Vol.2014-CE-127 No.9
2014/12/7

Text Glossing Methods for Computer-Assisted Language Learning

BURLAKIN VADIM[†1]　NAKAGAWA MASAKI[†1]

This paper presents two glossing methods to augment the self-study language learning approach based on "learning through intensive reading." Related studies have revealed that the coverage rate required to read foreign texts smoothly is fairly high. Furthermore, the vocabulary gains from foreign text reading are low. These factors limit the number of learners who are able to study by reading foreign text, and decrease learning effectiveness. In this study we address these problems of "learning through intensive reading", and propose two glossing methods: i) picture glosses and ii) Kanji decomposition glosses. We have implemented them into a prototype system and discuss preliminary results.

1. Introduction

Recent developments in computer technologies brought learning and teaching to a new level. Just a few decades ago learners had to use printed textbooks and dictionaries for second language (L2) learning, which seemed to be very effective. Bearing that in mind, it becomes clear how recent achievements in technology help students nowadays: electronic dictionaries, which save a lot of time compared to their printed "legacy" versions, interactive computer programs, or even online lectures. The latter is a technology which no one ever thought of just ten years ago: nowadays students can attend highly-ranked courses completely online and get a certificate ("*The future of higher education: How technology is changing today's classrooms*", Marie Glenn, 2008). Modern computer-assisted language learning systems (CALL), also provide students with tools for distance learning, virtual classrooms which allow students to easily communicate with instructors, language learners and native speakers. Electronic flashcard applications (*Anki* and its analogs) allow students to automate the card generation process, as well as let learners to include various multimedia for easier memorization.

However, studying with instructor in the classroom remains the most popular way of learning L2. This is why most of CALL systems are focused either on supplementary study, such as vocabulary learning, or on enhancing the student-teacher communication by using distance learning technologies or virtual classrooms. Another systems (for example, *Livemocha* online language learning community) focus on creating a social-network-style structure, allowing learners to directly communicate with native speakers, which is very effective, considering that in some remote areas there may be no native speakers for speaking practice. Some other applications (*Rosetta Stone*, or similar software) include complete study programs with predefined vocabulary lists, reading and grammar exercises and other materials, which can be considered as a natural evolution of electronic textbooks.

†1 Tokyo University of Agriculture and Technology

Research in this area has also revealed that for certain stages of language learning instruction is not a priority. This can be exampled by a lot of Japanese language courses, where Kanji, and sometimes vocabulary learning is a complete self-study. These steps of language learning process rely almost completely on the learner's diligence and aspiration for learning new words and characters. To achieve this goal learners often resort to the help of different methods and numerous computer applications. We have based our research on one of the popular methods that students use: *learning though intensive reading*.

Learning through reading allows to solve a number of problems which learners encounter during their studies, such as: a) ability to choose literally *any* L2 text materials for the learning purpose; b) ability to better understand character and vocabulary usage in different contexts. The latter solves the common problem found in the *spaced repetition* method (e.g. *flashcards* method), where context information for the entries is usually not provided.

Our goal is to take the *learning through intensive reading* method to a new level. Our proposed solution is a computer application which would provide various types of assistance and hints for the learners, which would help them to acquire vocabulary and characters easily. The two methods that we introduce in this research are aimed at providing even more context information related to a specific word: a) visual information (using *picture superiority effect* for learning vocabulary); b) Kanji character decomposition (understanding the word meaning from the characters it contains). Our goal is to combine these methods with the *learning through intensive reading* approach and to analyze its impact on language learning. Both of the methods and their implementations will be described in details in this paper.

Currently, our system supports Japanese language learning for English speakers with at least JLPT N4 level of Japanese language proficiency, but a multilingual system is considered as a future work.

2. Related Works

Prior to conducting this study, we have analyzed previous works on the following topics: a) language learning through intensive reading; b) intensive reading and CALL; c) vocabulary acquisition methods; d) picture superiority effect in language learning; e) glossing L2 text.

