Multi-Functional Software for Electronic Dictionaries

Hiroaki SATO, Senshu University, Kanagawa, Japan

Abstract

I have been developing a computer program named RebLin. RebLin has special functions other than searching electronic dictionaries. In this paper, I will show how multi-functional RebLin expands the possibility of electronic dictionaries. First, I will compare RebLin with other computer programs which handle electronic dictionaries, analyzing how each of the software deal with inflected forms and derivatives. Then, I will explain how RebLin retrieves useful information out of other electronic resources such as those on the internet, digitized movies and a lexical database. Finally, I will analyze a log file which records words which users input in the search field of RebLin, and show how they make use of the various electronic resources which RebLin provides them with.

1 Towards the Ideal Dictionary

[FILLMORE/ATKINS 95] describe the ideal dictionary as follows:

> The dictionary must be, in principle, capable of allowing its users access to *all* the information that speakers possess about the words in their language.

Their statement is, in a sense, self-evident. The problem is that compiling such a dictionary is extremely difficult and must take many years.

What I will describe in this paper is a temporary alternative approach to the ideal dictionary with the web-based computer program RebLin. Instead of trying to compile its own dictionary, RebLin makes the best of several electronic dictionaries and other materials on the web. Although RebLin uses some proprietary resources, half of the dictionaries in RebLin are freely available on the web. Since RebLin is a web-based program, users can access the free materials through the internet without installing them in the local computer.

RebLin is a "glue" computer program in the sense that it puts together several electronic resources in a simple interface. Since RebLin uses the basic functions of web browsers, any users who are familiar with them will find no difficulties in using RebLin.

2 Electronic Dictionaries in RebLin

RebLin uses six electronic dictionaries.

1. (a) Kenkyusha's English-Japanese Dictionary for the General Reader
   (b) An Encyclopedic Supplement to the Dictionary for the General Reader
   (c) Taishukan's Genius English-Japanese and Japanese-English Dictionary

(d) The Exceed English-Japanese Dictionary

(e) The American Heritage Dictionary

(f) Wordsmyth, the Educational Dictionary-Thesaurus

(1a-c) are proprietary dictionaries. I bought twenty-four licenses for each of the three bilingual dictionaries in order for students and teachers to use them legally in a computer room at Senshu University. (1d-f) are also propriety materials, but freely accessible on the web. RebLin puts (1d-f) together in the single interface in Figure 1, using JavaScript 1.2 of Netscape Navigator. In a publicly available part of RebLin, users can access (1d-f) seamlessly in this interface (http://www.senshu-u.ac.jp /~thc0408/reblinlite/rebLinLite02.html).

When they click on an English word which they want to look up, RebLin automatically copies and pastes it into the search field of the electronic dictionary, and executes the search.

