論文 / 著書情報
Article / Book Information

Title(English)	Japanese Composition Support System Displaying Co-occurrences and Example Sentences
Authors(English)	YOSHIHASHI Kenji, NISHINA Kikuko
Citation(English)	Proc. of the International Symposium on Large-Scale Knowledge Resources, Vol. , No. , pp. 119-122
発行日 / Pub. date	2007, 3

Japanese Composition Support System Displaying Co-occurrences and Example Sentences

NISHINA Kikuko, YOSHIHASHI Kenji***

**International Student Center, **Graduate School of Decision Science and Technology*

Tokyo Institute of Technology
knishina@ryu.titech.ac.jp

Abstract

This paper describes the development of a Japanese composition support system called Natsume, which, based on the idea of data-driven learning, provides learners with plenty of authentic example phrases and sentences when writing Japanese compositions. Natsume is being designed to support intermediate or advanced learners in writing Japanese compositions. The system shares databases and dictionaries with Asunaro—a reading-support system that we have also developed. Natsume is a unique system realized to display appropriate example sentences selected based on co-occurrence frequencies, genre, and learner proficiency level. This paper explains how example sentences are filtered from a copyright-free corpus including literary works and newspaper articles and displayed on the web interface.

Index Terms;Natsume, corpus, data-driven learning, proficiency level, Aozora Bunko, co-occurrences

1 . Introduction

Studies of data-driven learning suggest that showering language learners with authentic examp le sentences is an effective method of helping them to acquire vocabulary and master grammatical patterns [1][2]. Drawing on this notion, we are developing a system to present learners with large numbers of examples taking into consideration both important pedagogical and technical factors. The system is realized by sharing data with our existing Asunaro system [3], which is a reading-support system allowing learners to read Japanese written texts by presenting kana readings, meanings and syntactic struct ures.

From a learner's perspective, it is important that they encounter as much authentic data as possible so that they can recognize exact word meanings and natural usages of the target language themselves. Accordingly, our system seeks to provide learners with examples taken from a wide range of written documents, such as novels, essays, personal letters and academic reports and papers. It is also important for example sentences to be appropriate in terms of the learner's level of proficiency, with basic vocabulary items and sentence structures that are not too difficult. However, from a technical point of view, it is rather difficult to extract large quantities of appropriate example sentences, because a sentence may be open to multiple interpretations, making it hard to determine the exact meaning if it is taken independently of its context. This has serious implications in developing the system to provide the most appropriate example sentences.

Given these constraints, we propose a method of constructing the system so that it can present plenty of examples to learners, but which are sufficiently simple structurally. First, we analyze the authentic corpora that have been collected so far, as described in more detail in Section 3, using existing morphological and syntactic analyzers in order to construct a database of example sentences. Then, utilizing the analysis results, the system computes the levels of the vocabulary, the levels of the grammatical patterns, and the complexity of structures in the database. Finally, we annotate the database with level markers. Users of the system can then extract examples that have been filtered according to user criteria, such as their specialty and native language, when sentences containing target words are retrieved. In this paper, we focus on the initial stages of constructing the composition support system.

2. Overview of the Natsume System

2.1. Overview of the System

We start with an overview of the structure of the Natsume system. Like the Asunaro system, Natsume utilizes the EDR dictionary, but it also includes a database of co-occurrences, a corpus of learner writing errors, and example sentences appropriate for every learner level. As Fig.1 overleaf shows, the Natsume system has also been designed to draw on specialist vocabulary lists to support system users, although this paper will limit itself to discussing only the co-occurrences and example sentences. The inclusion of the learner error corpus reflects the importance of helping learners to become aware of common errors due to L1 knowledge transfer interference. A study of learner errors has shown that learners tend to make certain kinds of errors depending on their L1 [4].

