox"/> complete (12)	<input type="checkbox"/> copy (8)	<input type="checkbox"/> based (11)
<input checked="" type="checkbox"/> Students (29)	<input checked="" type="checkbox"/> ACT (16)	<input checked="" type="checkbox"/> transcripts (6)	<input type="checkbox"/> fee (4)
<input type="checkbox"/> required (32)	<input type="checkbox"/> year (19)	<input checked="" type="checkbox"/> freshman (6)	<input checked="" type="checkbox"/> essay (4)
<input type="checkbox"/> high (31)	<input checked="" type="checkbox"/> student (13)	<input type="checkbox"/> completed (11)	<input checked="" type="checkbox"/> schooled (4)
<input checked="" type="checkbox"/> College (33)	<input type="checkbox"/> Office (12)	<input type="checkbox"/> note (7)	<input checked="" type="checkbox"/> degree (6)
<input type="checkbox"/> must (28)	<input checked="" type="checkbox"/> programs (13)	<input type="checkbox"/> recommended (7)	<input type="checkbox"/> evaluated (4)
<input checked="" type="checkbox"/> application (30)	<input checked="" type="checkbox"/> program (13)	<input checked="" type="checkbox"/> score (12)	<input type="checkbox"/> meet (7)
<input checked="" type="checkbox"/> applicants (20)	<input type="checkbox"/> minimum (14)	<input checked="" type="checkbox"/> teacher (6)	<input type="checkbox"/> below (6)
<input checked="" type="checkbox"/> submit (24)	<input type="checkbox"/> official (11)	<input type="checkbox"/> selected (4)	<input type="checkbox"/> requirement (6)
<input checked="" type="checkbox"/> scores (19)	<input checked="" type="checkbox"/> courses (13)	<input type="checkbox"/> first (10)	<input type="checkbox"/> international (5)
<input type="checkbox"/> requirements (16)	<input checked="" type="checkbox"/> test (8)	<input checked="" type="checkbox"/> mathematics (6)	<input checked="" type="checkbox"/> applications (4)
<input type="checkbox"/> Boston (23)	<input type="checkbox"/> Board (8)	<input type="checkbox"/> instructions (6)	<input type="checkbox"/> documents (4)
<input checked="" type="checkbox"/> Admissions (19)	<input checked="" type="checkbox"/> counselor (8)	<input checked="" type="checkbox"/> Education (5)	<input type="checkbox"/> check (4)
<input type="checkbox"/> please (14)	<input type="checkbox"/> Common (10)	<input type="checkbox"/> consideration (4)	<input type="checkbox"/> visit (4)
<input type="checkbox"/> you (45)	<input type="checkbox"/> should (13)	<input checked="" type="checkbox"/> study (6)	<input type="checkbox"/> directly (6)
<input type="checkbox"/> English (16)	<input type="checkbox"/> State (7)	<input type="checkbox"/> personal (5)	<input checked="" type="checkbox"/> grades (4)
<input type="checkbox"/> following (15)	<input type="checkbox"/> above (11)	<input checked="" type="checkbox"/> laboratory (4)	<input checked="" type="checkbox"/> TOEFL (13)
<input type="checkbox"/> your (29)	<input type="checkbox"/> years (16)	<input type="checkbox"/> Standards (6)	<input type="checkbox"/> second (12)
<input type="checkbox"/> considered (14)	<input checked="" type="checkbox"/> GED (9)	<input type="checkbox"/> writing (5)	<input type="checkbox"/> sent (7)
<input type="checkbox"/> information (17)	<input checked="" type="checkbox"/> Arts (9)	<input checked="" type="checkbox"/> history (6)	<input type="checkbox"/> online (4)
<input type="checkbox"/> may (19)	<input type="checkbox"/> contact (7)	<input checked="" type="checkbox"/> Subject (6)	<input checked="" type="checkbox"/> credentials (4)
<input type="checkbox"/> apply (14)	<input checked="" type="checkbox"/> Studies (7)	<input checked="" type="checkbox"/> secondary (6)	<input checked="" type="checkbox"/> results (7)
<input type="checkbox"/> language (15)	<input checked="" type="checkbox"/> science (9)	<input checked="" type="checkbox"/> coursework (6)	<input type="checkbox"/> applying (7)

**Figure 3** WBC keywords for corpus seeded with *freshman* and *university*

The second unit of *EMW* is called “Family and Hometown”. The title is a reasonable description of the contents of the unit, which are designed to get students to share, using the target language, information about their backgrounds. The two keywords featured in the unit title seemed a reasonable point of departure for generating a vocabulary list; this was done, and the result is shown in Figure 5. This may be compared with Figure 4, which shows the vocabulary prescribed for that unit of *EMW*. This vocabulary is barely concerned with the topic at hand at all – this comes as no surprise when it is known that the list was extracted from a story about one person’s life (albeit a very interesting story).

<b>Vocabulary</b>			
<u>Nouns</u>			
lightning	orphan	porch	region
roots	suburb	tragedy	twin
<u>Verbs</u>			
support			
<u>Adjectives</u>			
agricultural	polluted	urban	

**Figure 4** *EMW I* Unit 2 vocabulary

The picture from Figure 5, however, appears just as bleak. Only 10 items have been

found: 2 of those are the originally specified seed words, and of the rest, only 4 could be said to relate to the topic. The search was performed in exactly the same way as in the *freshman/university* case, again querying 100 websites.

We should not be too disappointed at such a sparse list of keyterms; *freshman* and *university* are simply much better topic descriptors than *family* and *hometown*. A comparison with a standard Google search is instructive: all but two hits from the first page of a Google search for *freshman university* are official university web pages dealing, precisely, with the issue of “getting started at university”. Equivalent Google results for *family* and *hometown* link to all manner of things, including a genealogical site, articles about disabled children and the Tour de France, and advertisements for real estate and a used Barbie Doll set. *University* and *freshman* are more powerful as a pair of search terms (recall from Baroni et al’s results, given in Figure 1, that the same is true of *machine translation*).

Intuitively, the more specific a term is (the less polysemous it is, and the further down a hierarchical hyponym tree it is found), the more powerful it will be. Thus, a term like *person*, which would be close to the top of such a tree, is less powerful than the more specific term *freshman*. *Family* is a polysemous item which can refer to related groups of people or of other entities, and it is high up in the semantic hierarchy. On both counts, therefore, *family* offers less specificity than *freshman*, and consequently is less powerful as a search item.

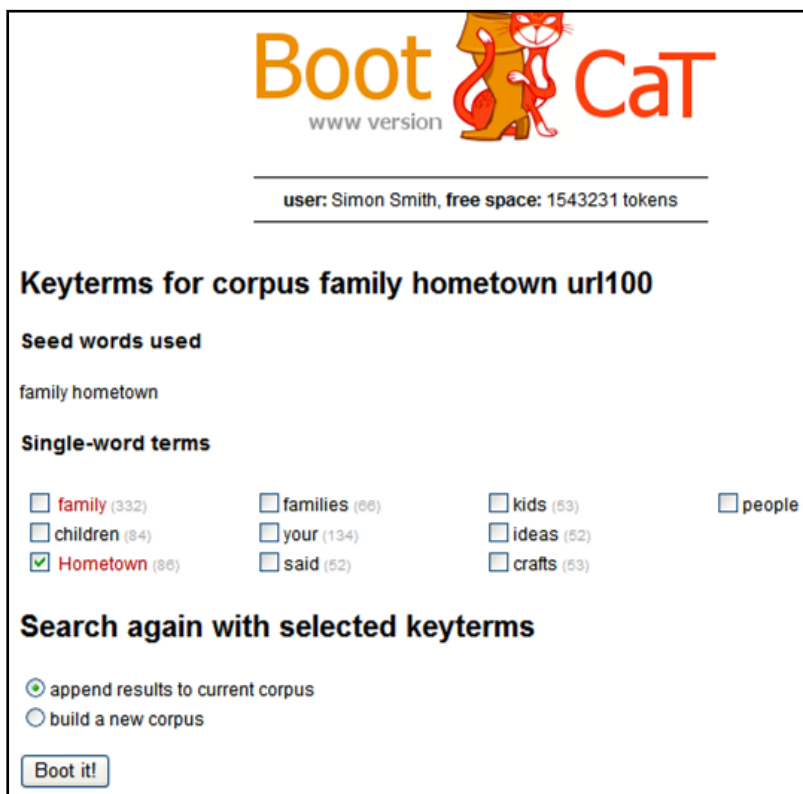


Figure 5 WBC keywords for corpus seeded with “family” and “hometown”

