A Japanese-English Parallel Corpus and CALL:
A Powerful Tool for Vocabulary Learning

Kiyomi CHUJO[1], Masao UTIYAMA[2], Chikako NISHIGAKI[3]
[1] *Nihon University*, Japan
[2] *National Institute of Information and Communications Technology*, Japan
[3] *Chiba University*, Japan

1   Introduction

Vocabulary building is an essential component to language learning, and new developments in technology provide new tools for learners. Using high-speed computers, it is now possible to use corpora in the classroom as a means to develop vocabulary and understand language structures in real language contexts. Although recognized by educators as a potentially useful tool, until now corpus application has been limited because the English concordance examples retrieved have been difficult for beginning level learners to understand (Tono, 2003; Umesaki, 2004). The recent development of English-Japanese parallel corpus programs have exciting potential for not only eliminating this barrier, but also in facilitating inductive corpus-based language learning, i.e., data-driven learning (DDL), in ways that have not been possible in the past.

To take advantage of the potential strengths in DDL, our team of associated researchers developed and combined a newspaper-based Japanese-English parallel corpus with a TOEIC-based CALL program to produce a set of corpus-based vocabulary learning activities for beginning level EFL learners. This paper will present a description of this CALL-Corpus program and will provide an overview of program development, and a case study including classroom implementation, student responses to the program, learning outcomes from the concordance-based activities in various learning contexts, and the optimum interface features between the corpus and the user.

2   Program Development and Implementation

2.1   Corpus and concordance program

The Japanese-English News Article Alignment Data, developed by Utiyama and Isahara (2003), is comprised of 260,000 translation pairs from the Japanese language *Yomiuri Shimbun* (having 6.1 million Japanese morphemes) and the English language *Daily Yomiuri* (having 4.9 million English words), and these pairs are automatically aligned.

We chose the multilingual concordancer ParaConc, developed by Barlow in 2002, to investigate the equivalences and contrasts between two languages. ParaConc is particularly useful since it permits a wide range of investigations of aligned texts between languages.

2.2   Target vocabulary for DDL

Since one of the goals of many CALL courses is to improve learners' TOEIC scores, we designed sets of corpus-based activities to teach learners how to use a corpus to identify recurring features of various words frequently appearing in TOEIC, and to show them how to induce generalizations from the samples. Using 640 words from our pre-existing CALL vocabulary teaching material 'TOEIC Vocabulary 1, 2, and 3' (Chujo, et al., 2003, 2004), we created 32 DDL lessons. Thirteen of these lessons were implemented in DDL classroom applications as a part of this study.[1]

2.3 Tasks

Since the targeted learners' English proficiency was at the beginning level, the majority of the tasks were fairly easy. These tasks were classified into six types, listed from the easiest (1) to the most difficult (6), and are described as:

(1) examining Japanese equivalents of an English word
   e.g. the English word *decline* and multiple Japanese equivalents *geraku*, *genshou*, *suitai*, *teika*, and *ochikomi*

(2) examining English equivalents of a Japanese word
   e.g. the Japanese word *joshi* and multiple English equivalents *boss*, *supervisor*, *superior*, and *subordinate*

(3) collecting frequently appearing citations of an English word
   e.g. the English concordance of *measure* and its expressions such as *assistance measures* (*shien-taisaku*), *emergency measures* (*oukyuu-taisaku*), *economic measures* (*keizai-seisaku*), and *economic stimulus measures* (*keiki-seisaku*)

(4) translating Japanese phrases into English
   e.g. In order to translate *hyakuman-yen iri-no fuutou* into English, the Japanese *fuutou* concordance shows the English equivalent *envelope* and its various recurring multi-word expressions, including "*an envelope containing … yen*".

(5) examining collocation patterns
   e.g. *accordance* appears in a collocation pattern *in accordance with* and the Japanese equivalents *ni-shitagatte* or *ni-motoduite*

(6) observing recurrences and inducing generalizations
   e.g. the occurrence of 'a rule or system' with *in accordance with*

2.4 Participants

The case study was conducted in three Japanese university beginning-level English CALL courses which met for one 90-minute class per week for four or five weeks. There were 24 freshmen participants in Class 1, and 28 freshmen in Class 2. Class 3 had 20 graduate students. The students' course objective was to use the CALL listening material and vocabulary learning material to improve their communicative proficiency.

2.5 Classroom procedures

During a 90-minute class, the DDL activities were integrated with the CALL individualized activities and were performed by pairs of students. In the first 15 minutes, a review test in both listening and vocabulary from the previous lesson was given, followed by a CD-ROM listening activity for 30 minutes.

Next, the easy-to-gradually-more-difficult DDL tasks described in 2.3 were done for 25 minutes. Learners worked in pairs with a self-explanatory DDL handout consisting of about ten tasks for exploring the usage of seven target words. Some target words were provided with several tasks. In pairs, one student performed the concordancing and the other wrote down the results on the handout. They consulted with each other about how to perform the concordance and how to interpret the patterns and tendencies in the data they found. Each pair was required to submit the handout after filling out the answers to each task question. They also were asked to jot down their daily reactions to using the DDL activity in order to assist us in evaluating the transitions in learner interest and ability. These evaluations are referred to as the "Daily Evaluations".

As soon as they finished the DDL activities, they had 15 minutes to use the CALL vocabulary CD-ROM to study the 20 target words which included the seven DDL targeted words. At the end of the lesson, a CALL CD-ROM quiz was given to confirm their understanding of the vocabulary learned that day.

2.6 Evaluation of the DDL activities

After completing all thirteen DDL lessons, students were asked to fill out a questionnaire ("Final Questionnaire") using a five-point rating scale and several open-ended comment questions.