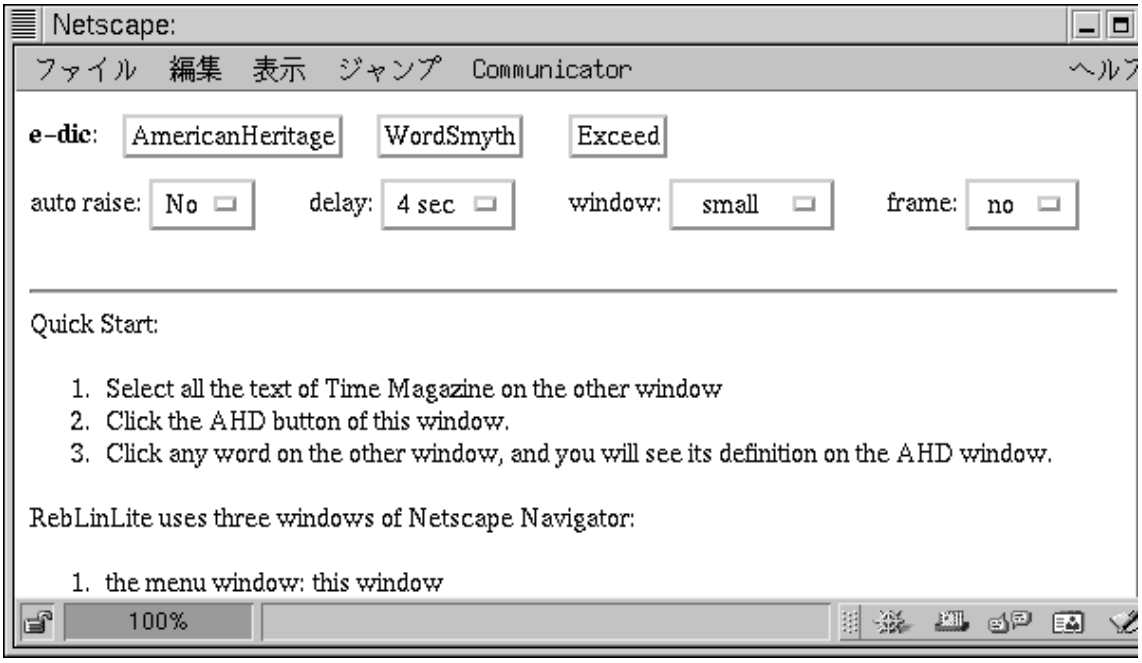

Figure 1: The interface for the dictionaries in (1d-f)

3 Processing Inflected and Derived Forms

There is a popular technique to consult the electronic dictionary. When reading electronic text on the computer, many users copy a word from the text and paste it to the search field of the electronic dictionary, instead of typing in the word from the keyboard. Good electronic dictionaries are supposed to have internal functions to change inflected and derived forms into basic ones so that users do not have to delete affixes manually, when they use this technique.

The Exceed English-Japanese Dictionary and Wordsmyth are among the poorest in this respect. Their internal search programs can process neither the verb suffix *ing* as in *redesigning* nor the noun plural suffix *s* as in *openers*; they just say something like “Sorry, word not found.”

The American Heritage Dictionary is fairly good at finding basic forms from inflected or derived ones, properly processing *redesigning*, *predicted*, *openers*, *constantly*, *longest*, and *thirstily*.

RebLin has its own function to refer to the proprietary dictionaries in (1a-c). Its stemming function is better than that of the American Heritage Dictionary. For example, RebLin can retrieve basic forms of the words listed in (2), while the American Heritage Dictionary cannot.

2. anticloning, machine-gunned, embeddedness, uncodified, nonincumbents, resettlers, anonymization, antinuptials, subtypes, intraracial, hypervigilant, tweakers, overardent, nonmatriculants