In using the *learning through intensive reading* method, the first question is: which learners are eligible for this method, and which are not? What level of language proficiency and what amount of text coverage is required for using this method? We can refer to Paul Nation and Robert Waring with their study "*Vocabulary Size, Text Coverage and Word Lists*" to answer this question. According to this study, a small number of words appear very frequently in foreign text, and if learner knows these words, the learner will be able to understand a very large proportion of the written or spoken text. As an example, we can refer to the statistics shown on Figure 2.1, which shows vocabulary size and corresponding text coverage rates for English language.

Vocabulary Size	Coverage Rate
1000	72.0%
2000	79.7%
3000	84.0%
4000	86.8%
5000	88.7%
6000	89.9%
15,851	97.8%

Fig 2.1: Text coverage rates for English language

While the same numbers may not be applicable for other languages, we can understand that with an intermediate L2 knowledge we can comprehend a large proportion of foreign text. However, both this study and other similar researches suggest that at least 95% coverage rate is required for adequate comprehension and pleasured reading. But, we need to keep in mind that no computer-assisted learning (CALL) technologies were used in these studies.

Another question is: how much vocabulary can the learner gain from text reading? If vocabulary gains would be very low, then the *learning through intensive reading* method would not have any reason for being. According to Nation and Waring, incidental vocabulary learning in reading foreign texts is 6 to 8%, depending on the type of text (fiction book, newspaper article, etc.). Another research conducted by Michael Pitts, Howard White and Stephen Krashen ("*Acquiring Second Language Vocabulary Through Reading*", 1989) reveal similar results: 6.4 to 8.1%.

While at least 95% coverage rate, as it is suggested, is required for pleasured reading of foreign text, using CALL technologies may considerably decrease this percentage, allowing even less-skilled learners comprehend foreign text. To analyze this phenomena, we can refer to Lara L. Lomicka's study "*To Gloss Or Not To Gloss*" (1998). According to her study, "Through hypermedia-annotated text, readers will be able to approach the text more globally, rather than linearly. To achieve a more global understanding of the text, other multimedia annotations such as images, sounds, cultural, historical and geographical references, and guiding questions could enhance comprehension". In her study she suggests that the ability to present the text in the variety of ways (e.g. visual or audio) and glossing may increase both text comprehension *and* easiness of reading. A large number of CALL applications serving this purpose, such as *Rikaichan* for Japanese language, prove that learners want foreign text to be glossed, rather than to use external dictionaries and sources, like in the "classic-style" printed-book reading. However, most of these systems do not have any extra features other than providing glosses with translations.

Using visual information is also a popular topic for discussion among researchers. Picture superiority effect is analyzed in details by Paivio's (1971, 1976) dual-coding theory. According to this theory, pictures are remembered better than words because they represent both visual *and* verbal codes. Shana K. Carpenter and Kellie M. Olson have studied this theory in deep in their paper "*Are Pictures Good for Learning New Vocabulary in a Foreign Language? Only If You Think They Are Not*" (2011). According to their paper, as well as other similar studies, picture superiority effect can be used efficiently, but to a limited category of words (called *concrete* words). These papers also gave a better understanding of how learner's mind works, which, in turn, allowed us to introduce another learning method specific for Japanese language. In our Japanese language learning system we use Kanji decomposition method, where we show meanings of each separate Kanji the word contains, but we do not show translation of the word itself. Similarly to picture superiority, in this method we try to create additional "hints" in the learner's mind for a better recognition of studied vocabulary entries. This method can be applied to non-concrete words, where picture superiority is not eligible. However, there are some limitations, which will be described later.

3. Study Scenario

3.1 Material Upload

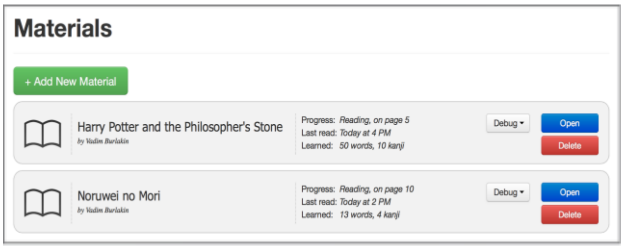

Fig 3.1: Material management interface

Learning process starts with material upload. Any raw text data of learner's choice can be used. In the prototype system, for testing purposes we have used the books "*Harry Potter and the Philosopher's Stone*" by J.K.Rowling and "*Norwegian Wood*" by Haruki Murakami. Figure 3.1 represents the material management interface.

