사전 편찬학 연구 제11집 1호(2001)
pp. 127~138

A lexicographer's workbench for new-word selection & identification: Web-server-based automatic identification of neologism candidates and the client editing system

NAKAMURA, Takahiro & TONO, Yukio
(Shogakukan, Inc., Japan · Meikai U, Japan)

연세대학교 언어정보개발연구원
Center for Linguistic Informatics Development(CLID)
Yonsei University

A lexicographer's workbench for new-word selection & identification: Web-server-based automatic identification of neologism candidates and the client editing system

NAKAMURA, Takahiro*
TONO, Yukio**

Abstract

We will demonstrate an interface which allows the lexicographer to view, edit, and store the results of an automatic extraction of new-word candidates from English newspaper resources on the Internet. There has been a growing demand among the Internet communities to access dictionary services or new-word databases via the Internet or via mobile phones. In order to meet such a demand, Shogakukan, Inc. has internally developed a system which extracts new-word candidates from large newspaper corpora and checks the candidates using a dictionary entry database primarily based upon a bilingual version of the Random House *English Dictionary*. This system also provides the lexicographer with a client database system to view contexts for the candidate words, write or edit the entry, and store & manage the entries. The system consists of three main units: (1) a server for browsing & downloading Internet texts, (2) a server which checks words against the entry

* Shogakukan Inc. 2-3-1 Hitotsubashi Chiyoda-ku, Tokyo 101-0081, JAPAN
takahiro@mail.shogakukan.co.jp

**Department of Foreign Languages and Cultures, Meikai University 8 Akemi, Urayasu, Chiba 279-8550, JAPAN / y.tono@meikai.ac.jp

database and extracts potential candidate words with great accuracy, and (3) a database client-server system, on which the lexicographer can view various examples of a potential candidate word in context and edit the sample entry. The demonstration includes a guided tour of the new-word candidate extraction process and the client editing system.

1. Background

The development of the electronic media environment continues to bring about fast and radical changes to the way information is delivered to end users. In the area of electronic dictionaries, however, the movement has been rather slow in the sense that the development of electronic dictionaries has, up till now, been a mere conversion of original paper dictionaries into the electronic format. The mode of delivery of dictionary content has to undergo a radical change, however, as we have increasingly different modes of reference media services such as hand-held dictionaries, personal digital assistants (PDAs), as well as the search-on-demand system available via the Internet or the mobile phone. Dictionary publishers are now being urged to optimize dictionary content to fit the specifications of the different media. Especially in the case of telecommunication services, users are demanding faster access, higher hit rates, and wider coverage of new words as added value. Therefore, dictionary publishers need to shorten the revision cycle, to meet the demands of potential new users.

Until recently Japanese lexicographers have depended primarily on published dictionaries of new words (e.g. *Third Barnhart Dictionary of New English*; *The Longman Register of New Words*; *The Oxford Dictionary of New Words*) in order to supply new entries for their dictionaries. The growth

of Internet communities and the growing number of web pages, however, mean it is now possible for us to access newspapers and periodicals directly. For example, according to Yahoo, more than 4,000 newspaper web sites are accessible via the Internet in North America. These constitute rich resources for the systematic identification and extraction of neologisms. Shogakukan Inc. has produced several unabridged dictionaries such as the *Random House English-Japanese Dictionary*, the *Encyclopaedia Nipponica 2001*, the *Great Dictionary of the Japanese Language* and the *Dai ji sen* (*Japanese Dictionary*), all of which contain more than 100,000 entries. We keep the dictionary data in an SGML-like format so that they effectively constitute a machine-readable lexical database from which we can extract revelant data. Thus it is easy to extract a list of all dictionary entries or encyclopaedic indexes, or a list of dictionary entries with their corresponding parts of speech. The database' can first be processed using English and Japanese morphological analysers so we can obtain word lists in a very flexible way. Because of the highly structured nature of our lexical database and the ability of our morphological analysers to handle input/output data flexibly, the right infrastructure was in place, and the neologism project was deemed computationally feasible.

2. Overview of the new-word search system

The main thrust of this neologism project was the building of a lexicographer's workbench, an integrated system of tools or modules in which every part of the new-word search process is neatly consolidated within the system. The different processes involved are: analysing text strings from web resources, filtering the results automatically through existing word lists

and presenting potential candidates from the filtered lists to human lexicographers together with concordance lines, and supplying the necessary information for further lexicographical processes. There are already a few projects of this kind among NLP groups (e.g. Baayen and Renouf 1996; Renouf 1993). Renouf and her colleagues have developed a neologism search system to identify the process of how new words are created, while our primary interest is to build a lexicographer's workbench to find new words.

