Introducing the Shogakukan Corpus Query System and the Shogakukan Language Toolbox

Takahiro NAKAMURA
Shogakukan, Inc.
2-3-1 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-0051
JAPAN
takahiro@shogakukan.co.jp
June TATENO
NetAdvance,Inc.
2-30 Kanda-Jimbocho, Chiyoda-ku,
Tokyo, 101-0051
JAPAN
june.tateno@netadvance.co.jp
Yukio TONO
Department of Foreign Language and Culture
Meikai University
8 Akemi, Urayasu,
Chiba, 279-8550
JAPAN
y.tono@meikai.ac.jp

Abstract
Since the British National Corpus has detailed bibliographical and demographic information as well as part-of-speech tags encoded in XML, there is high demand for sophisticated search functions using above-mentioned markups and tags as search keys. Current search software available, however, has not effectively utilized this feature. Shogakukan has developed two systems with which lexicographers can investigate word behaviour for better dictionary making. In this demonstration, we would like to introduce these two systems: the *Shogakukan Corpus Query System* (SCQS), which is used for on-line searches, and the *Shogakukan Language Toolbox* (SLTB), which is used for off-line searches using batch files. Both systems are based on a database, which has the Corpus Query Language at the front-end. This Corpus Query Language is very useful for flexibly writing annotation as search conditions. These two systems can be accessed from conventional Internet browsers.

1 Introduction

In Japan, corpus-based English-Japanese dictionary making has begun only recently. Since the BNC World Edition was released in 2001, there has been a growing demand among lexicographers to have sophisticated search techniques for large balanced corpora. Shogakukan has recognized this need and started to develop our search systems in the year 2000 and released them for in-house use in 2002.

2 Requirements

The system requirements are as follows:

1. There is a need to define a Corpus Query Language (CQL) so that various annotations of tagged corpus can be flexibly dealt with in search syntax. A CQL must be developed to describe mixed search conditions, such as conditions of subcorpora, phrase keys, and POS tags. For example, one should be able to search the pattern such as all instances of "give up" followed by gerund in the demographic section of spoken data. In addition, it should be possible to easily extend the CQL if the attribution of words, such as lemma and semantics codes, is added.

2. There should be a command-interface for off-line search using batch files. When developing dictionaries, if the list of queries is prepared in advance, we can get more reliable results because we can obtain them without individual differences of search skills. The On-line SCQS can use the same search engine as the SLTB.

3. It should be compatible with web browsers. There should be no need for the installation of special client software and maintenance. Our system is much less costly than systems that require installation and maintenance.

3 Functions

1. *Complex Full Text Search*: non-linear pattern search of combinations of subcorpora, key words and POS tags.
2. *Post-Processing*: KWIC concordance data that is searched once is saved on a server. Then it is sorted out and counted N-gram according to the search requirements.
3. *Collocation Table*: a collocation table can be created based on different collocation statistics such as raw frequencies, T- Scores, Mutual Information, LogLog statistics and attributes of words (POS tags, Lemma etc).

4 Graphical User Interface of SCQS

To allow pattern searches to be written intuitively, we provide a user-friendly interface and use the query table shown in Figure 1. The columns indicate the word order and in each row, the attribution of a word can be specified. Figure 1 shows the query pattern "*give up* + V-ing." Figure 2 shows the query pattern "*give up* + NP".

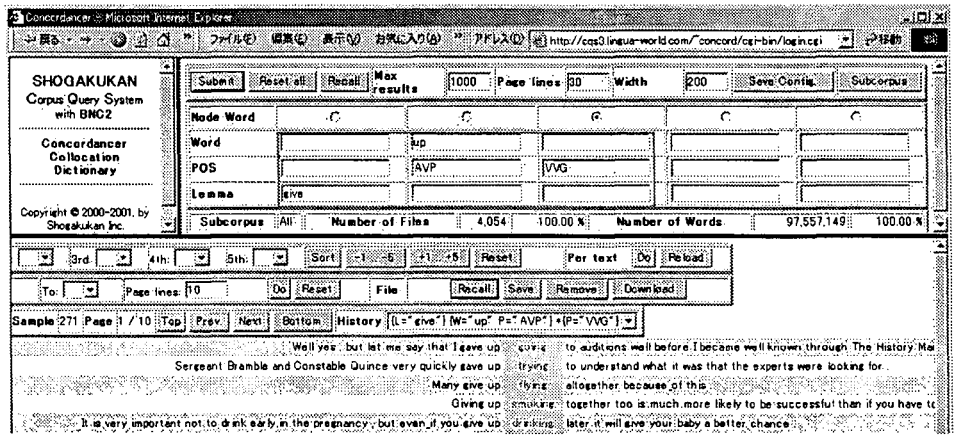

Figure 1: The Shogakukan Corpus Query System Web Interface ("*give up* + gerund")

