

EXPLORING PARALLEL CONCORDANCING IN ENGLISH AND CHINESE

Wang Lixun

The Open University of Hong Kong

ABSTRACT

This paper investigates the value of computer technology as a medium for the delivery of parallel texts in English and Chinese for language learning. An English-Chinese parallel corpus was created for use in parallel concordancing -- a technique which has been developed to respond to the desire to study language in its natural contexts of use. Specific problems of dealing with Chinese characters in concordancing are discussed. A computer program called *English-Chinese Parallel Concordancer* was developed for this research. The operation of the program is demonstrated through screen shots. The pedagogical application of parallel concordancing in English and Chinese is illustrated through examples from some teaching and learning experiments, and the Data-Driven Learning approach is applied and explored. It is hoped that parallel concordancing in English and Chinese will become a useful and popular tool for both English and Chinese learners in their second language learning.

INTRODUCTION

Parallel concordancing is a tool which has been developed to respond to the desire (fuelled by linguists such as Sinclair) to study language in its natural contexts of use. It allows us to place side by side for comparison two contexts produced for a given item -- phrase, word, or morpheme -- one being a translation of the other. It has many uses in translation studies and in translation pedagogy, such as in the compilation of bilingual dictionaries. However, in the present paper it is the pedagogical value of parallel concordancing which will receive attention.

The main research interest in this paper is in the use of parallel concordancing in the teaching of languages, specifically in its use as a form of consciousness-raising, of making learners aware of the differences between the target language and their own language (Rutherford, 1987). By comparing the contexts obtained for an item in one language, with the translations of the contexts in the other language, learners can see how the item is rendered according to varying contextual elements (Roussel, 1991). This can be useful pedagogically as, for example, it can help to prevent the L2 of more advanced learners from becoming fossilised and settling into the use of cognate but contextually inappropriate structures in the target language. It can help one to look at the way a given structure is used in different styles or registers, or by different age groups, or by native and foreign speakers (King, 1989).

Barlow, who developed the *ParaConc* (Barlow, 2001) program for parallel concordancing, claims that parallel texts (texts that are translations of each other) are a promising resource for a range of research projects related to language learning. Using parallel texts, as he puts it, "allows language learners to directly investigate (perhaps in response to queries posed by the teacher) the main correspondences between particular words and structures in two languages" (Barlow, 1996a). It helps beginning learners to create an awareness for the feel of a second language and also to obtain some concrete knowledge of correspondences. It also helps advanced learners to deepen their knowledge of words and phrases: to understand not just the main meaning or most common meanings of a word, but to understand a range of meanings and to perceive how context in terms of discourse and genre provides clues to the appropriate

meaning (Barlow, 1996a, 1996b). In this paper, some pedagogic applications of parallel concordancing are explored, making use of Barlow's insights and also the Data-driven Learning (DDL) approach (Johns, 1991, 1993, 1994), which will be discussed in the section "[Parallel Concordancing for Lexical Learning](#)."

To carry out parallel concordancing in English and Chinese, I constructed an English-Chinese parallel corpus and developed a software package, *English-Chinese Parallel Concordancer* (Wang, 2000). A concordance example of the word *xian4zai4* (now) is discussed in the paper, revealing an insight into different uses of the word, and how the findings can be applied in language learning. (*Xian4zai4* is Chinese Pinyin, the Roman transliteration of Chinese characters, which is used throughout this paper for the convenience of English readers. The numbers are tone markers.)

PROBLEMS OF DEALING WITH CHINESE CHARACTERS IN CONCORDANCING

Although parallel concordancing has been carried out between several European languages, it seems not to have been previously extended to non-alphabetic languages such as Chinese. This is due to fundamental differences in the language systems which create complex conceptual and computational problems of alignment. The most immediate differences between Chinese and the European languages are that Chinese is written in ideograms rather than alphabetic characters, and that it lacks the properties of most European grammatical systems. For example, it has no articles, no tenses, no participles or gerunds, no moods, and virtually no inflections. It even had no punctuation, until it was introduced from the West at the end of the 19th century.