3 Results and Discussion

3.1 Was the concordancing activity useful to learners?

Sixty-two percent of students agreed that the DDL activity was useful for vocabulary learning, versus seventeen percent of students who did not. On the whole, learners responded with a positive attitude to the experience. This is reflected by the fact that only three students of sixty-three responded negatively.[2]

This research question was further addressed by the learners' responses to the open-ended question "How would you like to use this tool?" Eighty-eight percent of students answered that they'd like to use it like a dictionary to find ample examples, to clarify the meaning of words, and to translate Japanese sentences into English. Since they are already used to using on-line search tools such as Yahoo and Google as part of their daily routines, we believe that if the concordance tool were available in a similar way at no cost, then learners would use it.

3.2 Did learners discover the one-to-many relationships between the two languages?

We asked students the open-ended question "What did you find by using parallel corpus concordancing?" One of the most frequent findings given by students (thirty-eight percent of students' answers) was that they noticed the one-to-many relationships between the two languages.

3.3 Did learners explore the lexicogrammatical and collocation patterns in English?

Students reported they had learned new words or specific information regarding the lexical and collocation relationships between words, as written in the students' reflective responses (Daily Evaluations) to each lesson. These included: "We were able to find frequently used lexical collocations"; "We found that English word combinations produce various meanings"; "We noticed that slightly different expressions are used to express similar meanings"; and "We thought it was a convenient tool for finding which parts of speech come after a word."

3.4 Was the concordancing tool easy to use?

Forty percent of students agreed and thirty-two percent disagreed with the statement "The tool was easy to use". Generally, students were able to manage easily with ten minutes of instruction in the first week. The general complaints from students about the concordance tool were (1) the concordancing program was unable to produce certain kanji characters so these looked garbled; and (2) students had to do the initial setting of both languages (English and Japanese) every week before they could start concordancing, and this was a nuisance.

Although the Paraconc concordancer used in this study was hugely beneficial for illustrating the potential for concordancing in data-driven learning, its limitations for Japanese-English applications are (1) it's expensive for classroom use; and, (2) it cannot handle certain kanji characters since it was not designed exclusively for the Japanese language.

To address these issues, we developed our own Japanese-English on-line bilingual concordancer, called *Text-Searcher*©, using the newspaper parallel corpus (Utiyama and Isahara, 2003). We also incorporated students' feedback on optimal user interface features. This first version is free and available at http://www.kotonoba.net/~snj/cgi-bin/text-search/text-search.cgi. Because it is Japanese-English-specific, it can display citations with the correct kanji. Credit for the original design and the development of Version 1 belongs to Masao Utiyama, and the development of a new search engine is underway in cooperation with Shinji Miura of Kobe University Graduate School. Future modifications will be made to later versions as our research continues to identify effective features.

4   Conclusion

Corpus-driven language learning using a Japanese-English parallel corpus allowed learners to compare the vocabulary and language patterns between two languages to not only acquire new vocabulary, but understand patterns and usage in both languages such as the one-to-many relationships between languages and collocation patterns. One of the strengths of concordancing use was the discovery aspect--- students are not "taught" these language patterns but find them by themselves (with guidance from handouts), and this kind of discovery can be both powerful and highly motivating. Corpus-based activities have generally been very difficult for beginner level students and as such, this kind of tool has had limited use. However, incorporating bilingual concordancing lessons with appropriate CALL vocabulary teaching material provides a successful learning tool for these EFL students.

The activities seemed to be well received by students and although the evaluation of the learning effect in this study was mainly based on learners' impressions, we were able to discern from their feedback that this system was both meaningful and useful. Further research will explore parallel-corpus applications to grammar in a CALL environment, and the on-going development of *Text-Searcher©*, an on-line bilingual concordancer.

Acknowledgements

Notes

1 The 640 words and ten of these DDL lessons discussed in this section are available at no cost at http://www5d.biglobe.ne.jp/~chujo/.
2 They complained that the activity was time-consuming, that they weren't able to understand its value, and that they preferred to learn words with the CALL method.

References

Barlow, M. (2003) *ParaConc: A Concordancer for Parallel Texts*. Houston: Athelstan.

Chujo, K. (2003) "Eigo Shokyuusha-muke TOEIC-goi 1 & 2 no Sentei to Sono Kouka [Selecting "TOEIC Vocabulary 1 & 2" for Beginning-level Students and Measuring Its Effect on a Sample TOEIC Test]." *Journal of the College of Industrial Technology*, *Nihon University*, 36, 27–42.

Chujo, K., T. Ushida, A. Yamazaki, M. Genung, A. Uchibori, and C. Nishigaki. (2004) "Visual Basic niyoru TOEIC-you Goiryoku Yousei Software no Shisaku (3) [The Development of English CD-ROM Material to Teach Vocabulary for the TOEIC Test (Utilizing Visual Basic): Part 3]." *Journal of the College of Industrial Technology, Nihon University*, 37, 29–43.

Tono, Y. (2003) "Corpus wo Eigo Kyoiku ni Ikasu [What corpora can do for language teaching]." *English Corpus Studies*, 10: 249–264.

Umesaki, A. (2004) "Eigo Gakushuu no Tameno Daikibo Corpus no Riyou [Using a large-scale corpus in the classroom]." Paper Presented at the 24th Japan Association for English Corpus Studies (JAECS) Conference, Tokyo, Japan.

Utiyama, M. and H. Isahara. (2003) "Reliable Measures for Aligning Japanese-English News Articles and Sentences." *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, 72–79.