### WBC Business Corpus

A number of *EMW* units deal with the world of business and international trade, especially in the senior year of the course. A good wordlist in that domain, therefore, would be particularly useful. Rather than generate a new corpus for the purpose of this study, we used an existing, much larger, WBC-generated Business Corpus, of about 10 million words.



For comparison, note that the size of the Machine Translation corpus created by Baroni et al (1986), used to generate the wordlist given in Figure 1, consisted of around 144,000 words, and the corpora we have so far described averaged about the same size. To generate a larger corpus, a larger number of seed words is selected. The Business Corpus was seeded with 50 words, selected by Kilgarriff on the basis of their intuitive relevance to the world of business, including *investment, capital, franchise* and *portfolio*.

The larger the corpus, the more documents salient to the subject area it will contain, and the better our chances of generating a good wordlist. The evidence from the Business Corpus bears these expectations out. Words found in the corpus were ranked by the ratio of the number of occurrences to the number of occurrences in a reference corpus, the 100m word BNC. Thus, given the relative size of the corpora, one would expect a non-business term (a word whose frequency in a business or general corpus is about the same) to be assigned a ratio of about 0.1. In the Business Corpus, around 20% of words have relative distribution ratios of 0.1 or above. The top 100 words are ranked by relative distribution ratio in Figure 6.

The reader will probably agree that almost all of the terms are of immediate relevance to the world of business and trade. 26 of them are not found in the Taiwan CEEC list; tellingly, these missing terms (marked “no” in Figure 6) are among the most intuitively relevant to the subject on our list.

It will by now be clear that corpus-derived wordlists are much more likely to succeed in representing a subject area than those compiled manually. If, however, lists such as the CEEC are to continue to serve as a curricular gold standard, it will be useful to learners if vocabulary items are classified as on- or off-list. The learner will then know whether they were exposed to a given item before (or perhaps whether it is likely to come up in an exam).

	<b>Ratio</b>	<b>In CEEC list?</b>		<b>Ratio</b>	<b>In CEEC list?</b>
franchise	3.08	no	asset	0.71	yes
license	2.26	yes	offering	0.69	yes
broker	1.59	no	percent	0.68	yes
commodity	1.57	yes	receipt	0.67	yes
prior	1.3	yes	prohibit	0.67	yes
fiscal	1.25	no	trading	0.66	yes
portfolio	1.05	no	underlie	0.65	no
bond	1.03	yes	program	0.64	yes
paragraph	0.97	yes	behalf	0.62	yes
equity	0.96	no	prescribe	0.62	yes
disclosure	0.94	yes	saving	0.62	yes
applicable	0.92	yes	regulatory	0.62	no
forth	0.9	yes	compliance	0.61	no
investor	0.89	no	investment	0.61	yes
shall	0.87	yes	stock	0.61	yes
transaction	0.87	yes	fee	0.61	yes
entity	0.85	no	contractor	0.58	yes
registration	0.84	yes	invest	0.58	yes
re	0.81	no	liability	0.58	no
exempt	0.78	no	dividend	0.58	no
faculty	0.78	yes	accounting	0.58	yes
designate	0.77	yes	provider	0.57	no
deem	0.74	yes	specified	0.57	yes
accord	0.72	yes	maturity	0.57	yes

	<b>Ratio</b>	<b>In CEEC list?</b>
exemption	0.56	no
expense	0.55	yes
terminate	0.55	yes
competent	0.55	yes
default	0.54	no
purchaser	0.54	no
purchase	0.54	yes
restricted	0.51	yes
amend	0.51	no
addition	0.51	yes
security	0.51	yes
eligible	0.51	yes
corporation	0.49	yes
obligation	0.49	yes
applicant	0.48	yes
renewal	0.48	no
employee	0.48	yes
fund	0.47	yes
prospective	0.46	yes
seller	0.46	yes
registered	0.46	yes
preferred	0.46	yes
lawyer	0.45	yes
counsel	0.44	yes
dealer	0.44	yes
shareholder	0.43	no
delivery	0.43	yes
portion	0.43	yes
enforcement	0.43	yes
sub	0.43	no
submit	0.42	yes
hearing	0.42	no
disclose	0.42	yes
appointment	0.41	yes
payment	0.41	yes
specify	0.41	yes
jurisdiction	0.4	no
revise	0.4	yes
selling	0.39	yes
compensation	0.39	yes
administrative	0.39	yes
written	0.39	yes
incur	0.39	no
certificate	0.38	yes
adviser	0.38	yes
hedge	0.37	yes
assign	0.37	yes
comply	0.37	no
retail	0.37	yes
respondent	0.37	no

**Figure 6** Business Corpus, top 100 terms ranked by ratio to BNC frequency

**Recursive bootstrapping with WBC: generating a second corpus from the first**

The seed words for the Business Corpus were chosen by the compiler by introspection and brainstorming. A better approach would be to select seed words from the corpus itself. This is achieved by first generating a corpus using one or two highly salient terms, such as *freshman* and *university*. The keyterm output from that corpus can then be used to seed a second corpus. The keyterms from the second corpus could be used to generate a third, and of course the process could be repeated recursively. The reader may have noticed WBC's invitation, illustrated in Figure 5, to "search again with selected keyterms".

Above, we showed the keyterms from our *freshman university* corpus. If the reader glances back at Figure 3, where those keyterms are shown, she will see that there is, against each keyterm, a checkbox. We bootstrapped a new corpus, using as seed words the items that were checked above. Figure 7 shows the keyterms which were extracted from it.

