Development of Multimedia Collocation Information Retrieval System for Learners of Japanese

Sangmok Lee and Shigeru Sato
Graduate School of International Cultural Studies
Tohoku University, Sendai, 980-8576 Japan
{sangmok, satos}@insc.tohoku.ac.jp

1. Introduction

Importance of collocation has been well recognized in the domain of English language education in Japan from early days as shown in an elaborate accumulation of collocation data published in the form of dictionaries[1]. In recent years large electronic corpora of English have been compiled for a variety of uses including educational purposes, which has lead to the emergence of large scale learning dictionaries widely used not only in the form of a book, but also on-line on the website[2]. In contrast to English, the environment for the learners of Japanese lags far behind in terms of the availability of electronic corpus.

This paper adopts a broader definition of collocation and reports on an attempt of developing a multimedia collocation retrieval system for the learners of Japanese as a foreign language, based on the mass supply of multimedia contents currently available on the web. Our definition of multimedia collocation is based on the text data combined with speech and video information recorded synchronous to the text data; namely, collocation in the multimedia context. We argue that this broader definition entails regarding as collocation the entire setting for the use of a particular expression that not only includes lexical and semantic usage but also covers the context where it is actually uttered. This kind of setting is arguably where an expression actually occurs in real-time with audio-visual confirmation of articulation, intonation, facial expression, and other paralinguistic information.

2. Multimedia Collocation in Education

It is obvious and well understood among language teachers that the use of a large scale corpus specifically facilitates learners' effort to understand the collocational meaning of expressions that are more than the combination of each of the consisting words. Definition of collocation in a narrower sense is "a word or phrase which is frequently used with another word or phrase, in a way that sounds correct to people who have spoken the language all their lives, but might not be expected from the meaning"[3]. Learning collocation entails not only acquisition of 'right/wrong' judgment but also of naturalness for a given phrase as well. There is a two-sided difficulty for learning collocation and acquiring naturalness of the native speaker. The native speaker does have intuition for his/her judgment but, as regards systematic description for a given expression, he/she may not be persuasive to the learner[4]. On the other hand, due to the limitation of time it is totally impossible for the learner to be exposed to the entire body of the language data.

In our efforts in making use of the present-day internet multimedia environment in language education we developed an on-demand web-based dictation system for learners of Japanese mainly targeted to Korean students[5]. One of our authors also participated in the project developing discourse teaching material using video-clips to be used in a server-client network setting[6].

3. Construction of a Multimedia Corpus

3.1 Speech-media Corpus

The speech-media corpus we have dealt with is an electronic compilation of Japanese textbooks for use in secondary school education in Korea. Using this speech-media textbook corpus we developed a preliminary collocation information retrieval system based on the text database with corresponding speech.

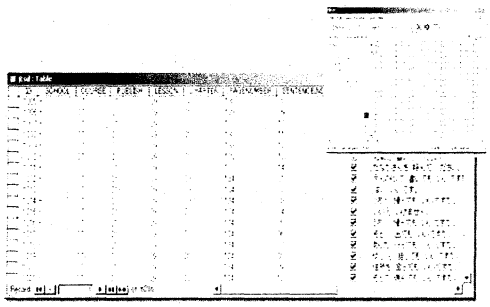

Fig. 1 Speech-Text Database: an Example

Each sentence has an index with information on the publisher, year of publication, chapter, and page. All the data and programs are written in Unicode, providing multilingual environment for clients accessing through the web. Speech files in WAV format are stored in hierarchical structures according to the index information, as exemplified in Fig.1.

3.2 Video-media Corpus with Teletext Subtitles

Teletext subtitling was originally designed to help the hearing-impaired enjoy TV/video programs, where transmission of subtitles utilizes unused lines of the 525 NTSC video frame scan lines to send the subtitle information. A decoder is necessary for subtitles to appear on the TV screen. We used a video capture card to separate the video signal from the subtitle signals, and along the latter we attached a time scale. (See Fig. 2)

The video-media (multimedia) corpus we used in our present experiment is the TV program "*Kōkō Kyōsi* (High School Teacher)", where subtitle information and its synchronous speech-video data were manually captured to compile a corpus for the present experiment. In order to locate a particular collocation expression, manual intake, using the attached time scale, visually locates the needed portion watching the closed subtitle data.

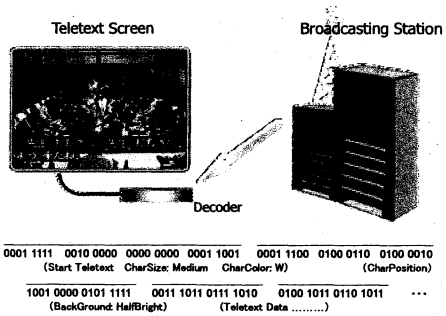

Fig. 2. Teletext Data in TV Broadcasting

4. Collocation Retrieval from Multimedia Database

According to the procedures in **3.2**, we developed a prototype of multimedia collocation information retrieval system with a limited vocabulary but with the following potential advantages for learners and teachers of Japanese:

(1) Presentation of collocation is available in the speech-video context.