Even in a language as English, the definition of "word" can be problematic. For example, is "crabmeat" one word, or two? However, it is even more problematic for a language as Chinese to define word. Written Chinese gives no indication of which characters are to be considered as words and which combine with others to form compound words. For example, according to standard Chinese grammar rules, *ban4* (half), *tu2* (way), *er2* (but), and *fei4* (give up) are four words, which should be separated by spaces. But most Chinese people consider this four-character combination a single word (give-up-half-way). This type of combination is very common in Chinese, having a similar function to that of an idiom in English, although the characters in it normally keep their original meanings rather than combine with others to form compound words. Also, unlike English idioms, the compound can function as an adjective, adverb, or verb, which might explain why people usually regard it as a single word.

If we want to take account of the non-correspondence between character and word, we must first develop some way of establishing when a string of characters can be considered a word. Then, in entering the Chinese text on computer, spaces can be inserted between these conceptual words to correspond to the standard graphical indication of a word in English. Thus *wo3* (I), *qi2* (ride) *zi4 xing2 che1* (bicycle) would be entered as *wo3 qi2 zi4xing2che1*.

However, there are a number of technical problems associated with this form of alignment. It seems impractical to design a computer program to insert spaces automatically, since two successive characters may be either one or two words according to the context. This means that the spaces have to be added manually, which is costly in terms of time and money. Furthermore, the end-user searching for a word with the retrieval software may conceive of words differently from the original corpus compiler and may have to make several attempts to match the compiler's input.

Given the technical and conceptual problems associated with non-correspondence alignment, it appeared that the only practical solution was to make an assumption of character-word correspondence and thus treat each Chinese character as a word. Having made this assumption, the inputting task was made easier by the Chinese word processor NJStar, which not only inserts spaces between Chinese characters automatically, but can also convert Chinese characters into Pinyin, which is very important for English-speaking people wanting to learn or pronounce Chinese.

CREATING AN ENGLISH-CHINESE PARALLEL CORPUS

Unlike other concordancing programs such as *Microconcord* (Johns, 1986) or *Wordsmith* (Scott, 2000), which can be used on any collection of texts, a parallel concordancer must be used on a corpus consisting of parallel texts in two or more languages. Before developing the concordancing program, then, it was necessary to select texts in order to set up an English-Chinese parallel corpus.

The corpus aims at helping intermediate English or Chinese language learners, such as university students, further improve their second language. Thus, the texts chosen were English or Chinese texts which are fairly easy to understand from the point of view of vocabulary, syntax, and discourse. University students are usually interested in genres such as novels, fables, essays, autobiographies, magazines, and general scientific articles, so these genres were taken into first consideration. To keep a balance, about half the source texts were in English and half in Chinese. Only written materials were collected, as it was too difficult for the present research to cover transcribed spoken materials. To ensure that the quality of translation was good, only published translations were selected. The corpus now contains about 1 million words in English and 2 million characters in Chinese. Table 1 shows the percentage of genres distribution in the corpus.

Table 1. Percentage of Genres Distribution in the Corpus

| Genre | novel | essay | fable | autobiography | scientific article | political address | magazine | other |
|-------|-------|-------|-------|---------------|--------------------|-------------------|----------|-------|
| % | 50 | 15 | 10 | 5 | 5 | 5 | 5 | 5 |

Initially, the method of inputting the texts was to scan in English texts and type in Chinese texts. Subsequently, Chinese texts were scanned with SunmiPage ScanInsert OCR software (Liang, 1997) and then edited. The texts used are either copyright-free or permission has been obtained from the authors.

After editing, the texts needed to be marked up. The purpose of marking up texts is to define sentence and paragraph boundaries so that a sentence in one text can be matched with its translation in the other by the parallel concordancing program. In order to keep the size of the text files as small as possible, minimal marking up was used: The only necessary element is <S> to identify sentence boundaries, as the program was developed in such a way as to recognise paragraph boundaries without special markers.

Electronically, each Chinese punctuation mark occupies two bytes, while each English mark occupies only one byte. A program was developed to automatically mark up Chinese text according to Chinese punctuation and English text according to English punctuation.