3.2 Reading

After the study materials have been uploaded to the system, the student is able to start the learning process. Figure 3.2 represents the reading interface.

Fig 3.2: Reading Interface

Since the goal of our system is to provide glosses and hints for the *each entity* of the text, we need to separate the text into sentences and individual words. This is done in real-time when the program loads the reading interface for a given material. The text is firstly divided into sentences - *pieces*, and the sentences are then divided into individual words or phrases - *units*. A *unit* is the longest entity for which a translation can be provided. For example, an expression 調子に乗る is a *unit* and will be shown to the learner as it is, because it has its individual meaning, even though it is constructed of two words and a Japanese particle の. An example of this process is shown on Figure 3.3.

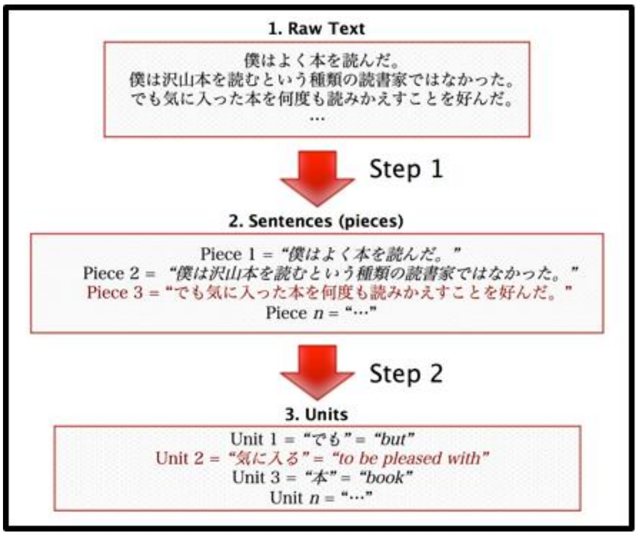

Fig 3.3: Text processing

Navigation through text elements is performed either by using the arrow keys on the computer's keyboard, or by mouse. By pressing up arrow key and down arrow key the user can navigate through the sentences, while left arrow key and right arrow key moves the cursor between individual words. In the similar way, hovering the mouse pointer over text's elements moves the reading cursor between text elements. If the learner encounters the word that he/she cannot understand, pressing Return button on the keyboard or left-clicking the word will show assistance.

Glosses are provided for each element of the text by using the following three methods: i) pictures; ii) Kanji decomposition; iii) translation.

Fig 3.4: Picture gloss

Fig 3.5: Kanji gloss

Fig 3.6: Translation gloss

The first method (*learn using pictures*) is based on the *picture superiority effect* and is similar to other "learn vocabulary through pictures" methods. Figure 3.4 represents a gloss used by this method. The difference of our system is that we *do not* show the translation to the user immediately. Instead, we show the pictures only, and ask the user to guess what is the meaning behind these pictures. By incorporating a "guessing game" into the learning process, we believe that: i) we can increase the learner's interest towards learning new vocabulary; ii) we apply an additional stimulation to the learner's mind which will help for better vocabulary recognition.

The *Kanji decomposition* method is based on the fact that a large number of Japanese words inherit their meaning from a combination of meanings of each individual character they contain. For example, the word 着陸 (*landing; to land*) is composed of two characters: i) *to arrive*; ii) *the land*. Similarly to the previously described learn using pictures method, we *do not* show the translation, but instead, show only information related to each separate character and ask the user to guess what would their combination mean (see Figure 3.5). In the example described above, the combination can be interpreted as "*to arrive to the land*", which would be enough for comprehension.