<input checked="" type="checkbox"/> students (342)	<input type="checkbox"/> two (102)	<input type="checkbox"/> process (57)	<input type="checkbox"/> taken (37)
<input checked="" type="checkbox"/> school (203)	<input type="checkbox"/> graduate (89)	<input checked="" type="checkbox"/> subject (49)	<input type="checkbox"/> provide (39)
<input checked="" type="checkbox"/> education (213)	<input checked="" type="checkbox"/> admission (58)	<input type="checkbox"/> skills (37)	<input type="checkbox"/> four (37)
<input checked="" type="checkbox"/> test (225)	<input type="checkbox"/> writing (104)	<input checked="" type="checkbox"/> science (33)	<input type="checkbox"/> reading (55)
<input checked="" type="checkbox"/> student (138)	<input type="checkbox"/> grade (54)	<input type="checkbox"/> including (53)	<input checked="" type="checkbox"/> Admissions (20)
<input checked="" type="checkbox"/> University (231)	<input checked="" type="checkbox"/> TOEFL (139)	<input type="checkbox"/> through (61)	<input type="checkbox"/> order (43)
<input type="checkbox"/> schools (147)	<input type="checkbox"/> required (65)	<input checked="" type="checkbox"/> submit (38)	<input type="checkbox"/> questions (43)
<input checked="" type="checkbox"/> score (150)	<input type="checkbox"/> take (90)	<input type="checkbox"/> subjects (59)	<input type="checkbox"/> examination (25)
<input checked="" type="checkbox"/> scores (148)	<input type="checkbox"/> apply (47)	<input type="checkbox"/> high (57)	<input type="checkbox"/> Center (37)
<input type="checkbox"/> English (142)	<input type="checkbox"/> may (111)	<input type="checkbox"/> exam (42)	<input type="checkbox"/> additional (27)
<input checked="" type="checkbox"/> study (79)	<input type="checkbox"/> educational (48)	<input type="checkbox"/> Art (77)	<input type="checkbox"/> related (49)
<input type="checkbox"/> your (267)	<input type="checkbox"/> curriculum (39)	<input checked="" type="checkbox"/> history (62)	<input type="checkbox"/> primary (29)
<input checked="" type="checkbox"/> program (103)	<input type="checkbox"/> higher (41)	<input type="checkbox"/> International (44)	<input type="checkbox"/> contact (31)
<input type="checkbox"/> level (105)	<input checked="" type="checkbox"/> Arts (58)	<input type="checkbox"/> include (38)	<input type="checkbox"/> colleges (28)
<input checked="" type="checkbox"/> college (79)	<input type="checkbox"/> available (66)	<input type="checkbox"/> complete (37)	<input type="checkbox"/> undergraduate (30)
<input type="checkbox"/> language (91)	<input checked="" type="checkbox"/> essay (66)	<input type="checkbox"/> receive (33)	<input checked="" type="checkbox"/> ACT (56)
<input type="checkbox"/> information (104)	<input type="checkbox"/> California (59)	<input type="checkbox"/> math (23)	<input checked="" type="checkbox"/> SAT (43)
<input checked="" type="checkbox"/> programs (70)	<input type="checkbox"/> each (63)	<input checked="" type="checkbox"/> Studies (52)	<input checked="" type="checkbox"/> applicants (34)
<input type="checkbox"/> based (100)	<input type="checkbox"/> three (57)	<input type="checkbox"/> Office (37)	<input type="checkbox"/> teachers (25)
<input checked="" type="checkbox"/> courses (94)	<input type="checkbox"/> research (60)	<input checked="" type="checkbox"/> secondary (47)	<input type="checkbox"/> used (58)
<input type="checkbox"/> course (112)	<input type="checkbox"/> edit (62)	<input checked="" type="checkbox"/> tests (30)	<input type="checkbox"/> areas (28)
<input type="checkbox"/> year (111)	<input checked="" type="checkbox"/> results (54)	<input type="checkbox"/> following (38)	<input type="checkbox"/> deadline (22)
<input type="checkbox"/> must (107)	<input type="checkbox"/> below (35)	<input type="checkbox"/> Chinese (54)	<input type="checkbox"/> instruction (31)
<input checked="" type="checkbox"/> academic (53)	<input checked="" type="checkbox"/> grades (32)	<input type="checkbox"/> Please (26)	<input type="checkbox"/> specific (28)
<input checked="" type="checkbox"/> application (89)	<input checked="" type="checkbox"/> degree (53)	<input checked="" type="checkbox"/> enrollment (54)	<input type="checkbox"/> eligible (17)

**Figure 7** Recursively bootstrapped *freshman university* corpus

In Figure 7, we have placed a check against the output keyterms which are the same as terms used to seed the corpus, for the reader's convenience (in the actual WBC output screen, such items are highlighted in red). We are encouraged by what new output wordlists of this kind show: it includes a number of new keyterms, such as *educational*, *curriculum* and *undergraduate*, which are salient to the educational domain.

### Multi-word terms

It is possible to expand the WBC corpora by generating multi-word term lists. The *EMW* vocabulary lists currently include only a few phrases, and we should be encouraging our students to learn vocabulary items in the contexts in which they typically occur. WBC can extract multi-word terms of two, three and four words, on the same principles as are employed for the simplex wordlists: the terms must be more frequent in the domain corpus than in the reference corpus. A stoplist of common words is applied, so that terms such as *a student*, no more salient to the domain than the simple *student* are ruled out.

From the corpora illustrated in Figures 3 and 7 (the *freshman university* corpus, and the corpus recursively bootstrapped from it), we generated multi-word term lists. The term list from the bootstrapped corpus is shown in Figure 8.

<input type="checkbox"/> high school (38)	<input type="checkbox"/> board of education (13)
<input type="checkbox"/> test scores (31)	<input type="checkbox"/> Study Abroad (13)
<input type="checkbox"/> State University (28)	<input type="checkbox"/> required to take (13)
<input type="checkbox"/> Open Enrollment (24)	<input type="checkbox"/> University of California (13)
<input type="checkbox"/> art history (24)	<input type="checkbox"/> TOEFL Secrets (12)
<input type="checkbox"/> based test (23)	<input type="checkbox"/> Art Building (12)
<input type="checkbox"/> Mother Tongue (22)	<input type="checkbox"/> Elective Programme (12)
<input type="checkbox"/> graduate school (21)	<input type="checkbox"/> written texts (12)
<input type="checkbox"/> your application (21)	<input type="checkbox"/> scoring system (12)
<input type="checkbox"/> language arts (18)	<input type="checkbox"/> Cultural Revolution (12)
<input type="checkbox"/> School of Education (17)	<input type="checkbox"/> Career Center (12)
<input type="checkbox"/> Ministry of Education (17)	<input type="checkbox"/> Rossier School (11)
<input type="checkbox"/> Writing the Essay (18)	<input type="checkbox"/> score recipients (11)
<input type="checkbox"/> Intercultural Studies (16)	<input type="checkbox"/> ISD 282 (11)
<input type="checkbox"/> General Test (18)	<input type="checkbox"/> local or regional board of education
<input type="checkbox"/> higher education (17)	<input type="checkbox"/> LSAT score (11)
<input type="checkbox"/> North Carolina (17)	<input type="checkbox"/> junior colleges (11)
<input type="checkbox"/> test date (15)	<input type="checkbox"/> your scores (11)
<input type="checkbox"/> financial aid (15)	<input type="checkbox"/> Junior College (11)
<input type="checkbox"/> Vienna School (14)	<input type="checkbox"/> test takers (11)
<input type="checkbox"/> Intensive English (14)	<input type="checkbox"/> English Language (11)
<input type="checkbox"/> School District (14)	<input type="checkbox"/> secondary schools (11)
<input type="checkbox"/> regional board of education (13)	<input type="checkbox"/> college students (11)
<input type="checkbox"/> regional board (13)	<input type="checkbox"/> law school (11)
<input type="checkbox"/> School of Intercultural Studies (13)	
<input type="checkbox"/> TOEFL score (13)	

