Using the BNC to create and develop educational materials and a website for learners of English

Danny Minn[a], Hiroshi Sano[b], Marie Ino[b] and Takahiro Nakamura[c]

[a] Kitakyushu University

[b] Tokyo University of Foreign Studies

[c] Shogakukan Inc., Electronic Dictionary Development Department

Abstract

A project team at Tokyo University of Foreign Studies (TUFS) developed a website that allows users to download sentence patterns for English educational purposes. These sentences were extracted from the British National Corpus (BNC) by writing search expressions using the Language Toolbox (LTB). Prior to the creation of the expressions, we surveyed major English textbooks used widely at Japanese educational institutes, for the purpose of covering the most often studied English sentence patterns. The data was stored in an XML database and was made available to users through the Internet. This project was carried out in collaboration with Shogakukan and Sano Laboratory at TUFS. The website was provided as a part of the Shogakukan Corpus Network (SCN) services.

1 Introduction

1.1 Research goals

The goals of this project were: (1) to develop an efficient method to create educational materials for learners of English based on language usage data, and (2) to offer the materials through a website.

In order to improve one's overall practical language ability there are at least four skill areas that must be improved: reading, writing, listening, and speaking. Our aim was to develop methods and materials that are useful for the improvement of grammatical ability related to those four skills. We have conducted a comprehensive survey of all types of English textbooks currently used by Japanese middle schools and high schools in order to see exactly what sentence pat-

terns are studied in the current educational curriculum. Using our results, we have teamed up with Shogakukan, Inc.'s Multimedia Division to search the British National Corpus (BNC) and extract English example sentences. Grammatical information was added to the collected data of extracted sentences to make an XML database. Then a website was constructed so that the XML database could be accessed through the Internet. This website was a service made available through Shogakukan Corpus Network (SCN).

1.2 Background

In Japan, the need for improvement in English education is not limited to the problem areas in educational curriculums. There is high demand in Japanese society for improvement, and it has even become a matter of public policy. For example, one of the themes of the 2004 "Modern Educational Needs Aid Program" (The Ministry of Education, Culture, Sports, Science, and Technology, also known as MEXT) was "Training Japanese to Use English in the Workplace."

Furthermore, as the business world becomes more internationalized and economic activities become more borderless, the demand for skilled workers who have good English ability becomes higher. It is even more of an asset if one has language skills that pertain to a specific field of work. The need for improvement of English ability within specialized fields and even in the English education field has been recognized. That demand is also seen in companies that need materials adapted to English for Specific Purposes (ESP). It is clear that more effective educational methods need to be established.

Conversely, the reality is that as the official Japanese curriculum calls for less rigid study, the amount of time allotted for English study will go down, as well as the amount of study items in textbooks. The influence of this relaxation in the curriculum will not be insignificant, and it creates a need for a framework to be established where improvement and effectiveness in English education are possible.

We do not desire to judge the merits of the decrease in educational content or that the amount should be gradually increased. Our aim, however, is to make the study of English more efficient, especially concerning the content that students are required to study. We believe that by including in educational materials the language spoken and written in the real world, we can improve efficiency.

1.3 Approach

If we considered the process of speaking in terms of a kind of problem solving, we might imagine some speakers asking themselves when confronted with a

speaking task and situation, "What kind of language is needed?" (linguistic knowledge), and "How do I use it?" (language usage knowledge). This subconscious planning is part of the ability that a learner tries to develop in language study. Up to now, most educational materials have been focused on the study, understanding, and acquisition of linguistic knowledge. As a result, English education in Japan has tended to lean too far towards linguistic abstraction.

For every speaking condition, there are expressions and language forms (including linguistic knowledge) that are used in certain speaking strategies (solution strategies). By studying these language forms, knowledge of language usage can be gained. That is, educational materials should not only include linguistic knowledge. They should also include information about language usage. It is possible to lessen the burden on learners by making their acquisition of required grammatical items more efficient and by including more usage data.

The efficiency-driven materials mentioned above are especially suited for ESP learning. They focus on the following: (1) vocabulary and sentence patterns that are specific to certain fields, (2) improvement of reading and writing ability, (3) content geared towards learners reaching their own language performance goals, and (4) efficient study methods.