The system is composed of three servers: (1) the Data Collection Server, (2) the NLP Server and (3) the Lexicographer's Workbench Server. The client machines will access these three servers to obtain the necessary information. The following sections will elaborate on the features of the NLPS and the LWS.

2.1 The NLP Server

After the Data Collection Server downloads HTML files from specified web sites, the NLP Server matches text strings in the collected pages against the Shogakukan Dictionary Entry Database (SDED), and produce a list of new-word candidates. This output also includes total frequency counts for each candidate word, frequencies for each web site where the word appeared, and the file names of the pages. The original sentences/contexts in which candidate words appeared will be output in example files. To be more specific, the process involves four stages: (1) extraction of the initial candidate words (ICW), (2) filtering of the list of possible new words, (3) output of the new-word candidate list, and (4) output of illustrative example files.

2.1.1 Extraction of the Initial Candidate Words (ICW)

At this stage, the text in a given HTML file collected from a specified website will be split into word tokens, which are then matched against the Shogakukan Dictionary Entry Database (SDED), a dictionary entry list based a bilingual version of the *Random House English Dictionary* and other dictionaries of new words published from Shogakukan. The accuracy of extraction is ensured by filtering the candidate list with several other lists, e.g. a list of unwanted words, prepositions, and past participles. The words extracted as possible candidates have the following patterns:

A) A sequence of one word or more enclosed within quotation marks (“ ”)
B) A series of more than two words whose first letters are capitalised. (It also allows for cases where there are prepositions or ‘prep + determiner’s inside the sequence.)
C) A sequence of more than one word that does not appear in the SDED.
E) A sequence of two words that have the pattern ‘noun + noun’ or ‘adjective + noun’
E) A sequence of two words that have the pattern ‘noun + doing’ or ‘adjective + doing’
F) A sequence of two words that have the pattern ‘past participle + noun’

Patterns (A) - (C) will extract proper nouns (e.g. personal names, organization names, product names, names of books or artistic works) while patterns (D) - (F) will extract noun compounds. This matching process will produce the list of Initial Candidate Words (ICW), which consists of

candidate new words and compounds along with their file names. At the time of writing, we have more than 900,000 items in the ICW, which is one of the most useful outcomes of this project.

2.1.2 Updating of the Reviewed Neologisms Database

"Reviewed neologisms" refers to those words that have been collected by the checked and confirmed as new entries by lexicographers. Every time the Data Collection Server produces a candidate list, there is always the possibility that words previously confirmed as new entries will be included again. They therefore need to be excluded from the candidate list although their counts need to be added to the existing frequency profiles for the words. Thus, what we call the Reviewed Neologisms Database (RND) must be updated every time the Data Collection Server does a new search cycle. The process involves the updating of frequency counts as well as file names in the list.

2.1.3 Output of the Finalised Candidate Words (FCW)

The list of finalised candidate words (FCW) is produced after filtering out the ICW from the RND. If a word enters into this finalised list, then the frequency count per website will be calculated to check if this particular word appears with sufficient frequency in many different websites (i.e. has a high dispersion).[1] The candidate words with sufficiently high frequencies and dispersions will be stored in the FCW database with total frequencies,

1) This thresold of sufficient frequencies will determined by the lexicographer based on personal experience.

frequencies per website, and filenames.

2.1.4 Output of example files

The final stage is the output of example files. Example files have the candidate word flagged, with the first line containing the URL, making it possible to refer to the original web page when viewed in the browser.