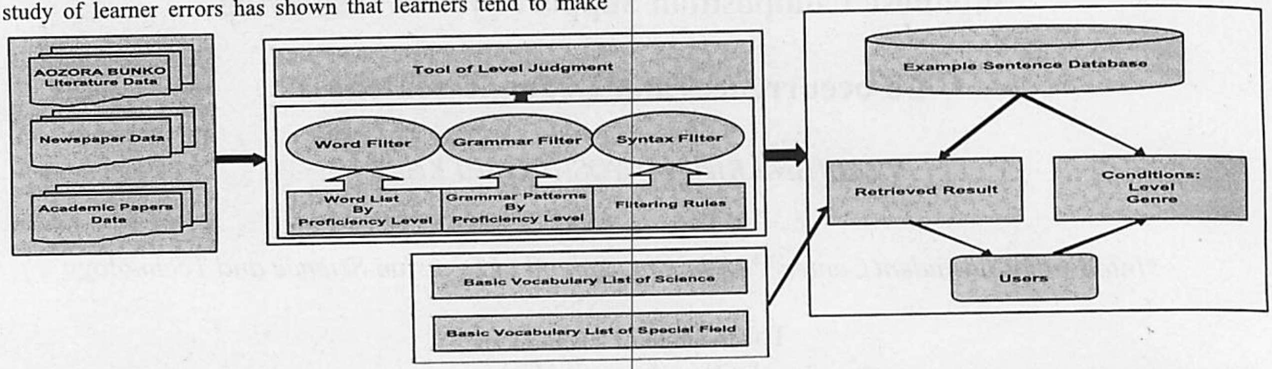

AOZORA BUNKO Literature Data
Newspaper Data
Academic Papers Data
Tool of Level Judgment
Word Filter
Grammar Filter
Syntax Filter
Word List By Proficiency Level
Grammar Patterns By Proficiency Level
Filtering Rules
Basic Vocabulary List of Science
Basic Vocabulary List of Special Field
Example Sentence Database
Retrieved Result
Conditions: Level Genre
Users

Figure 1 *Overview of Natume system*

The present system is designed to incorporate learner error data from compositions composed by Chinese learners. The system can provide feedback hints to the user on the system interface. A detailed explanation of the error analysis component of the system will appear in another paper

2.2. Interface

Fig. 2 is enlarged section from a screen shot of the Natsume interface, which we are currently developing. On entering the system, a learner is prompted to enter his or her level of Japanese proficiency, native language, and the genre of composition, so that system can retrieve and display appropriate co-occurrence words and example sentences. The learner can enter any noun that he or she wishes to use in composing a Japanese sentence, and does not even need to specify a predicate.

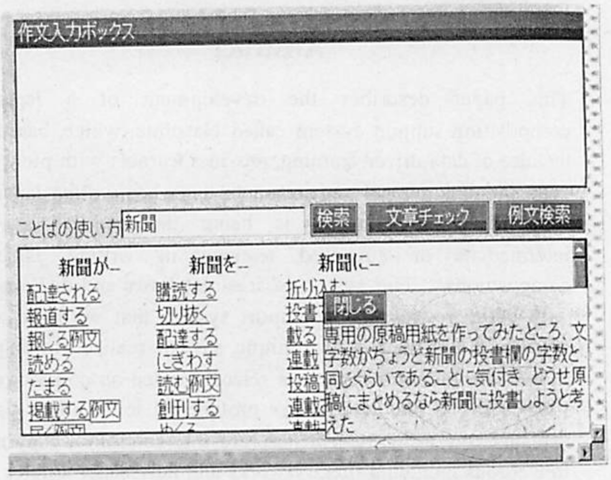
2.3. Displaying co-occurrence information

The system then presents candidate translation words together with frequently co-occurring particles and verbs in the lower frame of the interface. The system can also display the results of syntactic dependency structure analysis for a sentence, based on morphological analysis of a sentence with Cabocha. The system contains a database of dependency structure results for nouns and verbs created by analyzing 10 years of articles from the Mainichi newspaper with the syntactic analyzer Cabocha.

The following is an example of this syntactic analysis for *'shinbun ga'*, *'jiken wo'* and *'houjita'* as the core structure elements.

(shinbun ga(kinoo/ okotta) jiken wo) houjita
Shinbun ga kinoo machi de okotta jiken wo houjita.
(The newspaper reported on an incident that took place in the town yesterday)

The analyzed data is stored with Mutual Information (MI) scores, which indicate the frequency with which the two words appear together in a corpus. Verb and noun pairs within our data are computed and displayed as co-occurrence words [5] [6]. While the present system only retrieves nouns, future versions will be able to present learners with both co-concurrent verbs and nouns. For example, when a user inputs the noun *'shinbun'*, several verbs will be displayed according to the relevant case particles of *'ga'*, *'wo'* and *'ni'*.



ことばの使い方 新聞

新聞が–
配達される
報道する
報じる 例文
読める
たまる
掲載する 例文

新聞を–
購読する
切り抜く
配達する
にぎわす
読む 例文
創刊する

新聞に–
折り込む
投書
載る
連載
投稿
連載

専用の原稿用紙を作ってみたところ、文字数がちょうど新聞の投書欄の字数と同じくらいであることに気付き、どうせ原稿にまとめるなら新聞に投書しようと考えた

Figure 2 *Enlarged section of the Natsume interface*

Shinbun ga: haitatsu sareru, houdousuru, houjiru, yomeru
wo: koodokusuru, kirinuku
ni: orikomu, toushosuru

If the learner then clicks on a particular verb, the English meaning appears in a pop-up box. Moreover, several example sentences are presented when the user clicks on a reibun (example sentences) icon.