**Figure 8** Multi-word term list from *freshman university* bootstrapped corpus

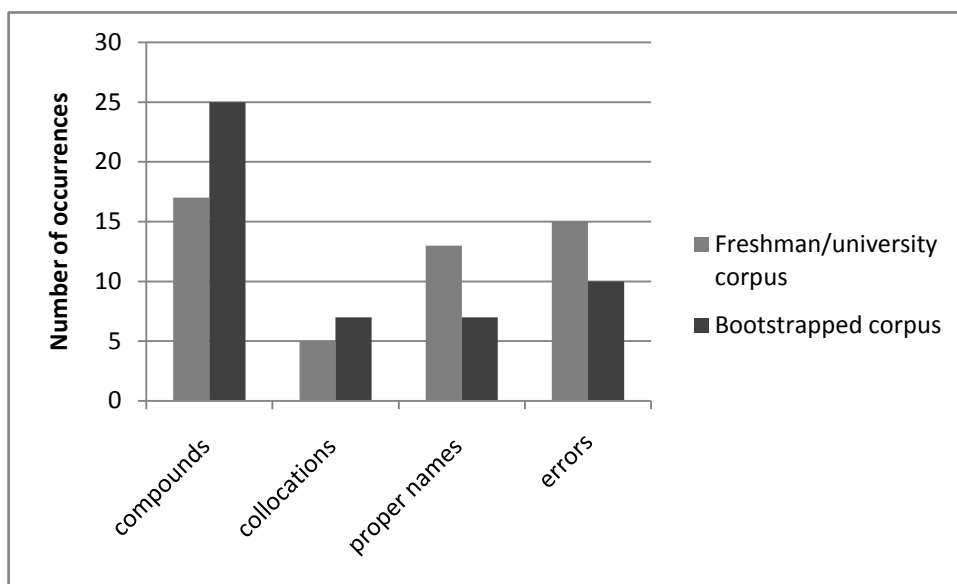
The reader will observe that the majority of the terms have a strong association with the educational domain, and indeed with the process of university application: stronger than that found in the simplex wordlist, in fact. We assigned the terms extracted to four categories:

1. **Compounds:** Lexicalized compound terms in the educational domain, such as *graduate school* and *higher education*
2. **Collocations:** Terms which clearly belong to the domain, and constitute a syntactic phrase group, but would not be found in a dictionary or lexicon, such as *scoring system* and *your scores*

### 3. Proper names

4. **Errors:** anomalous entries which are not syntactic units (such as *based test*) or do not belong to the domain (such as *Cultural Revolution*)

In Figure 9, we show the number of terms (from a total of 50 in each case) that we assigned to each of the four categories. It will be seen that the bootstrapped corpus yielded somewhat more impressive results. This corroborates the findings presented above: that corpus is a richer source of terms in the educational domain than the corpus built using only the two seed words *freshman* and *university*.



**Figure 9** Numbers of multi-word terms generated from two corpora

Whether proper names should be included in student vocabulary lists is a matter for debate; some of the terms extracted, such as University of California, are indeed in the educational domain. What is clear, though, is that the collocational items are just as important to learners as the lexicalized compounds. These collocations are part of the “large store of fixed or semi-fixed prefabricated items” which, according to Lewis (1997) are essential for the acquisition of language.

### **Future work: automatic cloze exercise generation**

At Ming Chuan University, we have found cloze exercises to be a useful learning and assessment tool. We are required to conduct formal English examinations twice per semester, and student numbers are large. Earlier research (Bachman, 1985; Hughes 1981) has indicated that cloze exercises can be used to assess a surprisingly wide range of language skills, including speaking; we lack the resources to examine

all our students orally, but cloze provides a practical substitute.

Currently, cloze exercises are prepared by hand. Not only is this time-consuming, but also the deleted item and distractors are chosen in an arbitrary way. A better solution would be to generate cloze exercises whose distractors are semantically related in some statistically demonstrable way. Ideally, the distractors would have features in common with the correct answer, determined by their similar distribution in a corpus, but would not normally occur in collocation with some other word in the sentence. By way of a simple example, take the cloze exercise “It’s a \_\_\_\_ day”. The correct answer might be sunny, and the distractors tepid, lukewarm and toasty.

Drawing on the resources of WebBootCat and the Sketch Engine, we will present an algorithm for the automatic generation of cloze exercises. The exercises can be used in class, in the lab, or at home, and could be incorporated into an interactive CALL interface, making students’ learning experience more enjoyable and fruitful.

## **Conclusions**

We have shown, in this paper, that it is possible to generate wordlists for vocabulary acquisition that are highly salient to particular topics. These lists are better than existing lists such as those found in the EMW textbooks. The direct learning of vocabulary in language acquisition is here to stay, especially in places such as Taiwan and Cambodia where language tests play an important role in decisions that affect the lives of students. We have shown one way to generate vocabulary to be memorized that is relevant to a lesson topic or has high frequency in texts on that topic.

## **Bibliography**

- Bachman, L. (1985). Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. *TESOL Quarterly*, Vol. 19, No. 3, pp. 535-556.
- Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In Proceedings of EAMT 2006, Oslo, 247-252.
- Biber, D., and Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly* 35.331-6.
- Chaffin, R. (1997). Associations to unfamiliar words: Learning the meanings of new words. *Memory & Cognition*, 25, 203 (24).

- College Entrance Examination Center. (2002). 大學入學考試中心高中英文參考詞彙表 [Daxue ruxue kaoshi zhongxin gaozhong yingwen cankao cihuibiao, High School English Reference Wordlist]. Retrieved March 6, 2008, from [http://www.ceec.edu.tw/Research/paper\\_doc/ce37/ce37.htm](http://www.ceec.edu.tw/Research/paper_doc/ce37/ce37.htm), English abstract from [http://www.ceec.edu.tw/Research/paper\\_doc/ce37/2.pdf](http://www.ceec.edu.tw/Research/paper_doc/ce37/2.pdf)
- Hughes, A. (1981). Conversational Cloze as a Measure of Oral Ability. *ELT Journal* 1981 XXXV(2), pp 161-168
- Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. Paper presented at EURALEX, Lorient, France, July 2004.
- Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern language Journal*, 73, iv, 439-464.
- Lewis, M. (1997). *Implementing the Lexical Approach*. Hove, UK: Language Teaching Publications
- Masamichi Mochizuki. (2003). JACET 8000: A Word List Constructed Using a Scientific Method and its Applications to Language Teaching and Learning. Symposium at ASIALEX 2003, Tokyo.
- Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American educational Research Journal*, 24, 237-270.
- Nation, I.S.P. & Coady, James. (1988). Vocabulary and reading. In: Carter, Ronald & McCarthy, Michael, eds. *Vocabulary and language teaching*. London: Longman, 97-110.
- Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: C.U.P.
- Pitts, M., White, H., & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language*, 5 (2), 271-275.
- Su, Cheng-chao. (2006). Preliminary Study of the 2000 Basic English Word List in Taiwan 23rd ROC-TEFL, Taiwan.
- Uemura, Toshihiko. (2005). JACET 8000 as a Tool of Grading and Evaluating English Texts. *Asialex 2005*, Singapore.
- Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A. & Healy, N. A. (1995).



Growth of a functionally important lexicon. *Journal of Reading Behavior*, 27, 201-212.

Appendix 2: Smith et al (2008d)

## **Automatic cloze generation: getting sentences and distractors from corpora**

**Simon Smith**

Ming Chuan University - Taiwan

**Scott Sommers**

Ming Chuan University - Taiwan

**Adam Kilgarriff**

Lexical Computing Ltd - UK

*Abstract. Cloze exercises are widely used in language teaching, both as a learning resource and an assessment tool. It has been shown that they can cultivate and test a wider range of skills than immediately meets the eye. Cloze has a particularly useful role to play in Taiwan, and other Asian countries, where students of English expect and are expected to memorize a lot of vocabulary. Cloze encourages acquisition of vocabulary through context, rather than the memorization of synonyms or translations. Unfortunately, it is time-consuming and difficult for teachers and materials designers to make up large numbers of cloze exercises.*

*The present paper briefly reviews the literature on cloze in language learning,*

*including systems which generate cloze items automatically, and an algorithm for automatically generating cloze exercises from corpora is presented. It is a bottom-up algorithm, which takes as input from the teacher-user a lexical item which will form the correct answer to the cloze exercise. It outputs a sentence, extracted from a corpus, which contains the lexical item (with the item itself deleted) and a set of distractors is generated. The distractors have a similar semantic distribution to that of the lexical item, but cannot replace it to form a correct answer in the context of the sentence extracted.*

**Keywords:** cloze, FBQ, Sketch Engine, corpus linguistics, ELT

## Introduction

As EFL teachers in Taiwan, we have found cloze exercises (or “fill in the blank” questions, FBQ) to be of great use in our classes, as an instructional as well as an assessment tool. This is especially true, we have found, for very large classes in which many students are reluctant to speak out. Most of the literature (including papers to be mentioned presently) deals with the role of cloze in language proficiency assessment. However, cloze exercises generated by the means we describe in this paper could be used for either purpose, with equal effectiveness.