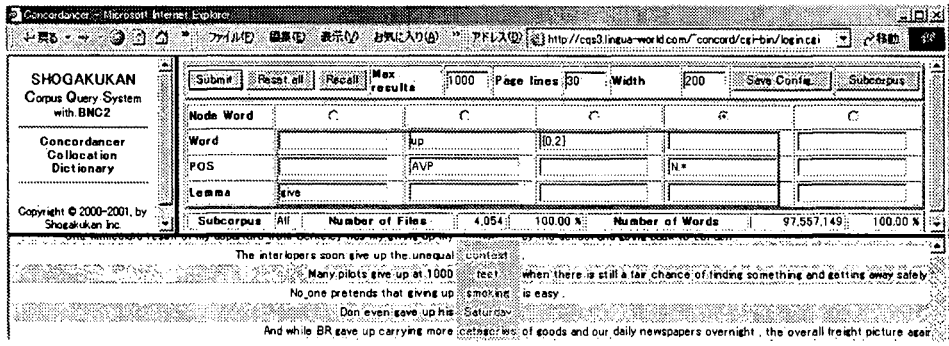

Figure2: The query pattern "*give up* + NP"

5 Graphical User Interface of SLTB

The windows of the STLB are divided into five frames as shown in Figure 3:
Upload of files: uploads scripts edited on a client computer onto a server.

① View of current directory: utilizes the viewer for download files.
② Command Menu: provides a dialog box for supporting command entry
③ Row for entering command-lines
④ Result Window Area: shows standard unix output and error output.

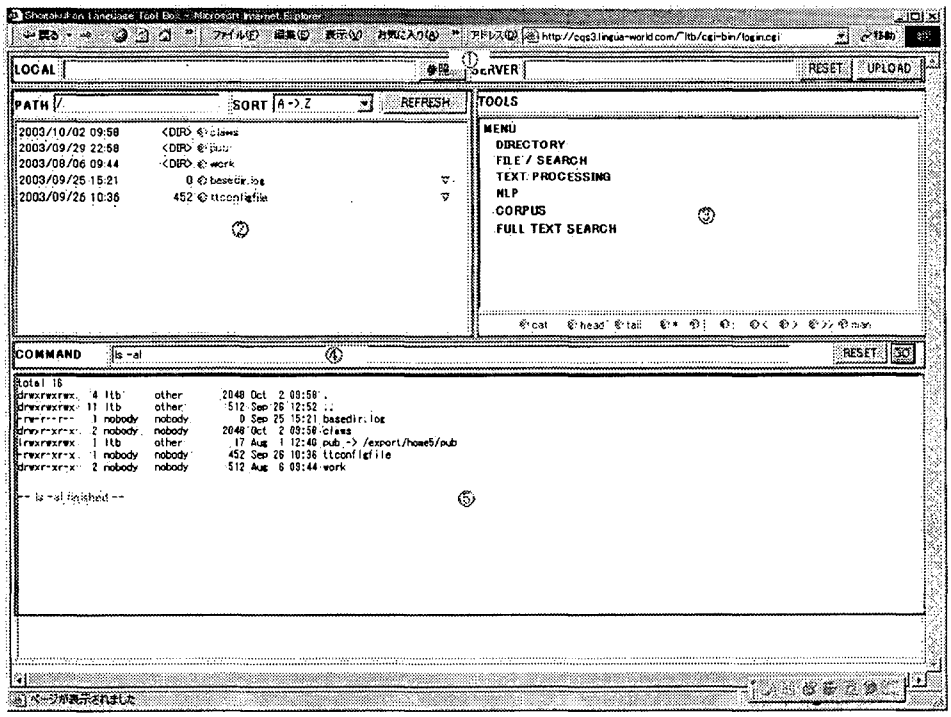

Figure 3 : The Interface of the Shogakukan Language Toolbox

6 Outline of the Demonstration

1. In the SCQS demonstration, we will show a complex pattern search, a combination of sub-corpora, lemma, POS tags and wild cards.
2. We will carry out a collocation search using an assortment of lemma, POS and statistics (T-Score, M.I., LogLog) and shows collocation tables.
3. In the STLB demonstration, we would like to proceed as follows.
   First, we will prepare a short script written search patterns by the SQL, and upload the file from a client machine to a server. Then we will run the script on the server and obtain the results. In addition, we will show how to get an N-gram list from the result file, and set the operation of data sets of KWIC.

7 References

Nakamura ,T and Y. Tono, 2003 Lexical Profiling Using the Shogakukan Language Toolbox *ASIALEX2003 Proceedings: Dictionaries and Language Learning: How can Dictionary Help Human & Machine Learning?* The Third Asialex International Congress, August 27-29, 2003, Meikai University, Japan, pp.170-176

Tono, Y., H. Iwasaki, T. Nakamura, M. Suzuki and E. Egawa, 2001 Shogakukan Corpus Query System in Collaboration with the American National Corpus Project. In Lee, S. (ed.) *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary.* The Second

Asialex International Congress, August 8-10, 2001, Yonsei University, Korea, pp. 231-235.