3. Corpus of Example Sentences

3.1 Schema for Display of Example Sentences

As just noted, reibun (example sentences) icons are attached to co-concurrent words within the frame. This section explains about the schema for retrieving example sentences. As the system is being designed to display appropriate sentences according to learner proficiency levels, we first classify the data by filtering out difficult words. Table 1 shows the structure of the database.

Table1 *Example sentences according to level*

1st line: Indicates level of each word
2nd line: Text title and sentence ID
3rd line: Sentence text
4th line: Sentence level and word counts for each level
(as a six-digit string)
Number 1: Level
Number 2: Counts for out of level words
(Only for nouns, adjectives and verbs defined by the CaboCha grammar)
Number 3: Count for level 1 words
Number 4: Count for level 2 words
Number 5: Count for level 3 words
Number 6: Count for level 4 words

Level 1 refers to advanced level, Levels 2 and 3 to intermediate, Level 4 to beginner learner [8]. Example sentences determined as Level 1 and Level 4 are shown below.

Level 4 example
# sensei:4 doko:4 iru:4 shiru:4
# kokoro-105
Shikashi sensei ga doko ni iru ka wa mada shiranakatta.
(But, I still did not know where the professor was)
4 0 0 0 0 4

Level 1 example
# koko:4 tada:2 sensei:4 kaku:4 honmyoo:1 uchiakeru:1
# kokoro-2
Koko demo tada senesi to kakudake de honmyou wa uchiake nai.
1 0 2 1 0 3

3.2 Usable corpus and level classifications

We have obtained the following Japanese corpora.

1) Aozora Bunko (Aozora Library).

A collection of literary works (novels, essays, and personal letters) that were written more than 50 years ago.

2) Newspaper articles

2-1 Mainichi Shinbun; 2-2 Asahi Shinbun; 2-3 Shinano Mainichi Shinbun (editorial columns and essays).

3) Academic papers

While we are employing a large quantity of newspaper articles for statistic purposes, they are not used for example sentences on the web due to copyright reasons. However, the Aozora Library collection is copyright-free, and while this material can be rather difficult and old-fashioned for Japanese learners, we can extract useable examples by filtering out difficult vocabulary and grammatical patterns from the data. Our procedure for selecting appropriate sentences and constructing the database for the Aozora Library collection has been as follows. Although the Aozora Library works include both Shin-kana texts (modern orthography texts) and Kyuu-kana texts (classic orthography texts), we collected 4,023 works after excluding classic orthography texts because it is not supported by current morphological analyzers. Table 2 presents data about the example sentences extracted from the Aozora Library. The four levels in the table are learner proficiency levels. The Japanese Language Proficiency Standard specifies vocabulary and grammatical patterns for all levels.

Table2 *Vocabulary, grammatical patterns, and Aozora Library sentences according to each proficiency level*

Level	Vocabulary (words)	Grammar (items)	Aozora Library sentences
1	10,000	402	19,658
2	6,000	303	45,736
3	2,000	132	13,366
4	800	50	11,686

Koko demo tada sensei to kakudake de honmyou wa uchiake nai.
koko:4 tada:2 sensei:4 kaku:4 honmyoo:1 uchiakeru:1
# kokoro-2

As there are two level 1 words, one level 2 words, no level 3 words, and three level 4 words, the sentence is marked with the six digit string 1 0 2 1 0 3, because the whole sentence is regarded as a level 1 sentence based on the highest level among the constituent words.

While we also take into consideration grammatical factors when classifying the levels of example sentences, such factors are not discussed in this paper. For example, although the example sentence,

'Shikashi sensei ga doko ni iru ka wa mada shiranakatta',

consists only of level 4 sentences, such as *'sensei', 'doko', 'mada', 'shiru', 'ka'*, the grammatical patterns of *"doko~ <predicate> ka* and *wa <predicate> nai* are level 3 patterns. Stricter controls will be possible after further synthesis of the vocabulary and grammar items.

4. Evaluation study

We have conducted an evaluation study for the system.

4.1 Method and procedure

Participants: 19 participants, ranging from proficiency levels 1 to 3 based on either JASSO tests results or self-evaluations (13 from Chinese character regions and 6 from other areas).

Materials: 10 different 'Kobo-chan' comic strip stories. We selected the 4frame comic strips because they have simple stories and because they are unlikely to prompt arbitrary responses from the participants.

Procedure:

1) Three kinds of materials were sent as e-mail attachments.

2) The participants read the comic strips and Japanese explanations of the stories. The story explanation sheet focused on 18 presented nouns with some blanks that were filled with their corresponding verbs. The 18 presented nouns were: deeto (date), mushi (insects), karada (body), shirushi (mark), himo

(string), tebukuro (gloves), remon no kaori (lemon fragrance), monosashi (ruler), omen (mask), koshi (hips), ashi (legs), booru (ball).