The example sentences and practice drills in textbooks may have been insufficient in number, but they served as good data to base our research on. Our research aimed to construct a method to automatically extract example sentences based on the characteristic patterns we found in textbooks.

In addition to increasing efficiency and lessening study load, we were concerned with the following problem areas: (1) because of the labor-intensive nature of creating teaching material, large amounts of varied material can not be made in a short time; (2) the quality of material largely depends on the creators' ability; (3) many of the creators of such material are not native English speakers, so the expressions included tend to be lacking in variety, and the quality of example sentences can not be immediately guaranteed.

1.4 Creating a website offering educational material for English learners

In a previous research project called N-Cube, the goal was to develop a language learning methodology that responded to the varied educational needs of learners (Sano 2002: 55–62, 2003: 34–44). N-Cube was a framework designed to aid language learning through the efficient creation of materials for ESP. In the framework, we employed: (1) the Cognition-Based Rule-Driven / Data-Driven (RD/DD) Interactive Learning methodology that is characterized by its focus on cognition and native language usage, and (2) language usage data and natural language processing technology.

The overall process and goal of our research is outlined in Figure 1. The scope of this paper is represented by (3) to (7) in the diagram. The Language Tool Box (LTB) is a workbench for corpus use developed by Shogakukan, Inc., represented by (4) in the diagram.

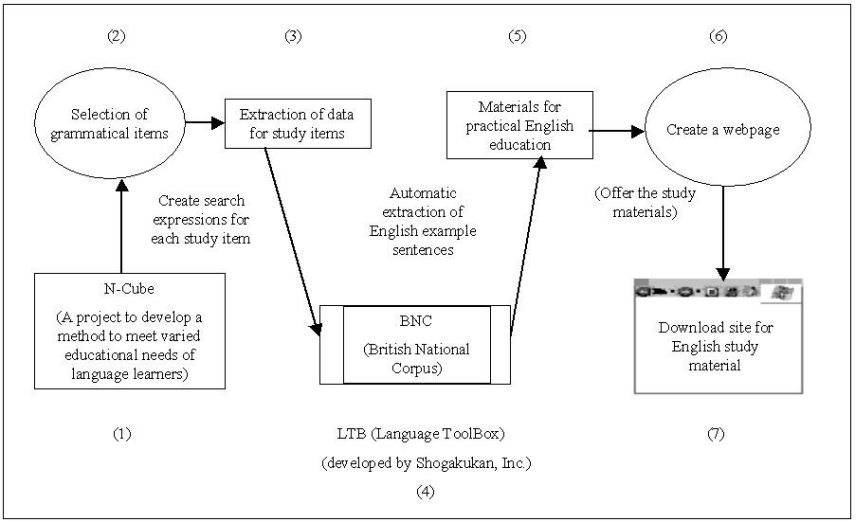

Figure 1: Overall process of the research

The sentence patterns that were characteristic of textbooks were recorded (3). Then the LTB was used to extract examples from the BNC that matched the patterns (4) and (5). Grammar tags were added to the extracted data to create an XML database. This facilitated the construction of the website offering English educational material (7). The website was offered as a service of SCN.

The SCN and LTB are explained in Section 2, and the writing of the search formulas used to correspond to the study items in textbooks is explained in Section 3, which also describes how the LTB was used to extract examples from the BNC. Improved work efficiency and the construction of the website that offers educational material is explained in Section 4.

2 Corpus extraction system

2.1 Shogakukan Corpus Network

In 2003, Shogakukan, Inc., established the SCN, a commercial site aimed at broadening the use of corpora for linguistic research and language education. Currently, the center of its service is corpus searching. When it was established, one of the goals was to contribute to the improvement of dictionaries and language textbooks through the advancement of corpus accumulation and technology.

In April 2004, the WordbanksOnline section of the site was released. There are also plans to release other services in the near future, such as the Japanese EFL Learner corpus (JEFL), the Professional English Research Consortium (PERC), and the American National Corpus (ANC). These planned services will all be accessible through the same search interface, and the corpora will be tagged via the same software.

Prior to the release of SCN, Shogakukan had been researching corpus-based dictionary editing techniques from the late 1990s. For the purpose of constructing an intra-company electronic editing infrastructure, Shogakukan started developing a corpus search system. At the center of this system was the LTB. The SCN is a more user-friendly, publicly available form of the results of Shogakukan's in-house research and development. The LTB that will be described in the next section is generally not for public use. It may be used with permission, however, on the condition that research results will be shared with SCN.

