Consideration of Language Learning Service with Visualized Vocabulary Map Derived from WordNet

Masatsugu Ono
Faculty of Engineering
Muroran Institute of Technology
Muroran, Japan
onomasa@mmm.muroran-it.ac.jp

Toshioki Soga
Faculty of Science and Technology
Chitose Institute of Science and Technology
Chitose, Japan
t-soga@photon.chitose.ac.jp

Masato Kikuchi
Faculty of Engineering
Nagoya Institute of Technology
Nagoya, Japan
kikuchi.masato@nitech.ac.jp

Tetsu Tanabe
Information Initiative Center
Hokkaido University
Sapporo, Japan
ttanabe@iic.hokudai.ac.jp

***Abstract*— The authors aim to propose highly effective, efficient and satisfactory service of language teaching and learning by means of educational technology from the viewpoint of usability engineering. In the field of computational/quantitative linguistics based on corpora data, many linguists have been numerically analyzing lexical characteristics for each English vocabulary and compiled them as a vocabulary list which is machine-readable since 1960s. On the other hand, a lexical database of semantic relations called WordNet has also been developed in the field of computing science. This study consists of three components. One is to enrich, refine and integrate already-existing vocabulary information in terms of "frequency", "difficulty", and "relativeness". Another is to develop the system of language learning and teaching service with the integrated vocabulary information. And the other is to evaluate the system from usability engineering point of view. So far, the authors reached and conducted the first two kinds of studies, succeeded in deriving the newly-integrated vocabulary list and visualizing the word relations in each experiment.**

Keywords—vocabulary lists, visualization, conceptual structure map, language learning service, user experience

I. INTRODUCTION

The purpose of this study is to improve present activities and situations of English language education for both learners and teachers. And the goal is to suggest taking a new strategy of vocabulary learning from the viewpoint of service science. The authors aim to propose highly effective, efficient and satisfactory service of English vocabulary teaching and learning by means of educational technology. They regard a series of actions on language education as a "service" as well as regard both learners and teachers as "users" in the field of usability engineering. They will evaluate the educational contents and services that they developed, based on user experience research.

In this study, they attempted to generate vocabulary maps in order to show English learners difficulty, frequency and correlation of each word. This paper describes and focuses on how to visualize such information of lexical data of English vocabulary.

II. BACKGROUND

A. English Language Education in Japan

A generation ago, students in Japan started learning English when they became junior high school students. But present students in Japan start learning it when they get to the 3rd grade in elementary school after the year 2020. They learn English as a first foreign language, not English as a second language. It means that they do not have any opportunities to encounter or be involved in real English situations in their daily life in Japan and also means that the period of their learning English at school has just been extended. Normally they have an English instruction from teachers with textbooks in English learning classes. They learn English vocabulary as "new words" in each lesson from textbook, which often makes them difficult to learn through their self-study. For Japanese people, their first language is Japanese which is quite different from English in grammar like word orders, pronunciation, etc. It is still very difficult for most Japanese people to master English due to such language difference.

B. Characteristics and Education of Vocabulary

According to Tamamura[1], vocabulary has its own characteristics. Some of the typical ones are shown as below.

- Many words can be reworded or rephrased.
- The number of words keeps on generating and increasing more.
- The rate and range of one word usage differ from another.
- A sense of language also differs between different languages.

Tamamura also describes that instructors should be familiar with the word set to be taught at the same time as well because words can newly generate and increase as one of the characteristics. It means that words to be taught should be collected and/or classified as synonyms, co-occurred words, derived words, etc.

Sonoda[2] produced a lexical dataset called Hokkaido University English Vocabulary List (HUEVL), in which 7453 words are selected as learning-purpose vocabulary at school.

In the list, each word is classified into 5 groups from level 1 (beginner: junior high level) to level 5 (advanced: university level). Sonoda selected those words as frequently-used as well as basic vocabulary and described that the list would be a kind of end-point (or destination) of vocabulary learning travel for learners and also warrant the instruction of vocabulary learning.

Aoyama[3] describes that the academic society of English education called COCET compiled a vocabulary list which included 3300 words for technical college students to learn during the period of their 5-year school days. A concept of the list is to include potentially required words for engineering use at work after their graduation. Another is to learn them with multimedia devices. The list was not only published as a word book but also developed as a smartphone app as shown in Fig.1.