THE DEVELOPMENT OF THE ENGLISH-CHINESE PARALLEL CONCORDANCER

Since 1997, I have been developing the *English-Chinese Parallel Concordancer (E-C Concord)*, and the first version was successfully completed in 2000. It works in a Windows95/98 environment, and can carry out sentence-by-sentence parallel concordancing in English, Chinese, and Pinyin. The main technical problem in developing a program for parallel concordancing related to the alignment method used for identifying equivalent sentences between texts. A major problem in aligning texts arises when the number of sentences in the source language differs from that in the target language. The situation could also arise where the number of sentences in a paragraph is the same, but the divisions between them do not coincide. A program called *Multiconcord* (Woolls, 1997) had previously been developed at the University of Birmingham, using an algorithm which automatically looks for disturbance between the two texts and re-establishes the matches by joining several short sentences together in one language to match a long one in the other. The algorithm gives satisfactory accuracy in aligning parallel texts in European languages (Woolls, 1998). However, an adaptation of this program to align texts in English and Chinese only achieved an accuracy of about 60%, based on an accuracy test carried out by Woolls and the author. The decision was then taken that for the present research the texts would be pre-aligned -- which of course

gives an accuracy of 100%. That accuracy is achieved at the cost of time-consuming manual pre-editing of the texts.