When both *learn using pictures* and *Kanji decomposition* methods are available, in the current implementation system shows Kanji gloss first. Then, if the learner cannot guess the meaning of the word, picture gloss is shown. If the user still cannot guess the meaning, by pressing the "Show translation" button we fall back to a regular translation gloss and show the word's translation.

However, both of the described above methods (*learn using* *pictures* and *Kanji decomposition*) have a number of limitations (will be described in the *Design and implementation* section), therefore, there may be situations where both of them will not be available. In these situations we also fall back to the *translation* method, and show a translation gloss (see Figure 3.6).

When assistance has been requested for a particular word (e.g. when we see that the user cannot understand a word), besides providing the glosses we also automatically add this word to the user's study list. This information will then be used in the *Practice* section.

3.3 Practice

In the Reading section of our system user enjoys reading his favorite book or magazine, at the same time incidentally learning new vocabulary and Kanji. However, the goal of Reading section is not only to provide a nice reading environment and entertaining guessing games, but also to collect the data about what the learner doesn't know. This data is stored in study lists and is used in the Practice section to check the recognition level of new vocabulary and consolidate user's knowledge of these new materials. During text reading we collect: a) vocabulary for which assistance was requested; b) example sentences, which are taken *directly from the text*.

Fig 3.7: Kanji-based flashcard

Compared to traditional vocabulary drilling techniques, for example a typical flashcard method, we use a slightly different approach for the vocabulary learning: *dynamic flashcards*. While preserving the general idea of flashcard approach, where each item has a "question side" and an "answer side", we try to add more word-specific information to the "question side" of each item. Furthermore, the flashcards are not persistent and may vary depending on the chosen study material and assistance methods available.

A common problem of learning vocabulary through regular flashcards is a complete absence of word's relationship to the real word situations and examples, as well as visual or audio information. A lot of students encounter a situation when a well-learned L2 word becomes very hard to recall in a real life situation. The reason behind it is that learning only word's definition (*verbal code*) does not create any associations and relations in the learner's mind. On the contrary, learning vocabulary in its native environment creates a lot of associations: in which context was the word used, how does it sound, how does it look like, and others.

The idea of dynamic flashcards is to try to recall words through these word-specific associations and analogies, rather than learning the word's definition only. To achieve this, we use a) images, in the same way that we use them in the reading process; b) audio guidance powered by text-to-speech engine; c) kanji compounds (Figure 3.7); d) context, where the word was encountered and can be used. For the latter we use the *exact same sentence* from the text where user had encountered the particular word, plus surrounding sentences to give more context information. This approach helps the user to create more natural associations with each word, which allows for a fast and effective recalling in real life situations.

4. Design and Implementation

4.1 End-User Client

End-user client represents a web-browser (in the current prototype version). However, most of the communication between the web server and the client are conducted through JSON APIs, which allows easy implementation of third-party clients (e.g. iPhone or Android applications).

4.2 Ruby on Rails Application

For website's core functionality, such as page rendering, navigation and user authentication, Ruby on Rails (RoR) platform has been chosen. The reasons for that is that: a) RoR provides all of the required functionality for creating a skeleton for a regular website out-of-the-box; b) RoR allows easy linking up with custom Ruby code and external APIs.

4.3 Text Segmentation

Since our system requires text segmentation to provide learning assistance, MeCab text segmentation engine (https://code.google.com/p/mecab/) coupled with a custom-built UniDic dictionary has been chosen for this purpose. MeCab engine is widely used in natural language processing, and is also used in variety of devices and systems including iOS and Mac OS X operating systems, where it serves as Japanese language input engine.

4.4 Picture Glosses

In our system we use images for more efficient study material acquisition. Therefore, our system requires an image database and a powerful search engine which would provide us with the most relevant results. For this purpose we use Google Image Search API. This API searches the Web for images by using keywords, and returns image URLs in the order of relevance, as well as other relevant information. The keywords are based on the filename of the image, the hyperlink text pointing to the image, and the text surrounding the image.

Our goal in this part of the system is to use a foreign word as a search keyword, and use Image Search API to fetch images which would represent the word's meaning. For example, for the word "apple" as a search keyword we expect to get an image containing a photo or drawing of an apple.