3) The p articipants filled in blanks with appropriate particles and predicate verbs based on the stories.

4) As multiple responses were acceptable, we judged a response as being correct if it was a reasonable answer and was grammatically correct.

5) We prepared a questionnaire of 10 items, asking the participants to evaluate the system on a 5-point scale after the test stage.

4.2 The results of the word filling test

Generally, the participants selected correct predicate verbs for the 18 presented nouns from among the candidates presented by Natsume. *'booru'* (ball)' had the lowest rate of correct answers. Although appropriate answers would be *"wo nageru"* (to throw) or *"wo nagekaesu"* (to throw back) after *"booru"* based on the comic strip situation, the participants gave various answers such as *"wo naguru"* (to hit someone) or *"wo sakarau"* (to move against). The system only displays simple verbs, but users will sometimes need complicated predicates involving complex verbs. In some cases, even though the participants selected appropriate verbs, they failed to construct grammatically sound sentences, making syntactic errors relating to voice, tense and aspect.

These results suggest that while Natsume can be very helpful in selecting simple co-occurrence verbs, it is less useful for complex verbs and difficult grammatical features. We will therefore explore methods to improve the system, such as classifying the relationships between case particles and predicates according to voice.

4.3 Evaluation results

As shown in Figure 3, the evaluation scores indicate that participants from regions that do not use Chinese characters were generally more positive about the system than participants from Chinese character areas. On the other hand, earners from Chinese character areas expressed more desire to use the system in the future. We may speculate that the learners from Chinese character areas would seem to be less interested in the system, because they feel they can grasp general meanings from their knowledge for Sino-Japanese words, but it is precisely this false sense of familiarity for Japanese words that leads them to make frequent grammatical errors concerning particles, voice, tense and aspect when they write documents in Japanese.

5. Conclusions

This paper has discussed the wide range of Japanese language corpora used in the construction of the Natsume composition-support system. Because the corpora that we can freely use for the system is somewhat restricted due to copyrights and learner proficiency levels, we have fully exploited non-restricted corpora by filtering and other processes to construct a database of usable texts. This database is a vital component of the Natsume system providing learners with an abundant source of authentic sentence examples. In order to examine the usefulness of the system, we conducted an evaluation study in which participants made selections from candidates displayed by the system. The study showed that the system was generally highly rated by the participants, particularly those from non-Chinese character area. The participants realized that Natsume is a useful tool not only for reviewing and preparing for Japanese language classes but also for writing papers and e-mails. However, there is clearly still room for improving the system interface. We also plan to incorporate grammatical information for proficiency levels in order to further refine the retrieval of appropriate example sentences. Moreover, we will explore ways of providing appropriate expressions according to different genre criteria, which learners want to study.

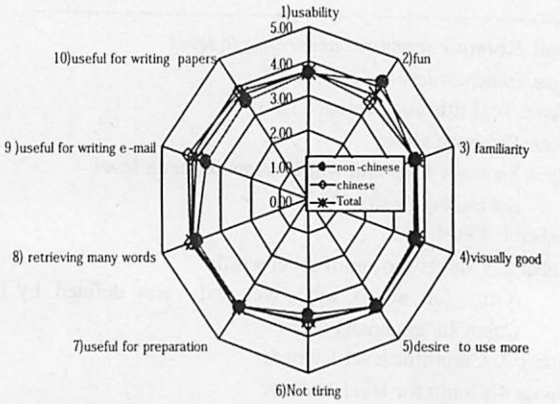


Figure 3 *Average participants evaluations*

6. References

[1] C. Tribble, G. Jones Concordances in the Classroom, Athelstan, 1997

[2] Granger, S., Learner English on Computer, Longman, 1998.

[3] Nishina Kikuko, Development of Multimedia Contents for e-Learning–Asunaro System Extension Symposium on Large-Scale Knowledge resources (LKR2005), pp.83-86, 2005

[4] Cao,H., Nishina,K.,and Chinese learners' acquisition of Japanese adjectival collocations: A composition error analysis and pedagogical implications, Nihongo Kyooiku pp.70-79, 2006

[5] Totsugi Norihisa, Development of CALL System for Japanese Composition Symposium on Large-Scale Knowledge resources (LKR2005), pp. 215-218, 2005.

[6] Church,K. Hanks, P. Word Association noun, mutual information and lexicography, Computational Linguistics 16: pp.22-29,1990

[7] The Japan Foundation and Association of International Education, Japanese Language Proficiency Test: Test Content specifications Vocabulary and Sentence Pattern Lists for Japanese Proficiency Test, Japan Foundation, 1994