2.2 The Lexicographer's Workbench Server

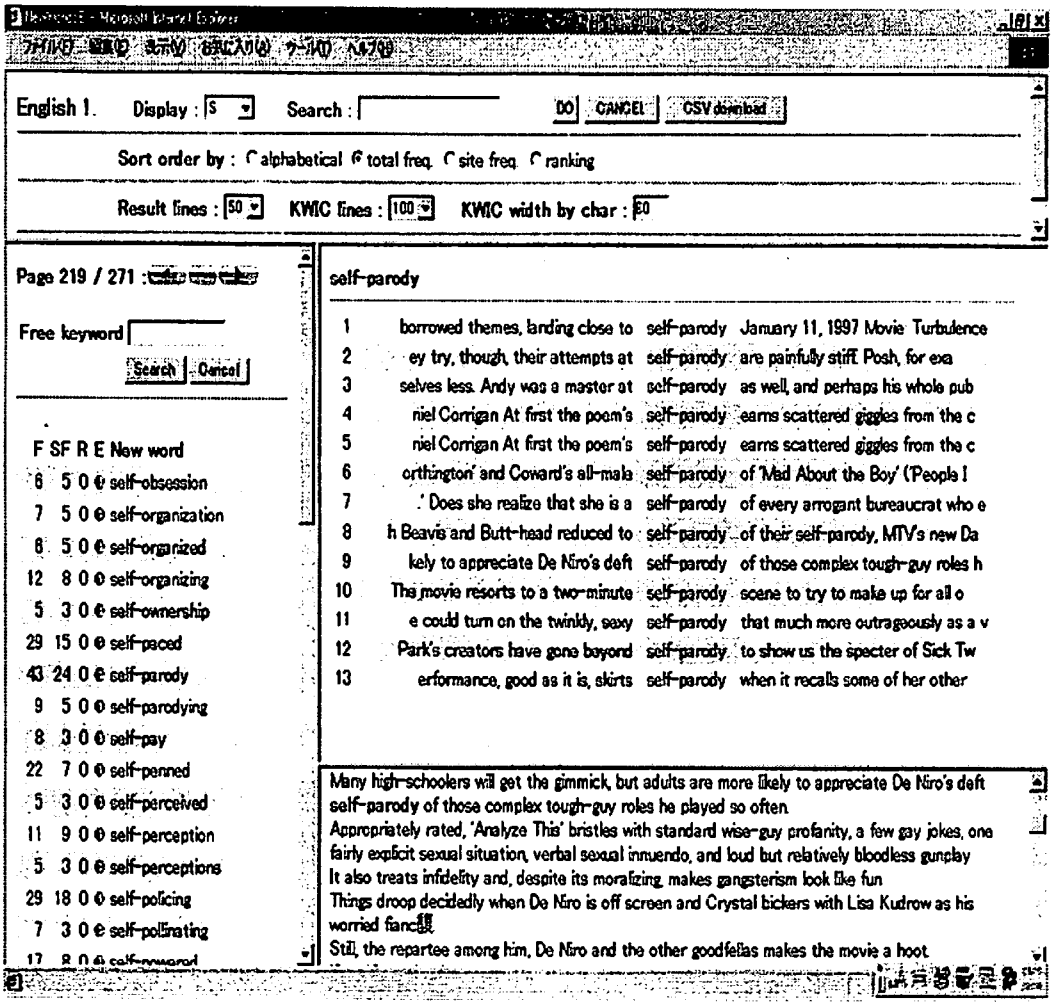
2.2.1 Two versions

The Lexicographer's Workbench Server (LWS) has been developed in order to facilitate lexicographers' process of checking and confirming the words in the FCW database as new words. It has two versions: an SQL server version and a web version (under development). The web version is for lexicographers to use when working independently away from the company, while the SQL version is for in-house work. The web version will be further developed so that in future we can hire native-speaker lexicographers working from abroad.