Cloze is defined by Jonz (1990) as “the practice of measuring language proficiency or language comprehension by requiring examinees to restore words that have been removed from otherwise normal text.” The idea is traditionally attributed to Taylor (1953), when it was used as a test of text readability. The term itself derives from the concept of closure in Gestalt Theory used to describe the human tendency to mentally complete figures even when parts of that figure are missing. Taylor and other cloze

researchers have used the term to describe a sample of naturally occurring text in which words are deleted and respondents asked to use semantic clues in filling in these deleted words. By the 1970s, the concept had been incorporated into educational assessment and subsequently into the assessment of English proficiency among second and foreign language learners (Alderson 1978, Oller 1973).

When constructing cloze tests, EFL researchers use a number of different procedures for text selection and word deletion. Deletion procedures generally follow one of three standardized formats. The historical format established by Taylor calls for the deletion of words at regular intervals regardless of their linguistic properties. A second, similar, approach is random word deletion. A third format uses the linguistic properties of words to determine which words get deleted. In this case, the focus might be syntactic (particular parts of speech, such as prepositions, are candidates for deletion) or, as in the work reported here, it might be on the semantics of deleted items.

### Manual cloze generation

It is difficult for teachers to think up cloze exercises from scratch. Having composed or located a convincing and authentic carrier sentence, which incorporates the desired **key** (the correct answer, or deleted item), it is also necessary to generate **distractors** (wrong answers suggested to the student). This is not a trivial task, as two important constraints apply. On the one hand, the distractors must be incorrect (inserting them in the blank must generate an incorrect sentence). On the other hand, the distractors must in some sense be viable alternatives for completion of the carrier sentence: near synonyms of the key, for example, or words typically found in similar collocational contexts.

A teacher who tries to generate distractors through intuition and introspection may, therefore, encounter the following paradox: if the distractor is too distant from the key, in a semantic distribution sense, it is likely that the student will find the correct answer very easy to deduce; if the distance is too close, sentences incorporating the distractors may turn out to be infelicitously correct.

If a corpus is consulted when manually generating distractors, the teacher may well have access to the necessary distributional information. Nevertheless, the process is time-consuming and tedious, especially if large numbers of items are required, and the advantages of automation are apparent.

### Automatic cloze generation

A growing amount of research has found that cloze can be effectively generated through automated systems. Hoshino and Nakagawa (2007) devised an NLP-based teacher's assistant, which first asks the user to supply a text. The system then suggests deletions that could be made, and helps the teacher to select appropriate distractors. Mostow et al (2004) generated cloze items of varying difficulty from children's stories. The items were presented to children via a voice interface, and the response data was used to assess comprehension. Both of these systems use longer texts, while Sumita et al (2005) describe the automatic generation of single sentence cloze exercises from the World Wide Web. Sumita et al obtain distractors from a thesaurus, and check to make sure that there are zero Google hits for hypothesized sentences in which the **key** (the correct answer) is replaced by distractors.

Our system is similar to that of Sumita et al, in that we select single sentences of

authentic language to build our cloze exercises, and that we look for words with similar lexical distribution to the key to serve as distractors. However, we do not constrain our choice of distractors to synonyms, or even near synonyms; indeed, key and distractor could perfectly well be antonyms, as long as they can occur in the same contexts. Another difference between the two systems is that the Japanese team use a published resource to find distractors, and extract carrier sentences from the web. We use distributional information from a corpus for both of these purposes.

Our system is designed to work in a bottom-up fashion. The teacher/user is first invited to select the correct answer (the key); that is to say, the particular lexical item of which they want to check or reinforce the student's understanding. As far as we are aware, this type of architecture is unique. Other automated systems, by contrast, require the user to select a text, and offer assistance in deciding which word to *delete*. This is significant for two reasons: first, because when we are writing cloze exercises for our students, we often use a vocabulary item as a point of departure.

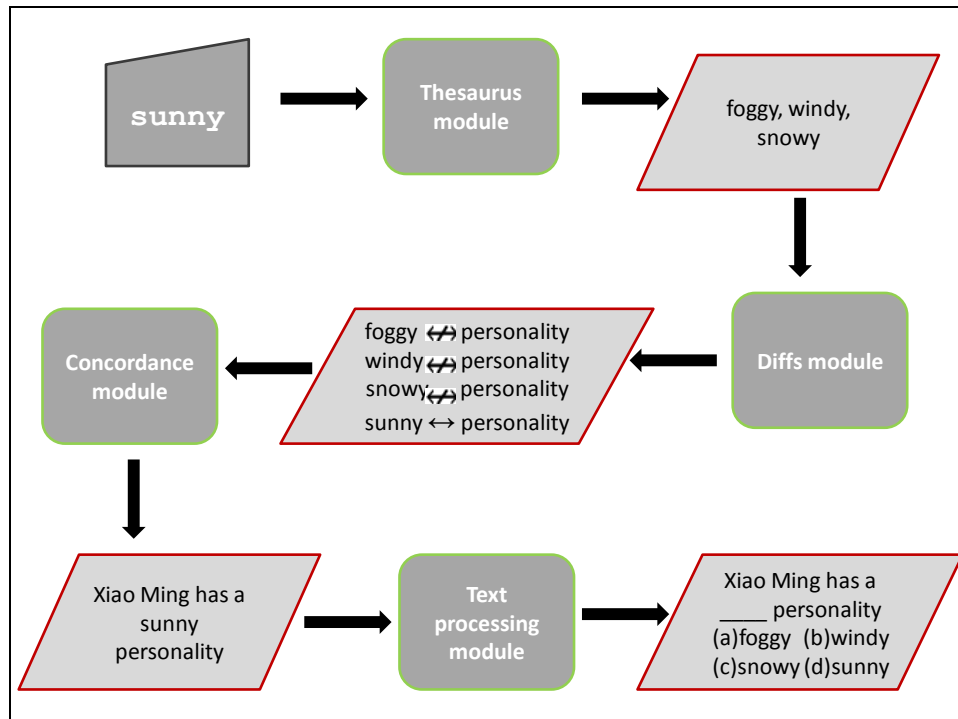
Secondly, our architecture is capable of generating large numbers of cloze items on a given topic ("Business", perhaps, or "Starting out at University"). In Smith, Sommers & Kilgarriff (2008) we reported how to extract corpora, on such topics, from the world-wide web, using WebBootCat (WBC; Baroni et al 2006). The corpora were then used to generate wordlists containing vocabulary salient to the topic. Such wordlists could be readily used as lists of keys to bootstrap collections of on-topic cloze exercises.

## System architecture

The system works like this. First, the teacher specifies the key, or a list of keys to be

processed. Assume, then, that the teacher/user wishes to teach or test the use of the adjective *sunny*, as used to describe personality. She would enter *sunny* into our system as her chosen key. The system will find words which have a similar lexical distribution to that of *sunny*, such as *rainy*, *windy* and so on. It will do this by establishing that these **potential distractors** (PDs) and the key are all found with some set of other words (**key and PD collocates**, KPDCs) such as *weather* and *climate*.