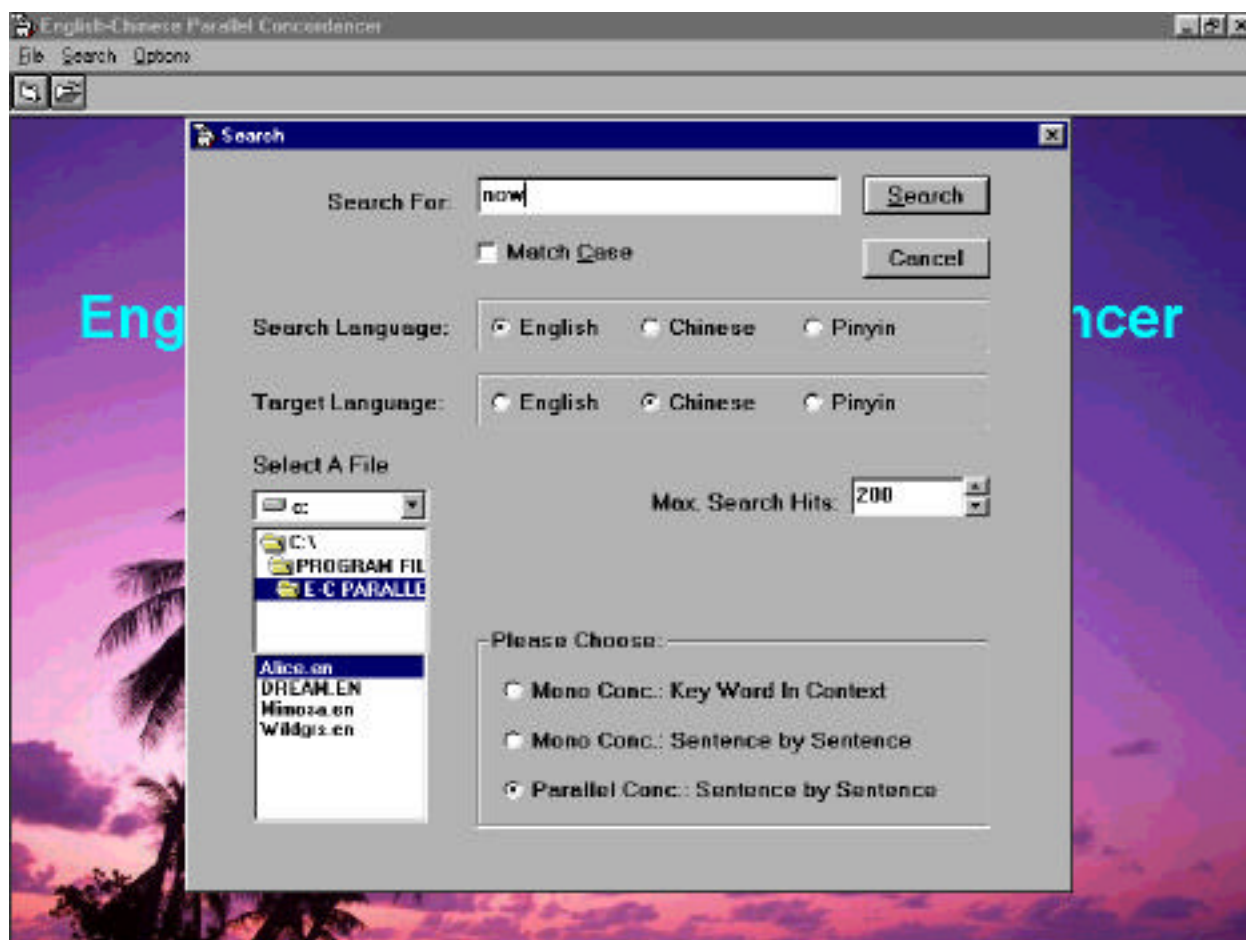


Figure 1. Screen shot of the search window of *E-C Concord*

The program allows the user to type in a search item in the "search box," and choose a Search Language and a Target Language. When entering an English or Pinyin search item, wild cards (*) are acceptable, so that "book*" can be "book," "books," "booking," "booked," and so forth, and "wang*" can be "wang1," "wang2," "wang3," or "wang4." Wild cards cannot be used with Chinese characters. The user needs to select one or more text files from the file list: These files contain the corpus data. The program provides three ways of concordancing: (a) Monolingual Concordance, Key-Word-In-Context; (b) Monolingual Concordance, Sentence-by-Sentence; and (c) Parallel Concordance, Sentence-by-Sentence. The user can also control the maximum search hits. After making all the necessary choices and pressing the "Search" button, the user will get a result such as shown in Figures 2 and 3.

The concordance output is in sentence-by-sentence format, which consists of pairs of English and Chinese sentences, one been the translation of the other in the pair. The text can be edited on screen and saved as text files for further studies.

PARALLEL CONCORDANCING FOR LEXICAL LEARNING

More than one and a half centuries ago, von Humboldt (1836/1988) pointed out that "we cannot, properly speaking, teach a foreign language: all we can do is create the conditions under which it can be awakened in the soul" (p. 236). Using Humboldt's insights, and based on the data generated by the concordancer, Johns (1991) proposed a new language-learning approach, which he called Data-Driven Learning (DDL). The DDL approach puts emphasis on the inductive acquisition on the part of students of grammatical rules or regularities through the process of analysing the patterns of language use of specially selected items as revealed through corpora (Johns, 1991; Tribble & Jones, 1990). Johns's remark "Every student a Sherlock Holmes" implies that the role of the learner has changed in DDL: A learner is a researcher, testing hypotheses and revising them in the light of data; a learner is a detective, finding and interpreting linguistic clues. DDL can focus on different aspects of language. This paper focuses on lexical learning using DDL. The following is an example of what a learner can detect by analysing parallel concordance data.

The lexical item studied here is the adverb *xian4zai4* (now), as it is a very common and important word, but one not satisfactorily covered by bilingual dictionaries. Some differences in the use of *xian4zai4* and "now" in the two languages are discussed below.

One hundred and twenty-eight examples were found in four different texts (novels). Forty examples were randomly selected from them, and were classified into several groups. The idea was to ask Chinese students at an intermediate English level to identify the linguistic bases of the grouping. In order to compare Chinese characters with English words more clearly, the Pinyin transcription identifies its separate "words."

The following abbreviations, as used by Li & Thompson (1981), were used in the examples:

| Abbreviation | T | O | CRS | PFV | ASSOC | GEN | CL | 3sg |
|--------------|-------------|----------|--|---------------------------------|---------------------------|------------------------|------------|-----------------------|
| Term | translation | original | currently relevant state (<i>le</i>) | perfective aspect (<i>le</i>) | associative (<i>de</i>) | genitive (<i>de</i>) | classifier | third person singular |

Some of the above abbreviations were used because certain Chinese characters, such as those for *de* and *le*, cannot be translated directly into English words. Furthermore, each of these two has two distinct meanings which depend on the context. Many Chinese classifiers cannot be translated into English, as they simply do not exist in English, where, for example, one speaks of "a herd of cows," but there is no classifier for a single cow. The third person singular pronoun *ta1* in Pinyin does not show the gender, so it cannot be automatically translated into "he" or "she."