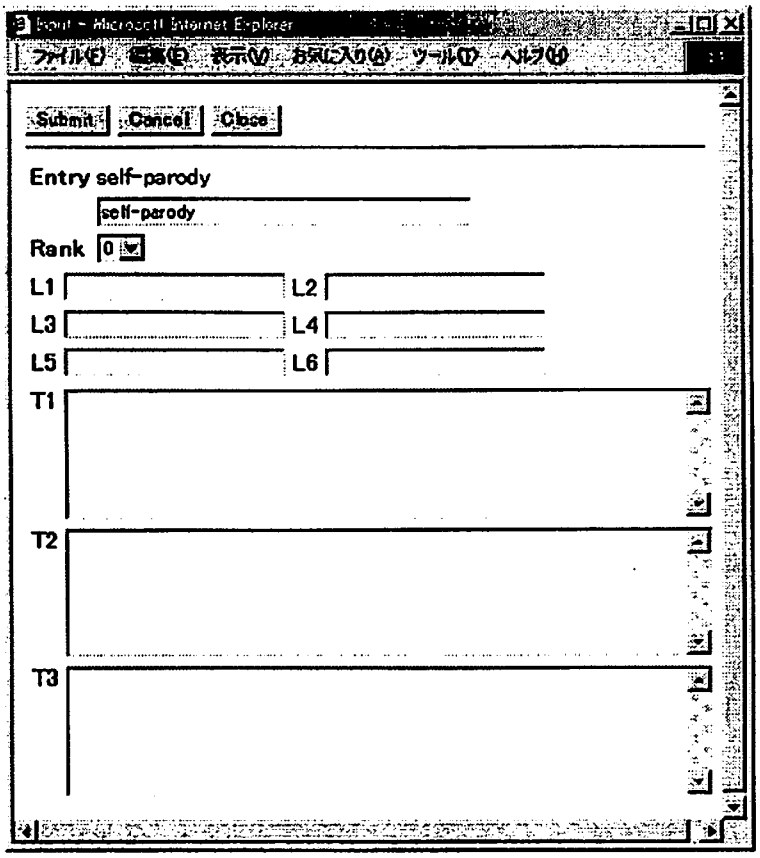
2.2.2 Editing viewer

The viewer for this editing process consists of the list from the FCW database in the left and KWIC lines and their contexts in each file in the right of the window (see Figure 1). The editing window is provided for inserting comments or translation equivalents (see Figure 2). Lexicographers'

comments and translation equivalents will be automatically stored in the database. This viewer can also output a list of the finalised or 'reviewed' words along with the lexicographers' notes into a CSV file.



<Figure 1> The screenshot of the client editing system



<Figure 2> The editing window for each new-word candidate

3. Evaluation and prospectus

3.1 Statistics

Table 1 shows the data on selected new words using our system. The number of HTML files collected by robots will remain almost the same if the target sites are fixed. The Finalised Candidate Words (FCW) and the words in the Reviewed Neologisms Database (RND) (c) are in inverse proportion. The increase of the RND leads to fewer FCW. After a certain amount of time, however, both lists will hit a ceiling and not increase

significantly. At this point, we would need to sort out new extraction patterns or try new websites for more hits. In the case of English words, we have revised search patterns twice before (see, for instance, the data for 12/2000) and have already observed the expected fall in the number of candidate words. Thus, we will consider using new search patterns, involving verb phrases and prepositional phrases.

<Table 1> The processed data by the neologism search system

Year	Month	FCW	Selected new words	RND	Percentages of new words
1998	10	208580	7194	7194	3.4%
1999	3	287574	12674	19868	4.4%
1999	5	171132	13192	33060	7.7%
1999	7	64421	6907	39967	10.7%[2)]
1999	8	45248	5550	45517	12.2%
1999	12	96880	3922	49439	4%
2000	5	3236	224	49663	6.9%
2000	6	5342	371	50034	6.9%
2000	8	3293	222	50256	6.7%
2000	12	16465	1100	51356	6.9%
Total		902171	51356	51356	5.7%

Notes: FCW = finalised candidate words; RND = reviewed neologisms database

The ratio of words which lexicographers selected as new words to the number of the Finalised Candidate Words (FCW) in the viewer is about 1:20 to 1:10. This means that out of 20 lines on the viewer, lexicographers actually confirmed about one or two new words. We have the impression that this ratio is high enough to be encouraging for lexicographers using

2) The change of native speaker staffs caused this increase of the selected new words.

the system.

3.2 Rate of production of new-word entries and future applications

It is estimated that the number of entries covered with this system by four or five lexicographers working seven hours a day for one month would be approximately 3,000 to 4,000 words. The lexicographers are supposed to produce a simple definition with gloss, POS information, illustrative examples, citation sources and comments. They are also preparing the database for the forthcoming web-based new-word information services, in which 40-50 words across different genres are newly written or rewritten every week to supply target users or learners with fresh vocabulary in the field of politics, economics, and sports on a regular basis. Our market researchers are confident that there is a genuine and sizeable demand for such services. Shogakukan will have a team of native speakers who will choose 150 - 200 words from the Finalised Candidate Words described above and write a description of each word, which is then translated into Japanese.

Approximately 50,000 new words have been identified since the implementation of this system at the end of 1998. Almost 900,000 words have been checked on the editing viewer system and the total amount of HTML data accumulated by the robots so far is 70 GB.

The finalised new words are exported as a CSV format into an Excel file and presented to the Dictionary Publishing Division. The information will be used for the production of revised editions of the *Pocket Progressive English-Japanese Dictionary*, the *Random House English Dictionary*, and the *Shogakukan Dictionary of New Words*. We are also planning to provide services for on-line dictionaries and mobile phones. The updated information

on new words will be one of the key factors to attract those mobile phone and internet users' attention. We hope that this kind of new-word search system will hopefully contribute to the automatization and streamlining of the entire dictionary-publishing process.

References

Baayen, R. Harald & Renouf, Antoinette (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language* 72(1).

Renouf, Antoinette (1993). A word in time: First findings from the investigation of dynamic text. In: Jan Aarts, Pieter de Haan and Nelleke Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation.* Amsterdam: Rodopi (Language & Computers #10), pp. 279-288.