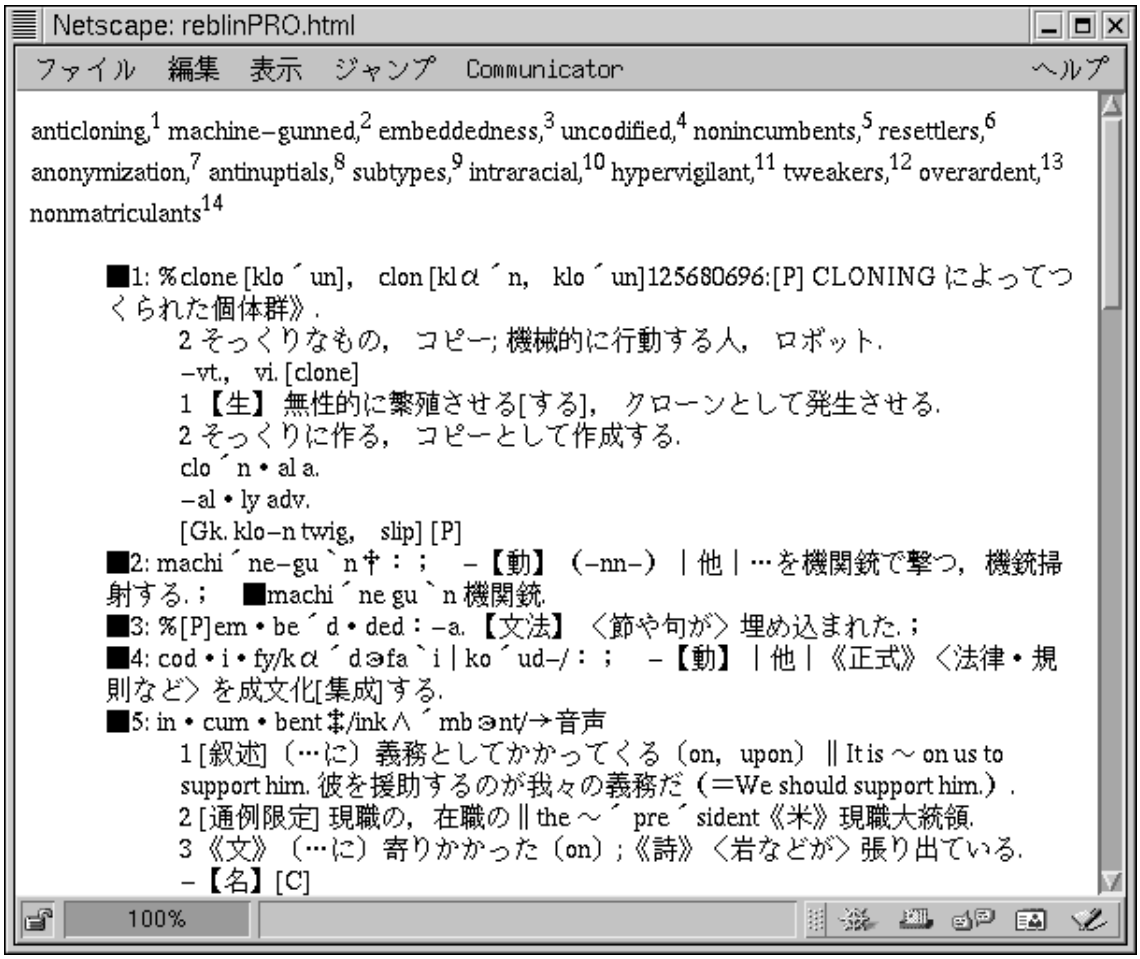

Figure 2: The output of RebLin

All the words in (2) are actually used in the electronic version of *the Time Magazine*. RebLin processes inflected and derived forms at three stages, stripping one affix at a time and checking for a match in the dictionary at each stage. RebLin shows the definition of *nuptial*, while ignoring that of *anti*, when it processes the word *antinuptials*. If users do not know the meaning of *anti*, they will not understand the whole meaning. But this is not a serious problem, because

the number of prefixes and suffixes is limited and advanced learners of English are supposed to know most of them. It must be the case that more users know the meaning of *anti* than that of *nuptial*.

4 Use of the Internet as a Corpus

Data from the internet is helpful when we search for proper nouns or new words which dictionaries do not contain. A lot of web pages also have color pictures with text, which are useful to understand the meaning. For example, one of my former students found several pictures concerning *Trans Am* on the web which clarified the meaning of the following dialogue in an American movie.

3. (a) Woman: How do I look at him?
   (b) Man: Like you just seen your first Trans Am.

In order to understand the latter sentence, one must know what a Trans Am looks like. I realized that one picture on the web was worth a thousand words in the dictionaries. I started incorporating the search function of the internet into RebLin soon after that. When a word is not found in the electronic dictionaries, the user can request a search for the word with the web search program of RebLin
(http://163.136.182.112/koav2.html).

Figure 3: A result of the web search

In Figure 3, RebLin show sentences containing the words with the URL's. The URL's are linked to the original texts so that users can access them by just clicking the links.

The web search of RebLin is useful to see collocation. For example, when users search for words following the expression *you bet your*, they can find several patterns with their frequencies.

order	freq	word
1	36	YOU BET YOUR life
2	8	YOU BET YOUR ass
3	7	YOU BET YOUR sweet
4	2	YOU BET YOUR current

Table 1: the search result of *you bet your*

5 Digitized Movies

Users can listen to the actual pronunciation of a word in some electronic dictionaries and spoken corpora. But there are no dictionaries or corpora in which users can see the scene where a sentence is uttered.

As discussed in [SATO 96], I have been compiling a large audio-visual database for English education since 1989, using 500 American movies. I have paid to a movie company license fees to use seven of them for educational purposes. Students can search for words or phrases in the seven movies, and see the scenes in which they are uttered.