Fig. 1. Examples of word book (L) and smartphone app (R)

C. *Effects and Problems of Vocabulary Lists*

Such books and apps as in Fig.1 are considered as good educational materials at a glance that make learning effective and efficient. However, there are some researchers who pointed out problems on vocabulary learning with them. Kawamura *et al*.[4] pointed out that students still ran short of learning time to familiarize themselves with new words even in elementary school. He cited students' quick-fix or temporal approach to vocabulary learning as an example due to tight learning schedule at school.

Kaneta[5] pointed out that already-existing vocabulary lists were apt to be selected subjectively and have both good and bad aspects. He said that the former was the lists increased educational validity and that the latter was subjective lists whose name is like "basic words for daily conversation" or "necessary words for English instruction" left a problem of objectivity in processing data.

III. Research Questions and Hypotheses

A. *Problems to Solve in Vocabulary Education*

There are several problems to solve in the field of vocabulary education. These are considered as making the activities of vocabulary learning ineffective and inefficient.

- Although quantity as an end point (or a destination) for learning words is indicated as a vocabulary list, it is not explicitly delivered to learners.
- Even if there is an end point, the specific order and procedure to learn words is still unclear. It is as if students would have to learn just like groping in the dark.
- Teaching as well as learning time is still limited in the classroom at school.

On the other hand, there are also problems to solve in the field of language learning service, which would make degree of user satisfaction lower.

- There are vocabulary learning systems using smartphone apps as well as websites. Students can learn freely if they have a smartphone or a tablet with them. But most of the systems are based on a one-to-one correspondence between Japanese and English. It might make flexibility or applicability low.
- Such systems are mainly based on textual information. There is little vocabulary information that visually or graphically makes users understand the applicability of words, such as conceptual diagrams.

B. *Research Hypotheses*

Based on the present problems in the former section, then the authors set up the following research hypotheses.

- If an appropriate service of language learning support that explicitly shows co-occurrence relationships and conceptual connections between words is provided for learners as a user, then it will become possible for them to learn new vocabulary with a greater awareness of the connections. And they will be able to reduce less amount of their mental burden in memorizing each new word.
- If the appropriate service through vocabulary grouping with visualized vocabulary maps is provided as a learning service, then learners as a user will be able to obtain higher motivation and satisfaction of language learning by contacting multiple words at once.

The service will increase the efficiency of learning work and make it easier to obtain the effect of long-term memory retention compared to before. Rather than learning words individually, it will be better to learn words that are connected by grouping together because the learners will find more routes to access words. So it can be easier to remember memories, reduce the burden of memory, and reinforce the efficiency of learning work.

IV. Development of Contents

A. *Vocabulary Size*

Lonsdale[6] shows an interesting statistical data for communicative language learning. He shows that 85% words of daily conversation are covered if a person acquires first top 1000 words in frequency and that 98% words are covered if he/she acquires first top 3000 words in frequency. It is shown graphically in Fig.2. The authors regard these number 1000 and 3000 as an important key for language education services.

This research is an exploratory research. In order to facilitate this vocabulary analyses, the authors started by focusing on as few core vocabulary words as possible. As mentioned in the previous section, with elementary school English in mind, the lexical frequency information to be handled is limited to nouns and verbs that are the main elements of English sentences. 2,105 words that are within the top 3,000 ranks in Table 1 are targeted, according to the vocabulary size of Lonsdale's scales.

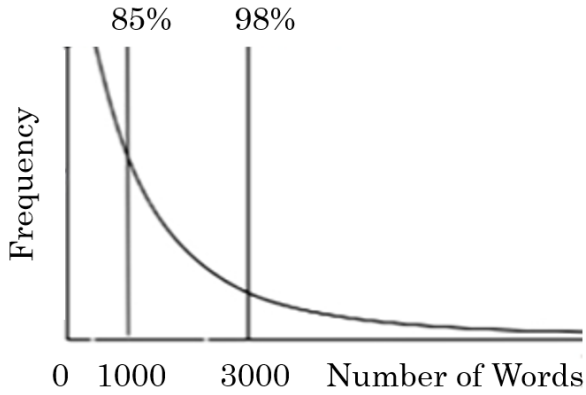

Fig. 2. Graph of Word Frequency and Counts

TABLE I. NUMBER OF WORDS IN FREQUENCY RANK IN BNC LIST

Ranked	Nouns	Verbs	Totals
in the Top 1000	419	218	637
in the Top 2000	954	419	1373
in the Top 3000	1477	628	2105
out of Top 3000	1785	653	2438
subtotal	3262	1281	4543
total		6318	

On the other hand, in narrowing down the learning vocabulary, the authors focused on the top 1000 and 3000 words, which are Lonsdale's scales, and analyzed 4659 words from levels 1 to 3 of HUEVL, shown in Table 2.