2.2 An outline of the LTB

The Shogakukan LTB is a server/client system that uses a corpus search engine through a command interpreter. The construction of the search commands (fcql) is based on Corpus Query Language (CQL). The commands offered are not limited to search functions, as there are other useful commands available such as the following (Nakamura 2003: 170–176, 2004: 147–152):

- kwic: search results are organized in KWIC form
- cluster, clusterC: search results can be narrowed by using queries regarding grammar categories and length of strings.
- colloc, collig: by selecting available grammar categories and word location variables, collocation statistics can be gathered.
- cql_and, cql_or, cql_diff, cql_symdiff: search results can be further narrowed using these Boolean commands
- random: a random sampling of examples can be extracted from search results

The output data of search results is produced in its “most detailed KWIC form” initially. It includes the placement number from the query, node-specific information, and morphological and syntactic information (word and sentence generic group information). From this data, it is possible, for example, to process it further to trace all of the occurrences of a chosen word within a specified grammar category. In the top line, the query information is added as a header.

The LTB is used with a browser, and it allows different functions for different frames. Operations are made simpler through the use of dialog boxes and the options and switches used for commands. Commands are sent to the server via Common Gateway Interface (CGI), and the processed results appear in the display area (standard output). Figure 2 is a screenshot of the LTB.

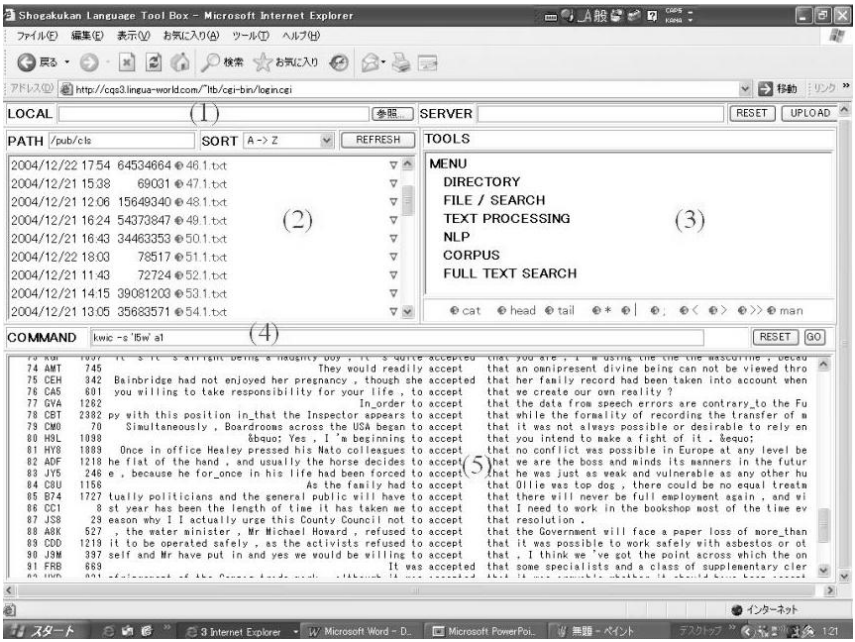

Figure 2: Screenshot of the LTB

(1) File uploaded from client to server.

(2) File organization on the server. Directories are displayed, and files can be moved between directories.

(3) Options and switches for various commands are selected in dialog boxes. Data can be classified and organized through the different commands.

(4) The interface uses a command interpreter (command line).

(5) Results are shown in the display area (standard output, standard error output).

2.3 CQL and its functions

In the BNC, the data was tagged with two types of information:

- classification and sentence attribute information in the header
- word attribute information

The header classification information could be used for searching sub-corpus information. The sentence attribute information indicated whether the sentence was from conversation or written language. The word attribute information was added by a tagging program (CLAWS).[1] The word attribute information included the lemma (dictionary form) and a code for part of speech data. Shogakukan has developed a language (CQL) where a corpus can be queried using the header and word attribute information mentioned above.

The output can be aligned so that various collocation patterns can be compared, such as by word (syntax) or by attribute (paradigm). For example, when one wants to investigate the pattern "give up + Gerund" while ignoring the paradigm change of "give", the following pattern is used: [lemma="give", lemma="up", part of speech code="VVG" (in the BNC, VVG is the "-ing" code)].