Next, the system looks in the corpus for a word which co-occurs with the key, but never with the PDs. This word is termed the **key only collocate** (KOC). In this example it could conceivably be *personality*, which co-occurs with *sunny* but no other weather adjectives. A sentence that includes the KOC *personality* along with the key *sunny* is then selected from the corpus. All that remains is to delete the key from the sentence, and supply key, distractors and sentence to the student in an appropriate format, as shown in the “Cloze generation system architecture” diagram.



Cloze generation system architecture

Thus, the carrier sentence, the key and the three incorrect answers (distractors) are returned by the system. Subsequently, in the interactive mode, the teacher would be asked if they were satisfied with the item, whether they wanted to generate a new item using the same key, or whether they were happy with the sentence but would like to create a new set of distractors.

Here is an example of a cloze item actually generated by our system.

(1) *They have an enviable \_\_\_\_ of blue-chip clients.*

**Ans: investment infrastructure asset portfolio**

The learner is asked to complete the underscored gap with one of the four answers given. The reader will agree that only the (key) answer *portfolio* is possible, and that if any of the three distractors were inserted, the sentence would become meaningless.

In this work, we make use of the Sketch Engine (SkE) suite of corpus query tools described by Kilgarriff et al (2004), and the ukWaC web corpus to which it provides access.

It needs to be made clear at this point that our system is not computationally implemented. The procedure for deriving the carrier sentences and distractors currently involves the manual implementation of rules which will be automated when we have the necessary time and resources available; we have taken care to set the system up in such a way that it can be readily programmed.

We now describe each step of the algorithm used for generating cloze items in detail.

### *Thesaurus Module*

The Thesaurus module of SkE outputs words which typically occur in the same context as the search term. We show below the SkE Thesaurus output for *portfolio* (the key for the cloze item presented at (1) above). The screenshot reveals that most of the words with similar distribution to *portfolio* are in fact not synonyms or near synonyms: only *collection* and *package* qualify in that regard. A number of the words, as one might expect, have to do with business and the world of investment, with *investment* itself and *asset* ranking high on the list. The presence of the word *curriculum* on the list reflects the fact that the term *portfolio* is now widely used in the education domain.



The three top-ranking list members – *investment*, *infrastructure* and *asset* are noted and retained for use as PDs (potential distractors).



The screenshot shows a Windows Internet Explorer browser window displaying a thesaurus entry for the word "portfolio". The browser's address bar shows the URL "http://beta.sketchengine.co.uk/auth/corpora/run.cgi/thes?corpname=pre". The page has a navigation menu with tabs for "Home", "Concordance", "Word List", "Word Sketch", "Thesaurus", and "Sketch-Diff". The main content area displays the word "portfolio" followed by "ukWaC freq = 66065". Below this, a list of related terms is shown, each with a frequency value. The top three terms are "investment" (0.261), "infrastructure" (0.261), and "asset" (0.249). The browser's taskbar at the bottom shows the Start button, several open applications, and the system tray.

Term	Frequency
<a href="#">investment</a>	0.261
<a href="#">infrastructure</a>	0.261
<a href="#">asset</a>	0.249
<a href="#">assessment</a>	0.248
<a href="#">database</a>	0.247
<a href="#">strategy</a>	0.244
<a href="#">resource</a>	0.242
<a href="#">sector</a>	0.241
<a href="#">management</a>	0.24
<a href="#">initiative</a>	0.238
<a href="#">collection</a>	0.238
<a href="#">provider</a>	0.236
<a href="#">scheme</a>	0.233
<a href="#">planning</a>	0.232
<a href="#">expertise</a>	0.232
<a href="#">package</a>	0.231
<a href="#">curriculum</a>	0.226
<a href="#">business</a>	0.225
<a href="#">marketing</a>	0.224
<a href="#">presentation</a>	0.223
<a href="#">programme</a>	0.223

SkE Thesaurus entry for *portfolio*

### Sketch Differences Module

We next consult the Sketch Differences display. The screenshot below shows sketch differences for *portfolio* and *investment*, in contexts where either can occur in the ukWaC corpus. Notice how the display divides the output into grammatical relations between keyword and collocate. The screenshot shows us that *portfolio* occurs 34 times in a PP\_IN relation with *excess*, while *investment* occurs in this collocation 25 times. Typical contexts are "... an investment/ a portfolio in excess of *n* million dollars".

**portfolio/investment** preloaded/ukwac freq = 66065/213788

**Common patterns**

<b>portfolio</b>	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	<b>investment</b>
------------------	-----	-----	-----	---	------	------	------	-------------------

<b>pp_in</b> 564 26826 0.4 6.0	<b>possessor</b> 1111 1155 4.2 1.3	<b>pp_of</b> 11870 2311 4.2 0.2
excess <a href="#">34</a> <a href="#">25</a> 4.5 3.5	client <a href="#">47</a> <a href="#">42</a> 2.5 2.3	asset <a href="#">107</a> <a href="#">40</a> 4.4 3.1
accordance <a href="#">7</a> <a href="#">30</a> 2.1 3.6	company <a href="#">232</a> <a href="#">138</a> 2.4 1.7	<b>fund</b> <a href="#">47</a> <a href="#">201</a> 2.1 4.2
order <a href="#">45</a> <a href="#">120</a> 0.9 2.3	fund <a href="#">20</a> <a href="#">54</a> 0.9 2.4	resource <a href="#">55</a> <a href="#">129</a> 1.3 2.6

**"portfolio" only patterns**

<b>possessor</b> 1111 4.2	<b>pp_of</b> 11870 4.2	<b>n_modifier</b> 13599 2.2	<b>object_of</b> 11357 1.5
harrah <a href="#">52</a> 9.6	client <a href="#">486</a> 5.8	multi-asset <a href="#">48</a> 6.8	compile <a href="#">135</a> 6.6
kiplinger <a href="#">8</a> 7.8	evidence <a href="#">750</a> 5.8	patent <a href="#">158</a> 6.5	submit <a href="#">309</a> 6.4
ECGD <a href="#">7</a> 6.4	coursework <a href="#">36</a> 5.1	product <a href="#">1344</a> 5.7	assemble <a href="#">48</a> 5.7

<b>a_modifier</b> 13238 1.4	<b>possessed</b> 337 1.3	<b>modifies</b> 17076 1.1	<b>pp_through</b> 66 1.0
diversified <a href="#">211</a> 8.8	inception <a href="#">18</a> 7.3	□ <a href="#">8830</a> 10.6	acquisition <a href="#">8</a> 2.0
enviable <a href="#">111</a> 7.7	content <a href="#">77</a> 2.6	of... <a href="#">138</a> 7.9	
balanced <a href="#">219</a> 7.3	career <a href="#">13</a> 0.8	holder <a href="#">699</a> 7.7	

Part of Sketch Differences entry for *portfolio* and *investment*

Of course, we are interested in situations where the two words do not share a collocate, and for this we glance down at the “portfolio only” patterns. Alongside each collocating word, in the Sketch Differences screenshot, is shown the frequency of the collocation (an underlined integer) and the **saliency** (an index of the number of times *portfolio* occurs with the collocating word, as opposed to other words, given to one decimal place).

We now search for the collocate appearing only with *portfolio* (and never with *investment*) with the highest saliency. We apply the condition that the collocate must be a correctly spelled English word, not a proper name. Thus, the non-alpha character □ with saliency of 10.6 is rejected, as is *harrah*, a proper name (saliency 9.6). The third-ranking in saliency (8.8), *diversified*, is selected, and marked as a potential Key Only Collocate (KOC).