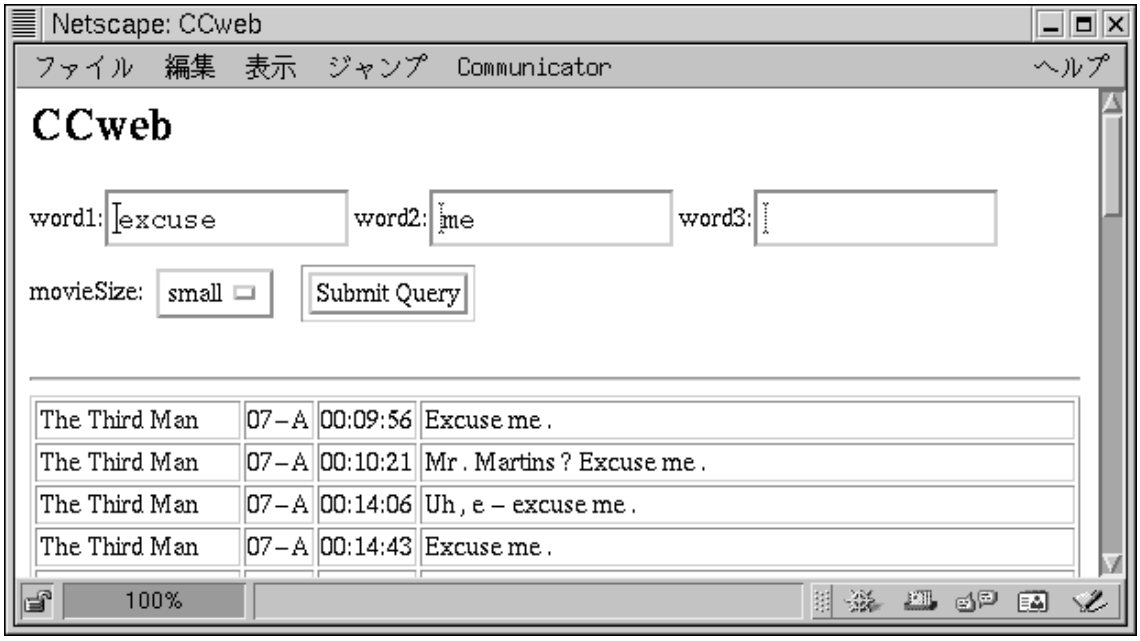

Figure 4: A search result of the movie database

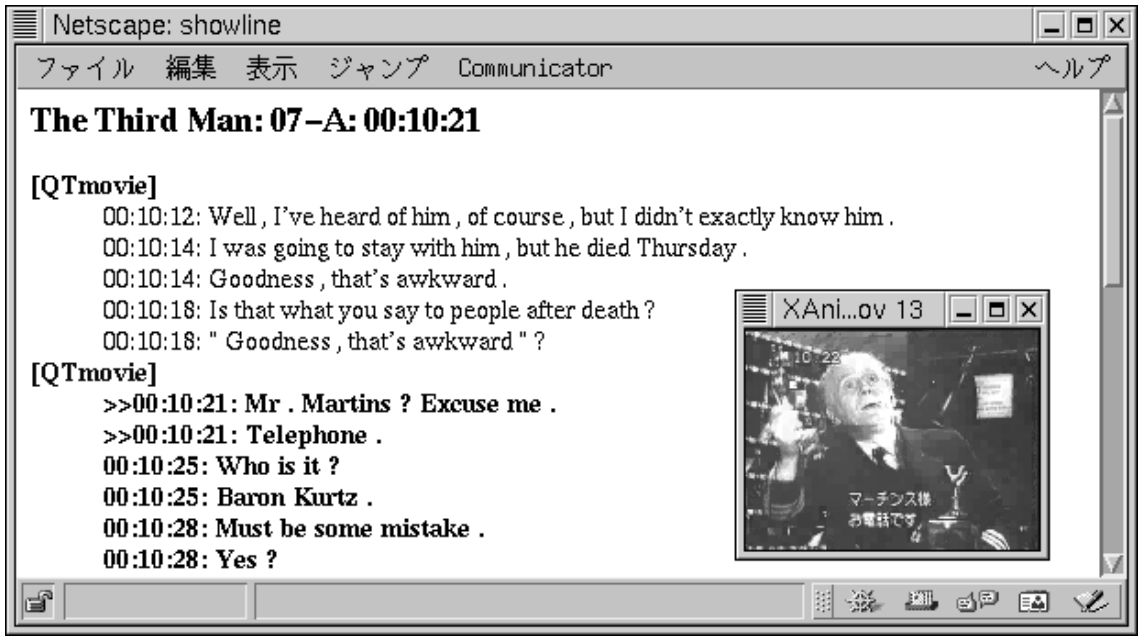

Figure 5: A digitized movie with its dialogue

Although most of the 500 movies in my personal database are new, the seven movies on the web are very old. Since the license fees of new movies are very expensive, I had to start with using the old movies for the movie database on the web. Some parts of the seven movies are accessible from outside Senshu University, and users can see several scenes with their dialogues on the web. The problem is that since the data size of the digitized movie files are large, those who use the slow modem have to wait several minutes to download the files.

The digital movies interact with the electronic dictionaries in RebLin. All the words in the movies are automatically linked to the dictionaries. Users can access the electronic dictionaries by just clicking on a word in a dialogue in the movies.

6 FrameNet

As [ATKINS 95] points out, complementation patterns are a typical component of a traditional dictionary entry. Since verbs have a rich variety of patterns, many dictionaries take a lot of space to show them, but such listings are usually not comprehensive. Though [LEVIN 93] shows to a great extent that the meanings of verbs determine their syntactic behavior, most dictionaries do not make clear the regularity between the meaning of a word and its syntactic patterns.