When the length of an item to search is not exactly known, approximated searches can be made. For example, searching for "give up + noun phrase" with "modifier/no modifier + noun" in the noun phrase will produce matches of noun phrases with a similar construction. The "modifier/no modifier" or other wildcards can be used in searches. Therefore, as illustrated here, CQL is a simplified language that can be used to create pattern matches of words and attributes in a two-dimensional arrangement.

2.4 Implementation of the database

Previously Shogakukan had worked on a program called OpenText (ver. 5) that had a SGML total text search engine. This previous experience was valuable while considering search functions and performance for the current project (Nakamura, 2004: patent pending).

A direct query can be made in OpenText using its PAT commands and region setting (by setting a tag name and making a sub-index file, a search can be limited to only matches with the set tag). A search can be undertaken in the header, by means of keywords and conditions. An example would be [conversational, "give up"].

Accessing header information as a simple record in a database was not a difficult problem. However, it was not possible to access part of speech, lemma, and wildcard information, and set parameters for sentence length using the PAT commands as they were. Because the PAT commands allowed setting length by byte amount only, searches with several wildcards or any kind of more complex queries would return too many erroneous matches.

In order to solve this wildcard problem, a method of setting the number of comparisons was set. At the same time, tokens were replaced by four-byte IDs and index files were created. Since the corpus that was used was not updated, the token-ID replacement chart was very useful.

By using the fixed length IDs, the data size itself was compressed. Also, by deleting verbosity, assuming that the average length of English words is four bytes, the file size of the BNC (about 1.3 GB) was reduced to about 30 per cent. For PC-UNIX environments where memory cost is low, all index files and sources could be deployed in the memory. Because operation was made possible without complex pattern searching and file accessing, response time was greatly improved.

3 Methodology for extraction of example sentences

3.1 Selection of grammatical items

A survey of the textbook market (based on sales figures) in Japan was conducted, and the top 31 textbooks were chosen (middle school English textbooks: 6, high school English textbooks, English I: 8, English II: 8, Writing: 9). Four English reference books that were widely used in Japan for English education were also chosen. Through careful analysis of the selected books, it was possible to find the grammatical items most often studied by Japanese students in class. There were two conditions for selection of items: (1) items that were common to the main textbooks, and (2) items that may not have been necessarily common to the main textbooks, but were common to many of the textbooks and judged to be as highly productive in learning language usage. Under these conditions, 153 grammatical items were chosen and organized. The 153 items included grammatical items from the simple "I am + noun" to the more difficult "SVC", "emphatic sentence pattern", or "subjunctive mood".

Basically, the 153 items were in the affirmative form. For each grammatical item, 14 subordinate items, such as negative forms or interrogative sentences, were made. Therefore, multiplying 153 items by 14 subordinate items resulted in 2,142 items. Of those, however, 747 patterns did not actually exist as proper English sentences, so after deleting them, the total number was 1,395 sentence patterns.

3.2 Search expressions and extraction of example sentences

CQL expressions were made for each of the 1,395 sentence patterns. For example, for "'How' exclamatory sentences" the following expression was used:

^{W="how"} {P="AJ0|AV0"} [0,10] {L="!"}$

This expression means to, "find a sentence that starts with the word 'how', continues with an adjective or adverb, is between 0 and 10 words in length, and ends with an exclamation mark." These types of search expressions were used to extract example sentences for all 1,395 sentence patterns.

3.3 Evaluation of extracted example sentences and adjustment of search expressions

In order to ensure the educational quality of the extracted data, the search expressions were improved. The BNC used in the LTB also has morphological analysis information that can be referenced. Using CQL queries, specified information about a word can be searched for, such as conjugated form, lemma (dictionary form), and part of speech. For example, for part of speech classification there is no distinction between intransitive and transitive verbs (in the BNC). An example would be the SVO (O = "that ~") sentence pattern, where there is "a transitive verb with a 'that' phrase in the object (O)". Even though the transitive verb (V) part of speech can not be set in a search, examples can be extracted through CQL expressions with 93 per cent accuracy because the object can be set. An example of results with low accuracy would be the SVOC sentence structure where there is "a noun as the object". Since the verb part of speech can not be set and O must be a noun, the CQL expression only achieved 50 per cent accuracy.