We next consider the second PD, *infrastructure*. The potential KOC *diversified* also does not occur in ukWaC in collocation with this PD, so it remains a candidate. However, when we move on to consider the third PD, *asset*, we find that *diversified assets* does indeed occur in the corpus. This means that *asset* cannot be used as a distractor for the key *portfolio* in the context *diversified portfolio*.

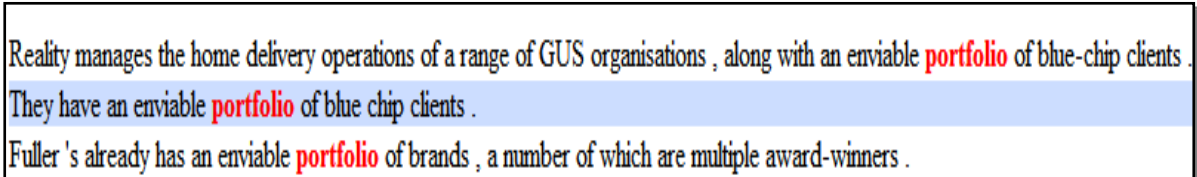
We therefore go on to consider the collocate appearing only with *portfolio* with the fourth highest saliency: this turns out to be *enviable*. This time, we find that the potential KOC does not occur in collocation with any of the PDs, so it is adopted as KOC.

So far, we have decided on the key, as well as the three distractors. We have also established that we wish our carrier sentence to include the collocation *enviable portfolio*. The next step is to determine what the carrier sentence will be: we do this by consulting a concordance.

### *Concordance Module*

The SkE concordancing software is equipped with a feature called GDEX (Husak et al, forthcoming) which favours sentences which are between 10 and 25 words long, containing only common words, and some other related constraints. GDEX sorts the order in which concordance sentences are presented, so that optimal sentences appear first. This means that the sentences which are most likely to be selected for dictionary examples or cloze exercises appear conveniently at the beginning of the concordance display.

From the concordance output from which the screenshot below is taken, we may now extract the sentence shown at (1) above. Note that if the user is dissatisfied with the first sentence, for any reason, they can be prompted to select the second or a subsequent sentence.

A screenshot of a concordance entry for the words 'portfolio' and 'enviable'. The text is displayed in three lines within a rectangular box. The first line is 'Reality manages the home delivery operations of a range of GUS organisations , along with an enviable portfolio of blue-chip clients .'. The second line is 'They have an enviable portfolio of blue chip clients .'. The third line is 'Fuller 's already has an enviable portfolio of brands , a number of which are multiple award-winners .'. The word 'portfolio' is highlighted in red in all three lines. The second line is highlighted with a light blue background.

Reality manages the home delivery operations of a range of GUS organisations , along with an enviable **portfolio** of blue-chip clients .  
They have an enviable **portfolio** of blue chip clients .  
Fuller 's already has an enviable **portfolio** of brands , a number of which are multiple award-winners .

Part of SkE concordance entry for *portfolio* and *enviable*

### *BNC cloze example*

In our experiments, we also generated (2), this time from the British National Corpus. Again, the correct answer choice is supposed to be *portfolio*.]

(2) *Albert E Sharp Fund Managers have launched AES European unit trust, which seeks long-term capital growth from a diversified \_\_\_\_\_ of European Securities.*

**Ans:** asset      portfolio      stock      holding

Unlike ukWaC, the corpus used to generate (1), the BNC does not contain any examples of the adjective *diversified* modifying any of the PDs. However, the concept of a “diversified **holding** of European Securities” does seem quite plausible; given two apparent possible answers, it is unlikely that many teachers would find (2) an acceptable cloze exercise.

The way in which the BNC was compiled means that it consists mostly of clean text, and relatively little noise, while ukWaC contains a fair amount of duplication and non-textual data. This might be taken as a compelling argument for preferring the BNC as a source corpus. However, the GDEX software does a good job of ensuring that the most meaningful sentences from a ukWaC concordance are presented first. What is more, if we posit that certain collocations have a vanishingly small chance of occurring – and that is the claim that one makes when setting the distractors for a cloze exercise – we should be using the very largest corpus available. The larger the corpus, the more exhaustive the evidence; and the less likely the system will be to generate unwanted **correct** distractors, such as *holding* in (2) above.

## Next steps

We have described an algorithm which is capable of generating a carrier sentence and distractors, given a user-supplied key (correct answer). We have shown how modules of the Sketch Engine corpus query tool can be used to generate these components.

As mentioned above, we will shortly prepare an implementation of the algorithm that will allow a user to supply a key at a computer, and be presented with a suggested cloze item. If the item is not satisfactory, the user will be able to run the program again and generate a new exercise.

Beyond straightforward programming, some work will be necessary to ensure that distractors match the key in terms of inflectional morphology (plural –s and the like). A review of any copyright issues involved will also be necessary.

Once implemented, this work can be put to good use immediately. Teachers who use the program will be able to generate authentic cloze items in very short order. As mentioned above, by supplying as input a list of vocabulary items pertinent to the topic of a unit or lesson, such as the “Business” or “Getting started at university” lists described in Smith et al (2008), it will be possible to produce a set of highly relevant cloze exercises. These exercises can be used for assessment, or simply as part of day to day teaching, making students aware of the collocational patterns in which the topic vocabulary commonly participates. The exercises can be used in class, in the lab, or at home, and could be incorporated into an interactive CALL interface, making students’ learning experience more enjoyable and fruitful.

## References

**Alderson, J. C.** 1978. “A study of the cloze procedure with native and non-native

speakers of English.” Doctoral dissertation, University of Edinburgh.

**Baroni, M., Kilgarriff, A., Pomikálek, J. and Rychlý, P.** 2006. “WebBootCaT: instant domain-specific corpora to support human translators.” Paper presented at *EAMT 2006*, Oslo, 247-252.

**Hoshino, A. and Nakagawa, H.** 2007. “Assisting cloze test making with a web application.” Paper presented at the *Society for Information Technology and Teacher Education International Conference 2007* (pp. 2807-2814). Chesapeake, VA: AACE.

**Husak, M., Kilgarriff, A., McAdam, K., Rundell, M. and Rychlý, P.** Forthcoming. “GDEX: Automatically finding good dictionary examples in a corpus.” Paper to be presented at *EURALEX*, Barcelona. July 2008.

**Jonz, J.** 1990. “Another turn in the conversation: What does cloze measure?” *TESOL Quarterly* 24(1): 61-77.

**Kilgarriff, A., Rychlý, P., Smrž, P. and Tugwell, D.** 2004. “The Sketch Engine.” Paper presented at *EURALEX*, Lorient, France. July 2004.

**Mostow, J., Beck, J. E., Bey, J., Cuneo, A., Sison, J., Tobin, B. and Valeri, J.** 2004. “Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions.” *Technology, Instruction, Cognition and Learning* 2: 97-134

**Oller, J. W., Jr.** 1973. “Cloze tests of second language proficiency and what they measure.” *Language Learning* 23: 105-8.

**Smith, S., Sommers, S. and Kilgarriff, A.** 2008. “Learning words right with the Sketch Engine and WebBootCat: Meaningful lexical acquisition from corpora and the web.” Paper presented at the 2008 *CamTESOL conference*, Phnom Penh.

**Sumita, E., Sugaya, F. and Yamamoto, S.** 2005. “Measuring Non-native Speakers’ Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions.” Paper presented at the *2nd Workshop on Building Educational Applications using NLP*, Ann Arbor.