Eight Chinese students in the University of Birmingham were asked to accomplish the following tasks concerning the adverb *xian4zai4* (now).

Task 1

Look at the following data:

1. T: *di2que4 shi4 zhe4 yang4: ta1 xian4zai4 zhi1 you3 shi2 ying1cun4 gao1 le5,*
truly be like this **3sg now** only have ten inch high CRS
O: And so it was indeed: **she** was **now** only ten inches high, ...
2. T: *shi4shi2shang4, ta1 xian4zai4 yi3 yuan3 bu4zhi3 jiu3 ying1chi3 gao1, ...*
in fact **3sg now** already much not less than nine feet high
O: in fact **she** was **now** rather more than nine feet high, ...
3. T: *ta1 wan2quan2 wang4ji4 le5 ta1 xian4zai4 bi3 tu4zi3 da4 shang4 yi1qian1 bei4,*
3sg completely forget PFV **3sg now** compare rabbit big up a thousand times
O: ...quite forgetting that **she** was **now** about a thousand times as large as the Rabbit, ...
4. O: *wo3 xian4zai4 yi3jing1 cheng2 le5 ming2 fu4 qi2 shi2 de5 gong1ren2 ...*
I now already become PFV name agree that fact GEN worker
T: I was **now** a bona fide worker ...

Question: What underlying pattern can be detected in the above parallel texts?

What the students found was that, in the Chinese examples, *xian4zai4* immediately follows the subject, while in the English ones, *now* follows "subject + be." They were then asked whether this was always the case. They carried out more concordancing and found that there was no such structure as "subject + verb (be) + *xian4zai4*" in Chinese in the corpus. The conclusion they drew was that Chinese speakers should pay special attention to the structure "subject + verb (be) + *now*" in English, as this structure does not exist in Chinese. They also suggested that English speakers learning Chinese should avoid adding an unwanted verb (be) to a Chinese sentence.

Task 2

5. T: *"xian4zai4 gai1 dao4 hua1yuan2 li3qu4 la5!"*
now should go to garden into!
O: "And **now** for the garden!"
6. T: *"kuai4dian3, xian4zai4 jiu4 qu4!"*
quick **now** immediately go
O: "Quick, **now**!"
7. T: *ba3 ta1de5 tou2 tai2 gao1 -- xian4zai4 na2 bai2lan2di4 lai2 --*
make his head raise high **now** fetch brandy come
O: Hold up his head -- Brandy **now** --

Question: Why are the English versions of the above sentences so much shorter than the Chinese ones?

The students found that in the English sentences various subjects and verbs around *now* were not present. For example, "And now (I should head) for the garden," "Quick, now (you go there immediately)," and "(You go and fetch some) Brandy now." In the Chinese translation, however, the words struck through were presented, such as "should go to ... into" in Example 5, "immediately go" in Example 6 and "fetch ... come" in Example 7. The students concluded that in Chinese the adverb *xian4zai4* could not be used independently, and some words not present in the English sentences were required in the Chinese translation. They realised that certain structures which are acceptable in English are not acceptable in Chinese, and vice versa. It seems that in the above Chinese sentences, 'the law of least effort' was not followed.

Task 3

8. O: *wo3 xian4zai4 bu4 chi1 zhi1shi4 wo3 bu4 xiang3 chi1 ta1 ba4 le5.*
 I **now** not eat only I not want eat 3sg CRS
 T: But I didn't choose to just yet.
9. O: *wo3 xian4 zai4 shi4 "zu3 zhang3" le5, geng4 zhu3yao4 de shi4*
 I **now** be group leader PFV even mainly GEN be
 T: Because I was "group leader" and, even more, ...
10. O: *wo3 xiang3 ta1 bu4 shi4 sui2 kou3 zhe4yang4 shuo1 de, ke3neng2 shi4 you3yi4shi4di4*
 I think 3sg not be casually in this way speak ASSOC may be intentionally
yao4 rang4 wo3 zhi1dao4 wo3 xian4zai4 bu4 tong2 yu2 guo4qu4 de shen1fen1.
 want let I know I **now** not same past ASSOC status
 T: I suspected that he said this to let me know my changed status.
11. O: *na3me5, wo3 xian4 zai4 sheng1huo2 yu2 qi2jian1 de zhe4ge4 xin1 de sheng1cun2*
 then I **now** live in between ASSOC this new ASSOC living
huan2jing4 shi4 zen3yang4 de5 ne5?
 surroundings be what GEN?
 T: So what about my life in these new surroundings?