The most common method of extracting example sentences is based on sentence structure parsing results. However, with the benefit of having a large-scale corpus of 100 million words, compared to morphological analysis, the accuracy of sentence parsing results is lower. Even by improving the accuracy of parsing analysis, the accuracy of the results does not improve significantly. Therefore, we based our method on morphological analysis results by: (1) creating search

expressions and extracting corresponding example sentences, (2) evaluating the results and finding the collocations that caused erroneous results, (3) based on that evaluation, making additional search expressions that took out the erroneous results, and (4) including the additional search expressions, doing the extraction again to create the example sentences without the previous erroneous results. An outline of this refinement of the extraction results is represented in Figure 3. The top portion represents the results after the additional expressions were made to remove erroneous matches.

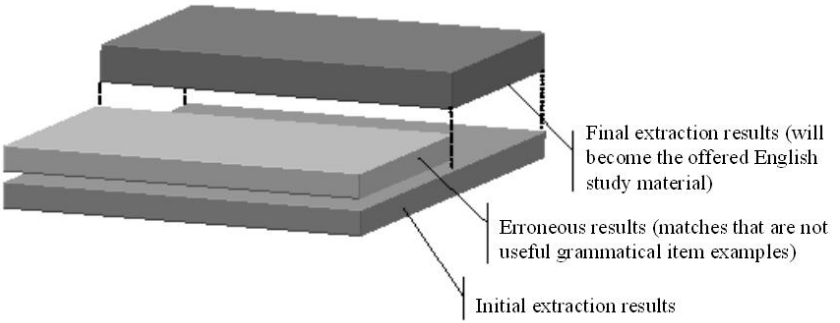

Figure 3: Extraction evaluation and improvement of accuracy

4 Managing the data with XML

4.1 Extraction through the use of search expressions and HTML coding

The work involved in making the search expressions, extracting the data, and converting it to HTML was done concurrently by several people. Because they had to share their work and results, it was necessary to have a shared environment using XML to manage the data.

Figure 4 shows the work flow of extracting the example sentences. As the diagram shows, several project team members created and entered their share of search expressions and extracted example sentences. Each person did not, however, access the LTB directly to do their search. They edited and updated the search expressions in a shared XML file. A batch file was used to execute the actual search of the LTB by using a script to do all of the search expressions in the XML file. By managing the data through the use of XML, everyone could always see updated information and loss of information was prevented. This also standardized the quality of data.

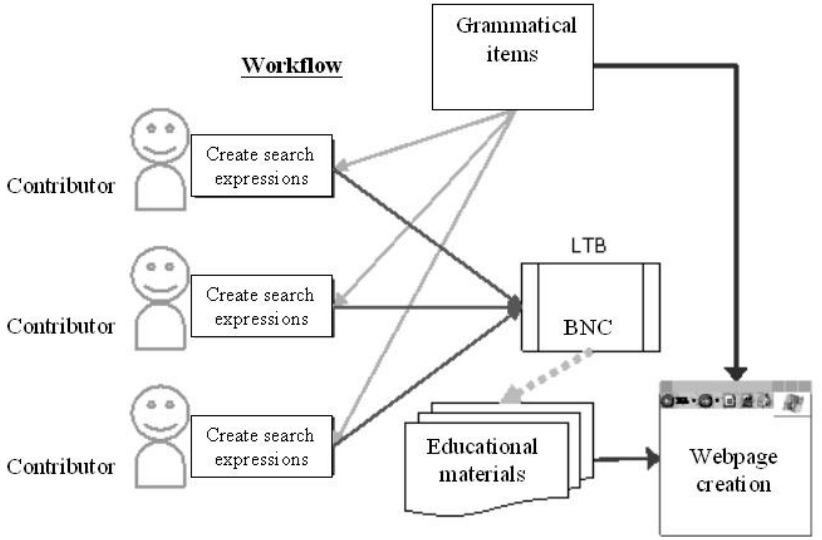

Figure 4: Work flow of extracting the example sentences

At the same time, not only the search expressions but also the grammatical items and their explanations could be managed by using tags. Also, because the data could be used with an XSLT template, the HTML conversion could be done automatically and visual confirmation was made easy. Any revisions in the data were done in XML, so it was easily converted to HTML (see Figure 5).