TABLE II. NUMBER OF WORDS IN DIFFICULTY BASED ON HUEVL

Difficulty Level	Counts	Level
Level 1 only	785	Junior High
Level 2 or less	2563	Senior High
Level 3 or less	4659	Univ Exam
Level 4 or less	6179	Univ Basic
Level 5 or less (Total)	7453	Univ Advanced

In the BNC vocabulary list, the authors extracted the relationships between hypernyms and hyponyms in WordNet, targeted at only nouns and verbs in the top 3000 with high frequency of use as well as in the difficulty vocabulary list of HUEVL, both of which are regarded as highly important in common. Table 3 shows the comparison of vocabulary frequency and occurrence distribution of learning vocabulary.

TABLE III. DISTRIBUTION OF FREQUENCY AND DIFFICULTY

Freq \ Diff	Jr. High	Sr. High	Univ Exam
Verbs Top 1000	129	197	197
Verbs Top 2000	148	308	330
Verbs Top 3000	154	389	439
Nouns Top 1000	184	366	379
Nouns Top 2000	256	693	772
Nouns Top 3000	274	901	1058
V+N Top 1000	313	563	576
V+N Top 2000	404	1001	1102
V+N Top 3000	428	1290	1497

In terms of English vocabulary in start-up level learning, nouns have more than verbs within the top 1000 frequencies. The ratio is about 1:1.43. In addition, when the frequency is expanded to the ranking down to the top 3000, the difference in the number of nouns and verbs in terms of lexical occurrences widens further. The ratio becomes 1:1.78. It was found that the verbs increased by about 310 words from the junior high school level to the university entrance exam level, while the nouns increased by 874 words. The need to memorize more nouns than verbs may be reflected in the increase in the amount of individual human knowledge.

B. Integration of Vocabulary Lists

As for the datasets of frequency, many researchers have tried to compile various vocabulary lists for language analyses. Kucera and Francis[7] compiled a dataset based on a text data which consisted of a million words called Brown Corpus. This is the first computational analysis of English vocabulary in the world in 1960s. Kilgarriff[8] compiled another dataset based on a hundred million word corpus called British National Corpus. It includes 6318 different words each of which occurs over 800 tokens in frequency with its part-of-speech tags. In Japan, Sonoda compiled a vocabulary list for English learners from language education point of view, as already described above.

TABLE IV. VOCABULARY LISTS FOR INTEGRATION

Types	Lists	Counts
Lists of Frequency	Kucera & Francis (1967)	-
	Kilgarriff (1995)	6318
	Sugiura 1 (2002)	5776
	Sugiura 2 (2002)	7286
Lists of Difficulty Level	HUEVL (1996)	7453
	JACET 8000 (2003)	8000
	COCET 3300 (2007)	3300
	ALC SVL 12000	12000
Lexical Database	WordNet	150000

Kikuchi and Ono[9] attempted to integrate the following vocabulary datasets of Table.4. They conducted text processing through Perl programming with a linux computer and succeeded in compiling the integrated datasets of English vocabulary lists which includes 15885 types of word in total, as shown in Fig.3.

```
2 3 - v3 2 3 04 govern
- 4 7 -- - - -- governess
- - - -- - - 08 governing
- 1 1 n1 1 1 03 government
- 6 - -- - 7 08 governmental
- 6 4 n3 2 3 03 governor
3 - 7 n6 2 6 05 gown
3 3 3 v3 2 2 05 grab
3 4 6 n5 2 4 03 grace
- 6 4 -- 3 8 05 graceful
- - - -- - 8 07 gracious
2 2 2 n3 2 2 02 grade
- - - -- 4 - 08 grader
```

Fig. 3. Example of integrated vocabulary lists

Each vocabulary list has its own characteristics. COCET 3300 was edited as a science-oriented list. Two Sugiura lists were based on the vocabulary which appeared in the high school textbooks of English subject approved by the Japanese Ministry of Education. They show vocabulary in educational purposes. The words in SVL 12000 were selected for every level of Japanese learner of English from primary school to

postgraduate level in terms of "usefulness" and "importance" as well as "frequency", according to ALC Press[10].