Question: What is missing from the English versions of the above sentences? Why?

The students easily found that *xian4zai4* occurs in the Chinese text but *now* did not appear in the English translation.

The students observed that the English translation in [Example 8](#) simplified the original Chinese sentence. There were two sentence structures parallel to each other in the Chinese sentence, the first stating the fact that "I now (do) not eat," the second telling the reason "I (do) not want (to) eat." Having further studied the extended context of the sentence in the original text, the students realised that the narrator of the sentence was in a state of starvation most of the time, so to be able to choose whether to eat or not was very satisfying, and the feeling was expressed through the parallel sentence structure. The English translation used prospective contrast, and it simplified the sentence. The students felt that it was not as expressive as the original Chinese sentence.

In [Example 9](#), the students argued that *now* was not used in the English translation because the past tense "was" was clear enough and *now* was not necessary. In the Chinese version, the combination "*now* ... *le* (PFV)" served the same purpose as "was."

In [Example 10](#), the students found that the Chinese version used contrastive structures twice: "casually in this way speak" versus "intentionally want let I know" and "now" versus "past," but neither appeared in the English translation. They argued that contrastive structures were frequently used in Chinese to make the meaning of sentences absolutely clear, but in English quite often such structures were not used so as to make sentences simpler.

In [Example 11](#), the students found it logically reasonable that the word *now* did not appear in the English translation: One could not live in the past in "new surroundings." Although it sounded redundant, the word *xian4zai4* should not be omitted from the Chinese sentence.

Having studied examples where *xian4zai4* occurred in the Chinese original but *now* did not appear in the English translation, the students carried out more parallel concordancing looking for examples where *now* occurred in an English original but *xian4zai4* did not appear in the Chinese translation. The following are some examples they found:

12. O: "**Now**, Dinah, tell me the truth: did you ever eat bat?"
 T: "**wei4**, dai4na4, gen1 wo3 shuo1 shi2 hua4, ni3 chi1 guo4 bian1fu2 mei2you3?"
wei (draw attention) Dinah to me say real words you eat PFV bat not
13. O: ...her face brightened up to think that she was **now** the right size for going through the little door into that lovely garden.
 T: xiang3 dao4 ta1 **mu4qian2** de shen1cai2 zheng4hao3 neng2 tong1 guo4 na3 shan4 xiao3
 think 3sg **in front of eyes** ASSOC size right can go through that CL little
 men2, ke3yi3 jin4ru4 na3 ke3ai4 de hua1yuan2, ta1 xi3 xing2 yu2 se4.
 door can enter that love -ly garden 3sg joy reflect through (face) colour
14. O: She found that she was **now** about two feet high, ...
 T: ta1 fa1xian4 **ci3ke4** zi4ji3 shen1 gao1 da4yue1 liang3 ying1 chi3...
 3sg find **this moment** self body height about two feet
15. O: "**Now** tell me, Pat, what's that in the window?"
 T: "**hao3le**, gao4 su4 wo3, pa4 te4, chuang1zi3 li3 na3 dong1xi1 shi4 shen2me?"
all right tell me Pat window in that thing be what

Having studied the examples, the students realised that *xian4zai4* is not the only translation of *now*, it can be translated as *mu4qian2* ("in front of eyes"), *ci3ke4* ("this moment"), and possibly other words, and sometimes now is used as a word for drawing attention rather than for referring to time: *wei4* ("well" or "listen") and *hao3le* ("all right"). Discoveries like this certainly help learners to be more aware of different uses of words in different contexts. Their L2 is less likely to become fossilised, and they will be able to see more of the subtle differences between meanings, and will try to avoid using cognate but contextually inappropriate structures in the target language.

The above discussion shows the possibility of using parallel concordance data as teaching materials for Data-driven Learning purposes. The teacher can either put data into groups for students to study, or ask them to carry out concordancing on a particular lexical item, analyse the data, and ask them to submit what they have found through the analysis.