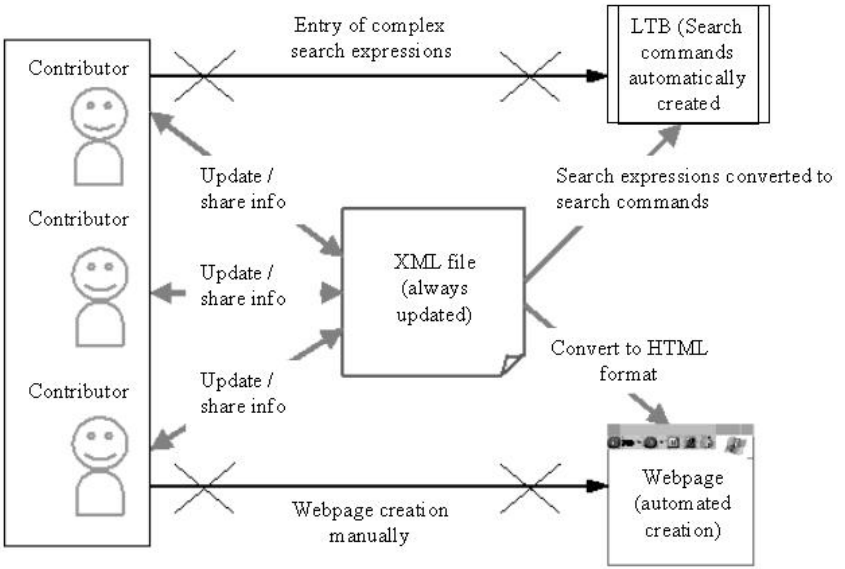

Figure 5: Managing the data

4.2 Designing a website to offer the English educational material

Currently, the target for the website offering the educational material would be teachers of English who teach non-native middle school, high school, and college students. Depending on a user's needs, one could find data to match specified sentence patterns to supplement his/her materials or make new materials with examples of English used in the real world.

Many textbooks have several grammatical items listed and highlighted in their table of contents. Therefore, it would be easy for users to use the grammatical items as keywords in their searches. The website is designed so that users can navigate through the grammatical items, confirm the sentence pattern explanations, and download the example sentences needed (see Figure 6).

Each grammatical item was given a corresponding number and name. Depending on the textbook, explanations for grammatical items varied somewhat, so for the purpose of clarity, terms used in the item definitions that appeared in each type of textbook were added as grammatical explanation data. Moreover, example sentences that correspond to chosen grammatical items can be downloaded from the site.

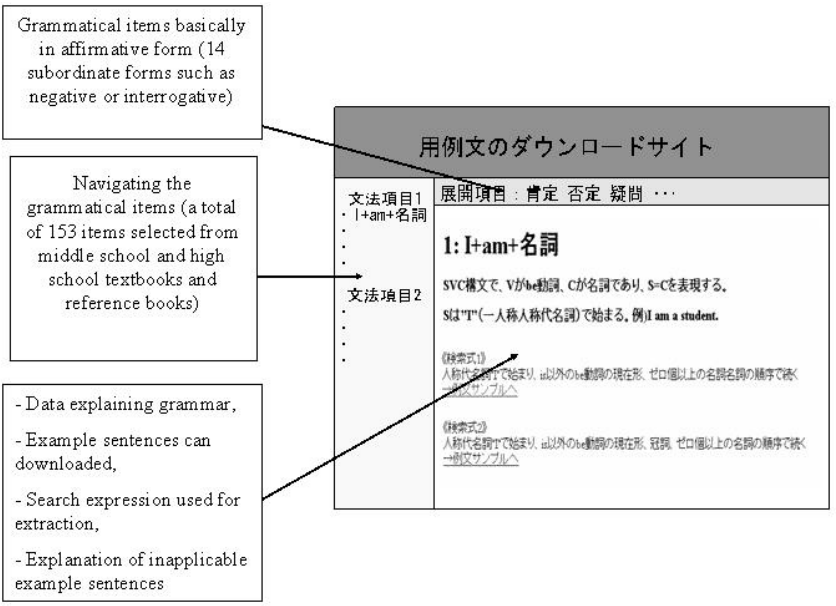

Figure 6: The design of the website

4.3 The constructed website

An example of the constructed website is shown in Figure 7.