(2) Access to multimedia collocation data activates learners' memory and facilitates learning process.

(3) Relational database employed to renew and maintain the data is easy to use and reliable for the teacher and/or system administrator.

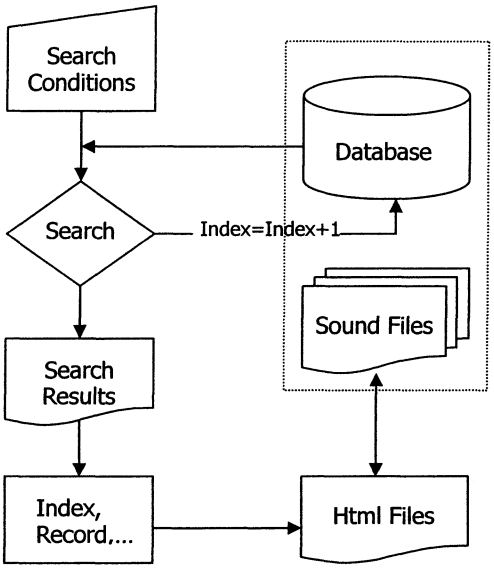

Fig. 3. Collocation Search Text-Speech System

4.1 Collocation Retrieval from Speech/Text

Figure 3 is a flow diagram of a system of text/speech collocation retrieval referring to speech files for their replay. In response to an input of search conditions by the learner the system returns an HTML file with reference to a corresponding speech wave file.

Text collocation emerges in the co-occurrence of words that are either physically adjacent to one another or placed discontinuously in a phrase/sentence. In order to deal with the discontinuity encountered in learning Japanese, wild cards are defined as in Fig.4.

ほら，*ほら*，韓国|かんこく，*から*まで*，*だろう，
*は*です，仮に*も，いかにも*ように

Fig.4. Wild Cards for Retrieval

4.2 Retrieval of Multimedia/Text Collocation

Figure 5 shows a user interface in a video/speech collocation retrieval session. Figure 6 is a diagram of the multimedia collocation retrieval system in a server-client environment, where the server administers the database, retrieves collocation information, and

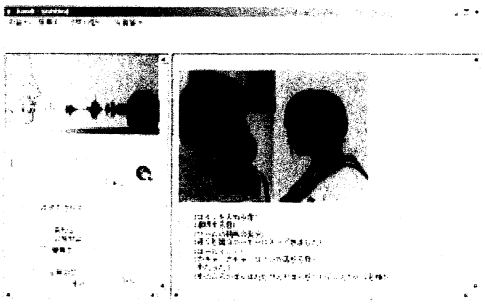

Fig. 5. User Interface in Collocation Retrieval

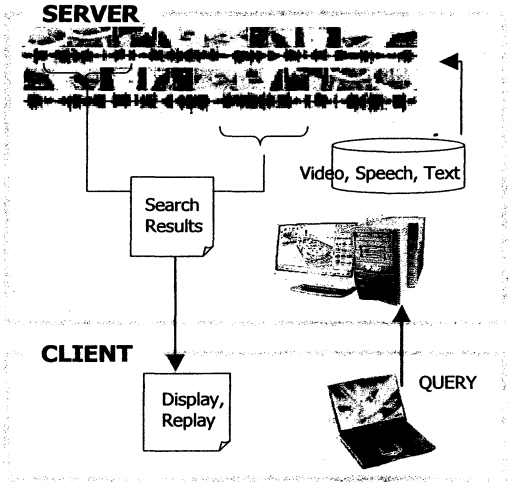

Fig. 6. Multimedia Collocation Retrieval System

analyses the results. The relational database comprises video with subtitles, and supplies data at the request from the client. The client, on the other hand, makes collocation query, displays retrieval and analysis results, and replays video and speech files.

5. Conclusion

The present paper started with the extended concept of multimedia collocation, and reported on the development of collocation information retrieval system for learners of Japanese in an on-demand real-time server-client environment. We are currently developing methods to evaluate the efficiency of such learning environment for classroom use by students of Japanese in Korean high schools.

Acknowledgments

This study was supported in part by the 21st Century Center of Excellence (COE) Program entitled "A Strategic Research and Education Center for an Integrated Approach to Language and Cognition", Tohoku University. Our thanks go to Kaoru Horie for his discussion with us on multimedia language education.

References

[1] Ichikawa, S. et al.: *The Kenkyusha Dictionary of English Collocations*, Tokyo: Kenkyusha, CD-ROM (1996)

[2] http://titania.cobuild.collins.co.uk/

[3] http://dictionary.cambridge.org/

[4] Granger, Sylviane: *Learner English on Computer*, London/New York: Longman (1998)

[5] Lee, Sangmok et al.: Development of Dictation System on the Web for Learners of Japanese, *Proc. 7th Annual Meeting of ANLP*, pp.441-444 (2001)

[6] Yasui, Akemi et al.: Development of a Web-based Multimedia Teaching Material and its Evaluation, *Proc. Autumn Meeting of the Society for Teaching Japanese as aForeign Language*, pp.109-114 (2001)