C. Extraction of WordNet Vocabulary Data

Vocabulary datasets are derived from the field of computing science as well as linguistics. One of the most significant computing datasets is WordNet, which was firstly developed by Fellbaum[11]. It consists of about 150,000 words of nouns, verbs, adjectives and adverbs, and also includes about 210,000 patterns of each word spelling and definition. Now everyone can download the database for free from the designated website of Princeton University. According to the website, these words in WordNet are grouped into sets of cognitive synonyms called synsets. And the synsets are interlinked by means of conceptual-semantic and lexical relations.

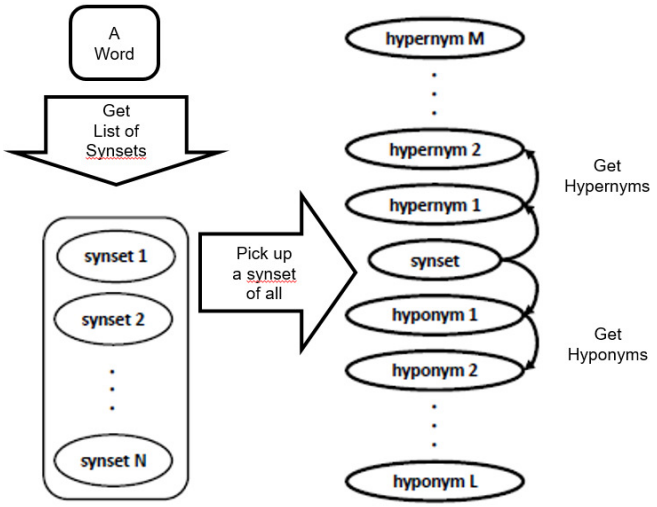

Fig. 4. Flow of extracting conceptual structures from WordNet

D. Visualization of Vocabulary Map

When language learners try to study their target language, they tend to focus on not only a single word itself but also a series of words like "chunks" and "collocations" as well as grammatical and idiomatic phrases. They also tend to demand paraphrastic or interchangeable expressions so that they can substitute an alternative word or phrase for what they want to say first. WordNet will be a possible solution because it has synonyms as "synsets", as it were, a thesaurus. It also has another information of hypernyms or hyponyms as each word relationship. Such information will help the learners with their foreign/second language acquisition.

The research group including the authors[14] attempted to utilize WordNet to enrich the database with hypernym-hyponym relation of each vocabulary. In order for language learners to understand the word relation, the researchers tried to visualize the hierarchical word relation, as shown in Fig.5. Based on electrical dataset in WordNet, they firstly set up a tentative word search program of WordNet under linux environment through Python programming and then extracted visualized images for each word derived from the integrated vocabulary database.

TABLE V. ACCENT COLORING FOR ANALYZED WORDS

HUEVL	JACET	BNC	Counts	Colors
LV 1, 2, 3		N1+V1	576	Red
		N2+V2	526	Aqua
		N3+V3	396	Lime

To extract and visualize the vocabulary data as a vocabulary map from WordNet, the list of JACET in addition to BNC and HUEVL was used in order to highlight important words. First, 2105 words were selected as shown in Table1, which are ranked in the top 3000 in BNC list. Then, 1497 words were extracted out of the 2105 words that occur in common from level 1 to 3 corresponding to junior high level to university exam level in both JACET and HUEVL lists. Finally, the last 1497 words of conceptual structures between each word were extracted and visualized from WordNet as well as color-highlighted by level as shown in Table 5. and Fig. 5.