In 1997, Charles J. Fillmore started compiling an electronic resource named FrameNet (http://www.icsi.berkeley.edu/~framenet/), a database which will in principle provide an exhaustive account of the semantic and syntactic combinatorial properties of each "lexical unit", which corresponds to one word in one of its uses. FrameNet shows not only the syntactic patterns of verbs, but also for nouns and adjectives as well.

The FrameNet database will be freely available as a research resource on the web in May, 2000. I made several computer programs for the FrameNet database while I worked at the International Computer Science Institute in Berkeley, California from April, 1999 to March, 2000. Since these programs are easily merged into RebLin, I plan to incorporate some part of the FrameNet database into RebLin, in order to provide information on complementation patterns. Electronic dictionaries cannot provide this type of information, which is especially useful for English instructors preparing teaching materials.

7 How Users Use RebLin

Since RebLin is web-based software running on a web server, the server program automatically saves users' access records in a log file. For example, it records the IP addresses of users, the dates when they access RebLin, the words that they look up, and the electronic resources that they use.

This log file is invaluable. It records fairly well the users' behavior when consulting dictionaries, a process which is usually difficult to monitor. The log file records the users' process of consulting dictionaries more directly than statistics calculated from questionnaires to users on their dictionary use. The log file also shows how users access various electronic resources. I will talk about some of their behavior in the presentation.

8 Concluding Remarks

I plan to incorporate new materials into RebLin. They will be from my personal audio-visual database which contains 500 American movies, and from proprietary dictionaries in the electronic format.

I have already finished writing the software to handle the new materials on the web. Those materials would be instantly available on RebLin, but I have to pay their license fees first to the copyright holders.

The fees are quite expensive. Luckily, the Ministry of Education of Japan has been supporting my research since 1991, giving me $100,000 for the license fees of the seven movies in 1997, $8,000 for those of the three proprietary dictionaries in (1a-c) in 1998. How large and quickly RebLin grows in the next few years depends mostly on the generosity of the ministry and other funding sources.

Acknowledgement. I would like to thank Dr. Collin Baker of the International Computer Science Institute at Berkeley, California for his comments on an earlier version of this paper. This paper is in part a result of research that I did at that institute from 1999 to 2000.

References

[ATKINS 95] Atkins, Beryl T. (1995) "The Role of the Example in a Frame Semantics Dictionary,"in *Essays in Semantics and Pragmatics: in Honor of Charles J. Fillmore*, ed. by Masayoshi Shibatani, Sandra Thompson. J., 25-42, Benjamins Pub. Co., Amsterdam; Philadelphia.

[LEVIN 93] Beth Levin (1993) *English Verb Classes and Alternations : a Preliminary Investigation,* University of Chicago Press, Chicago.

[FILLMORE/ATKINS 95] Fillmore, Charles J. and Beryl T. Atkins (1998) "FrameNet and Lexicographic Relevance," in *First International Conference on Language Resources and Evaluation, Proceedings,* ed. by Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada, 417-423, Granada, Spain.

[FILLMORE/ATKINS 92] Fillmore, Charles J. and Beryl T. Atkins (1992) "Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors," in *Frame, Fields and Contrasts,* ed. by Adrienne Lehrer and Eva Feder Kittay, 75-101, Lawrence Erlbaum Associates, Mahwah, New Jersey.

[SATO 96] Sato, Hiroaki (1996) "A Multimedia Database of Movie English," in *Multimedia Language Teaching,* ed. by Sandra Fotos, 80-89, Logos International, Tokyo.

Electronic Dictionaries in RebLin

- The American Heritage Dictionary (2000), Lycos, Inc., (http://www.lycos.com/referenceroom/)
- An Encyclopedic Supplement to the Dictionary for the General Reader (1994), Kenkyusha, Japan.
- The Exceed English-Japanese Dictionary (1999), Sanseido, Japan, (http://dictionary.goo.ne.jp/cgi-bin/ej-top.cgi)
- Taishukan's Genius English-Japanese and Japanese-English Dictionary (1998), Taishukan, Japan.
- Kenkyusha's English-Japanese Dictionary for the General Reader (1984), Kenkyusha's, Japan.
- Wordsmyth, the Educational Dictionary-Thesaurus (1999), Wordsmyth Collaboratory, (http://www.wordsmyth.net/searches.shtml)