CONCLUSION

Technically, parallel concordancing between English and Chinese has been established successfully, and further tasks can be developed and experimented with students at different level to increase their, and their teachers', familiarity with the methodology. It is highly possible that the *English-Chinese Concordancer* (Wang, 2000) can be extended to Japanese and Korean, as like Chinese, they use ideograms rather than alphabetic letters. Experience suggests that the parallel concordancer is one of the most powerful tools that computer science can offer to language researchers. The distinctive feature of the Data-driven Learning approach to inductive language teaching is that the language data are primary, and the teacher does not know in advance exactly what rules or patterns the learner will discover. DDL with the support of parallel concordancing will help the learner to develop in-depth knowledge of lexical meaning and use based on evidence from authentic language.

ABOUT THE AUTHOR

Wang Lixun was born in China. He was awarded a PhD in Computational Linguistics at the University of Birmingham, UK, in 2000. His research interests include computer-assisted language learning; corpus linguistics; Web-based language learning. He has developed the software *English-Chinese Parallel Concordancer*, *Bilingual Sentence Shuffler*, and *MatchUp*. He has also developed his [homepage](#) and the [ECLEPT](#) Web site. He currently works in the School of Arts and Social Sciences at The Open University of Hong Kong.

E-mail: lxwang@ouhk.edu.hk

REFERENCES

- Barlow, M. (1996a). Parallel texts in language teaching. In S. Botley, J. Glass, A. M. McEnery, & A. Wilson (Eds.), *Proceedings of teaching and language corpora 1996* (UCREL Technical Papers Volume 9; pp. 45-56). Lancaster, UK: University Centre for Computer Corpus Research on Language.
- Barlow, M. (1996b). Corpora for theory and practice. *International Journal of Corpus Linguistics*, 1(1), 1-37.
- Barlow, M. (2001). ParaConc [Computer software]. Houston, TX: [Athelstan](#).
- Humboldt, W. von. (1836/1988). *On language: The diversity of human language-structure and its influence on the mental development of mankind* (P. Heath, Trans.). Originally published as the introduction to *Über die Kavi-Sprache auf der Insel Java (1836-1840)*. Cambridge, UK: Cambridge University Press.
- Johns, T. F. (1986). Microconcord: A language-learner's research tool. *System*, 14(2), 151-162.
- Johns, T. F. (1991). Should you be persuaded -- two samples of data-driven learning materials. In T. F. Johns & P. King (Eds.), *Classroom concordancing* (English Language Research Journal 4; pp. 1-13). Birmingham, UK: Birmingham University.
- Johns, T. F. (1993) Data-driven learning: An update. *TELL & CALL*, 1993(2), 4-10.
- Johns, T. F. (1994) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Approaches to pedagogic grammar* (pp. 293-313). Cambridge, UK: Cambridge University Press.
- King, P. (1989) The uncommon core: some discourse features of student writing. *System*, 17(1), 13-20.
- Li, C., & Thompson, S. (1981). *Mandarin Chinese*. Berkeley, CA: University of California Press.
- Liang, X. M. (1997). [SummiPage ScanInsert OCR](#) [Computer software]. Singapore: Computek Enterprises Pte Ltd.
- Roussel, F. (1991). Parallel concordances and tonic auxiliaries. In T.F. Johns & P. King (Eds.), *Classroom concordancing* (English Language Research Journal 4; pp. 71-103). Birmingham, UK: Birmingham University.
- Rutherford, W. E. (1987). *Second language grammar: Learning and teaching*. London: Longman.
- Scott, M. (2000). [WordSmith Tools Version 3.0](#) [Computer software]. Oxford, UK: Oxford University Press.
- Tribble, C., & Jones, G. (1990). *Concordances in the classroom: A resource book for teachers*. London: Longman.

Wang, L. X. (2000). *English-Chinese Parallel Concordancer* [Computer software]. Birmingham, UK: University of Birmingham.

Woolfs, D. (1998, July 24-27). Multilingual Parallel Concordancing for Pedagogical Use. *Teaching and Language Corpora 98* (pp 222-227). Oxford, UK: Keble College.

Woolfs, D. (1997). *Multiconcord* [Computer software]. Birmingham, UK: CFL Software Development.