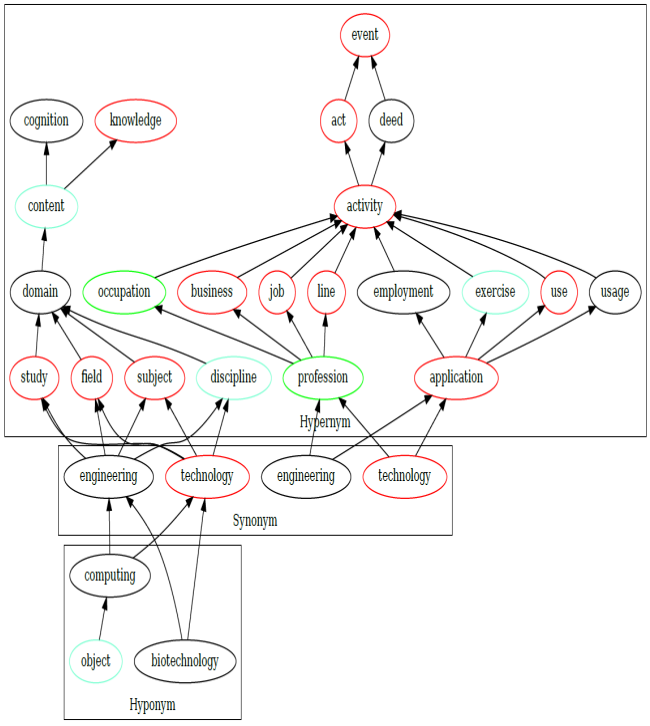

Fig. 5. Conceptual Map: Visualization of Word Relationship

V. CONCLUSION AND FUTURE STUDIES

This paper showed how to select important words for vocabulary learners, and how to extract and visualize vocabulary map with them from WordNet from the viewpoint of contents development. The derived data and map of words relationship will be used for future language learning service in the next step. Service system needs to be developed and also evaluation model for effectiveness, efficiency, and satisfactory will have to be examined as future studies.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP22K02825.

The authors are grateful to Mr. Haruka Hosokawa and Mr. Mikiyasu Nishiyama for collaboration on the early stages of this work.

REFERENCES

[1] F. Tamamura, The 7th Vocabulary Studies – Japanese Word Characteristics and Vocabulary Education, Japanese Education Report 30, The Japan Foundation, 1998, pp.10-11. (in Japanese)

[2] K. Sonoda, Daigakusei-you eigo goihyo no tameno kisoteki kenkyu[Fundamental Research for English Vocabulary List for College Learners], Language and Culture Studies Series 7, The Institute of Language and Culture Studies Hokkaido University, 1996, p.200. (in Japanese)

[3] A. Aoyama, Development of "COCET3300" and its Practice in English Classes, Research reports of Toyama National College of Technology 40, Toyama National College of Technology, 2006, pp.15-24. (in Japanese)

[4] K. Kawamura, M. Takasu, R. Okamura, T. Okai, Short-Duration Lesson Components for the Acquisition and Use of Words and Sentences in Elementary School English Classes, JES-Journal 18(1), The Japan Association of English Teaching in Elementary Schools, 2018, pp.166-181. (in Japanese)

[5] T. Kaneta, Kyoiku Goihyo no Hensen ni Miru Corpus to Kyouiku no Setten[A Contact between Corpus and Education with respect to Educational Vocabulary Lists], Web Magagine Lingua 24, Kenkyusha [online] https://www.kenkyusha.co.jp/uploads/lingua/prt/15/KanetaTaku1506.html (accessed Feb. 14, 2023)

[6] C. Lonsdale, How to learn any language in six months, 2004, [online] https://youtu.be/d0yGdNEWdn0 (accessed Feb. 14, 2023)

[7] H. Kucera, W.N. Francis, Computational Analysis of Present Day American English, Brown University Press, 1967.

[8] A. Kilgarriff, BNC database and word frequency lists, 1995, [online] https://www.kilgarriff.co.uk/bnc-readme.html (accessed Feb. 14, 2023)

[9] M. Kikuchi, M. Ono, Redefinition of English Words Difficulty in Language Learning through Reanalyses of Vocabulary Lists, Proceedings of 19th Kosen Symposium in Kurume, National Institute of Technology, 2014. (in Japanese)

[10] ALC Press, What is Standard Vocabulary List 12000?, 2022. (in Japanese) [online] https://www.alc.co.jp/vocgram/article/svl/ (accessed Feb. 14, 2023)

[11] C. Fellbaum, WordNet: An Electronic Lexical Database, The MIT Press, 1998.

[12] H. Hosokawa, M. Nishiyama, M. Ono, S. Anda, M. Kikuchi, T. Ozono, T. Soga, T. Tanabe, Extraction and Visualization of WordNet Datasets based on English Learning Vocabulary Lists, Technical Report of IEICE, The Institute of Electronics, Information and Communication Engineers, 2022, pp.64-69. (in Japanese)

[13] M. Ono, T. Soga, Enrichment and Refinement of Language Learning Vocabulary Database, Proceedings of CIF22, Chitose Institute of Science and Technology, 2022.