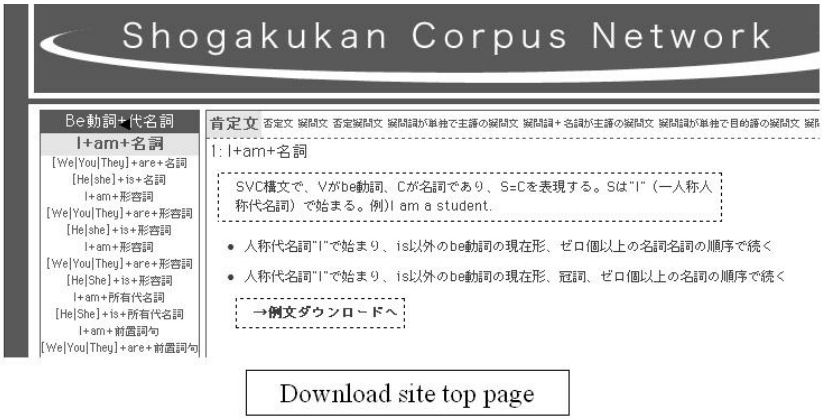

Figure 7: The website display

On the right-hand top side of the display, there are 14 links that show explanations regarding the material when they are clicked. Included in these explanations is information about each sentence pattern, the example sentences provided, and reference data for users actually using the data for teaching material. Evaluation results regarding the extracted data are also given. More specifically, the reason why a certain percentage of example sentences are inapplicable to the selected grammatical item is explained. The erroneous examples are qualitatively analyzed and explained. Finally, the search patterns used for each selected item are also given here for users' reference.

5 Future tasks for the project

The next step in this project will be to make the website open to the public and evaluate the site using user data and suggestions. The type of input needed from users would be data such as the number and kinds of grammatical items desired, the number of example sentences per item needed, and the level of satisfaction concerning the sentence length and vocabulary level settings. The improvement of the site will reflect the needs of the users.

Notes

1. The CDROM version of the BNC that was in general use did not have lemma information included. CLAWS was used in order to add lemma information.
2. Special thanks: This project was sponsored by the following: (1) The Ministry of Education, Culture, Sports, Science, and Technology (MEXT), 2004–2006 Research Grant (Fundamental Research (B)(2)), "All-digitized textbook data analysis and large-scale Japanese corpus construction", (Research Head: Hiroshi Sano), and (2) 2003 Shogakukan Inc., Multimedia Division Research.

References

Arai, Masayuki, Ami Watanabe and Hiroshi Sano. 2004. English language educational materials based on linguistic usage and development of a corresponding website (translated title). *Japanese Society for Information and Systems in Education, 29th Annual National Conference Proceedings*, 257–258.

Iwakura, Takayuki, Masayuki Arai and Hiroshi Sano. 2004. Using linguistic usage data to create English language educational materials. *FIT2004, 3rd Annual Forum on Information Technology*, 519–522.

Nakamura, Takahiro and Yukio Tono. 2003. Lexical profiling using the Shogakukan Language Toolbox. In M. Murata, S. Yamada and Y. Tono (eds.). *ASIALEX 2003 proceedings. Dictionaries and language learning: How can dictionaries help human & machine learning*, 170–176.

Nakamura, Takahiro, June Tateno and Yukio Tono. 2004. Introducing the Shogakukan Corpus Query System and the Shogakukan Language Toolbox. In G. Williams and S. Vessier (eds.). *EURALEX 2004 Proceedings. The 11th EURALEX International Congress*, 147–152.

Nakamura, Takahiro, Hiroshi Aizawa and Ryouji Watanabe. 2004. *Search methods and devices in natural language text* (translated title). Patent pending special application no. 2004–047377.

Sano, Hiroshi. 2003. An educational aid system for making ESP-suitable contents (translated title). *CALL Research and Applications for Foreign Languages*. Council for Improvement of Education through Computers (CIEC), Foreign Language Education Research Group, 34–44.

Sano, Hiroshi, Marie Ino and Yoichiro Uno. 2002. Developing a learning environment adaptable to diverse needs through a linguistic educational aid system (translated title). *Information Processing Society of Japan, Informatics Symposium Proceedings*, 55–62.