**Taylor, W. L.** 1953. “Cloze procedure: A new tool for measuring readability.” *Journalism Quarterly* 30: 415-433.

## **The authors**

*Simon Smith is an Assistant Professor in the English Language Center of Ming Chuan University, Taiwan, where he teaches EFL and Linguistics, and conducts research on the use of corpora and corpus tools in the teaching of English and Chinese. He has many years of English teaching experience in Taiwan and China private and university centres. He holds a BA in Chinese & Linguistics, from Leeds; an MSc in Machine Translation, from Manchester/UMIST; and a PhD in Statistical Language Modelling, from Birmingham. In his postdoctoral year, at Institute of Linguistics, Academia Sinica, he worked under Professor Huang Chu-Ren on Chinese corpora and Chinese lexical semantics, including the Chinese Wordnet project.*

*Scott Sommers is a Lecturer with the Ming Chuan University English Language Center. He has written and spoken on a wide range of topics including the history of English in Asia, commercial education, and social aspects of blogging. His research interests in ELT include assessment, data-driven learning and needs analysis. His personal blog, Scott Sommers' Taiwan weblog can be found at [www.scottsommers.blogs.com](http://www.scottsommers.blogs.com).*

*Adam Kilgarriff is a research scientist working at the interface between corpus linguistics, language technology, lexicography and language teaching. He has published over fifty articles in this area. He has since worked as computational linguist for Longman Dictionaries (1992-95) and as a research scientist at Brighton University (1995-2003). Since 2003 he has been running his own company, Lexical Computing Ltd., which develops language corpora and provides web services for language research and dictionary publishers. He is a consultant to Oxford University Press and Macmillan Publishers amongst others. He has chaired the Association of Computational Linguistics Special Interest Group on the Lexicon (2000-2003), served on the Board of the European Association for Lexicography (2002-06) and, since 2007, has been the founding Chair of the Association of Computational Linguistics Special Interest Group on Web as Corpus. He has served as invited speaker at conferences in three continents.*





## 出席國際學術會議心得報告

計畫編號	NSC 96-2411-H-004-048
計畫名稱	Sketch Engine 為語言學習的工具
出國人員姓名 服務機關及職稱	史尙明,銘傳大學 or 國立政治大學,助理教授
會議時間地點	<b>Phnom Penh, Cambodia, 23-24 February 2008</b>
會議名稱	<b>4<sup>th</sup> CamTESOL Conference on English Language Teaching</b>
發表論文題目	Learning words right with the Sketch Engine: Meaningful lexical acquisition from corpora and the web

### 參加會議經過

My paper, on automatic techniques for vocabulary building, was well attended (despite the large number of parallel sessions) and received. I decided to make it a participative session, including a task for attendees to work on. Attendees would have preferred an interactive demo of the vocabulary generating toolset, but that was not really possible due to local technical constraints.

I attended a large number of other presentations on both vocabulary and on the use of technology in ELT. A talk by Suksan Suppasetserree & Kiattichai Saitakham revealed that Thai vocabulary learning strategies are similar to those employed in Taiwan, suggesting that there is wider applications for the tools being designed by the PI. Beniko Mason's paper on vocabulary acquisition through story listening emphasized the importance of learning naturally occurring vocabulary.

A paper by George MacLean questioned some learners' willingness to embrace technology in language learning. This may be true of the Cambodian context, but it seems that in Taiwan computers are very much taken for granted.

### 與會心得

This was a very large TESOL conference, in another Asian country. There were many opportunities to meet researchers, teachers and potential collaborators from all over Asia. It was organized by an Australian educational service, and met international standards in almost all respects. A large number of presenters and delegates were from overseas, and the principal keynote speaker was David Nunan, one of the leading of the Task Based approach to language learning (and currently the world's top-selling EFL textbook writer).

## 出席國際學術會議心得報告

計畫編號	NSC 96-2411-H-004-048
計畫名稱	Sketch Engine 為語言學習的工具
出國人員姓名 服務機關及職稱	史尙明,銘傳大學 or 國立政治大學,助理教授
會議時間地點	<b>Lisbon, Portugal, 4-6 July 2008</b>
會議名稱	8th Teaching and Language Corpora Conference
發表論文題目	Automatic cloze generation: getting sentences and distractors from corpora

### 參加會議經過

This was a relatively small conference, one of only two (the other being PALC, which I attended and presented at last year) on the topic of corpus studies in language learning. There were a number of people well-known in the field at my talk. My audience was supportive and enthusiastic, and I was asked by several people to supply further details. One criticism was that the work could have been better motivated pedagogically. However, there are distinct needs that need to be met in Taiwan and other parts of Asia which that critic might not have been aware of.

There were many interesting talks on corpus-based approaches to learning, for example talks on data-driven learning by Alex Boulton of Nancy and by Kiyomi Chujo & Kathryn Oghigian of Tokyo. A number of new computational tools were described, including Michael Barlow's CorpusLab and the new MICASE (EAP corpus from the University of Michigan) online interface.

### 與會心得

I was originally due to give a pre-conference workshop, on learning Chinese using the Sketch Engine, but the workshop had to be cancelled because not enough people signed up for it. This was a pity, not only because it meant that probably the first ever workshop of that type did not take place, but because a lot of preparatory work had gone into it, including some quite substantial modifications to the Chinese corpus, for which a program had to be specially written.

The plan had been to take the Chinese class of verb-object compounds (開車, 吃飯 and the like) to teach participants a little basic Chinese, and show how the Sketch Engine can be used to bring out the importance of collocation in the language. The planned workshop is described at <http://talc8.isla.pt/workshops.html#mandarin>

The conference was a great opportunity to meet delegates with whom I had previously only been in

email contact, including people working with and on the Sketch Engine. One small criticism is that the conference was very much geared to the learning of English through corpora, with less focus on other languages.

## 出席國際學術會議心得報告

計畫編號	NSC 96-2411-H-004-048
計畫名稱	Sketch Engine 為語言學習的工具
出國人員姓名 服務機關及職稱	史尙明,銘傳大學 or 國立政治大學,助理教授
會議時間地點	Barcelona, Spain, July 10th-15th 2008
會議名稱	Tallers de lexicografia: Lexicom 2008
發表論文題目	none

### 參加會議經過

This one-week workshop was an opportunity for me to learn more about the Sketch Engine from the people who designed and programmed it. I was already conversant with the basic functions of SkE, but I will be able to use the knowledge of advanced programming functionality that I acquired to implement the vocabulary list generation and cloze item generation systems that I'm currently working on.

### 與會心得

The workshop also was an opportunity to share with others some of the work on Chinese that I have done with Sketch Engine. In particular, I have a new potential collaborator in China who is interested in the corpora that I have built and use.

## 出席國際學術會議心得報告

計畫編號	NSC 96-2411-H-004-048
計畫名稱	Sketch Engine 為語言學習的工具
出國人員姓名 服務機關及職稱	史尙明,銘傳大學 or 國立政治大學,助理教授

會議時間地點	Barcelona, Spain, July 15 - 19 2008
會議名稱	XIII Euralex International Congress
發表論文題目	none

### 參加會議經過

One key paper that I wanted to hear was that by Adam Kilgarriff, the designer of the Sketch Engine, on the use of a new feature called GDEX. This is designed to present concordance lines, from a corpus, in the likely order of usefulness to a user. This is done by ranking sentences according to length, the kind of lexis they contain (compared to a reference corpus) and several other factors. This will make a very important contribution to the choice of sentences used for context, in my cloze generation project. Other interesting papers were on collaborative dictionary editing by Internet forum, and on generation of word profiles from a large, balanced corpus of German.

### 與會心得

This was a large and elaborate international conference, with many famous names from lexicography and corpus linguistics in attendance, and presentations in a number of different languages. There were many opportunities to learn from speakers, and also to encourage and help younger researchers, by asking relevant questions and giving praise where appropriate.