We have conducted an experiment by using different types of words as the search keywords and analyzed the returned images. Different types of words include: a) concrete nouns; b) abstract nouns; c) verbs; d) adjectives. We used three manually-selected words of each type and collected four images for the each word (total 48 images) from the Google Image Search API. For the *concrete nouns* we were able to get very relevant results (e.g. images which truly represented the subject; Figure 4.2). For other categories, however, the resulting images were vague and therefore impossible to use in our system (Figure 4.3).

However, as the results returned by Google Image Search API are unpredictable and uncontrollable, in some cases we were getting images not related to the original subject, even for concrete nouns. For example, for the word "apple" we were getting results both related to the apple fruit and to the Apple company. To eliminate this case we show three consequent images representing the subject, and ask the learner to unite the concepts behind each of the images into a single meaning.

Based on the results of the experiment, we decided to use the *learn by using pictures* method only for *concrete* nouns, since it was the only case when the guess rate was sufficient. To discrete concrete nouns from other words automatically in the system, we used the JUMAN dictionary (http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN), which separates nouns into different categories. Categories include: places and regions, transportation-related words, food and drinks, animals, human-built things, and others. We then embedded this information into Unidic dictionary and recompiled it, so that we could obtain category information in the above-described MeCab + Unidic combination automatically.

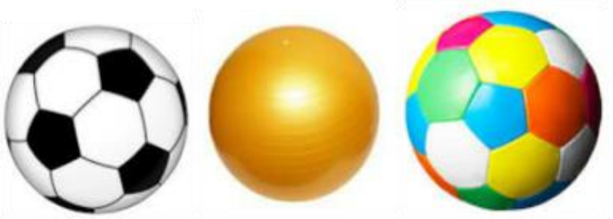

Fig 4.2: Search results for the word "ball"

Fig 4.3: Search results for the word "knowledge"

4.5 Kanji Glosses

In Japanese language, as well as in other ideographic languages, a large amount of words consist of several independent characters, each of which represents an independent sense (object, action, idea). The word, in turn, inherits these individual senses, which form the word's target meaning. This can be exampled by a Japanese word 住所 which consists of two characters: i) 住 (*to live*); ii) 所 (*place, location*). Together these characters form a word *"address; residence"*.

This feature of ideographic languages opens up new approaches for language learning. Instead of learning the characters and the words formed by them separately, the learning process can be combined. We use the same method as in the above-described *learn by using pictures* approach for the words consisting of two to three Kanji characters. We show senses of each individual character to the learners and ask them to combine these senses into a single concept. This technique allows learners: a) to remember the meaning of each individual character; b) to remember which characters are used in an each specific word; c) to create additional associations and analogies (*codes*) in their minds and thus increase recall efficiency. This approach especially helps to learn abstract words, where the *learn by using pictures* method is unavailable.

However, in some cases this method cannot be used. An example of such cases can be the word 動物 (*animal*), which consists of characters i) 動 (*to move*); and ii) 物 (*thing; object*). In this case individual characters do not have a strong semantic relationship to the target meaning, which leads to an abstract concept of *"moving objects"*. It is true, however, that the entity *"animal"* does have a connection to this concept and can be described as a *"moving object"*, but this relation does not work backwards.

To determine which words are eligible for this method we use WordNet to analyze the semantic connections between senses of individual characters and the word's meaning (Figure 4.3). If the characters' senses have a strong connection towards the target word's meaning, then we assume that this method can be used for this particular word.

For our project, we tested the algorithm against all JMDict entries consisting of two to three kanji characters, and embedded the results into the Unidic dictionary.

Fig 4.3: Analyzing semantic relationships

4.6 Multi Glosses

As above-mentioned, when several glossing methods are available for a particular word, we consecutively show different types of glosses. In the current implementation we show Kanji gloss first, picture gloss second, and finally, show the translation gloss (when other glosses are unavailable). Although this behavior is arguable, we are able to test the impact of every single glossing method on the learning efficiency.

However, as a future work of this project we are also considering to use so-called *multi-glosses*. A *multi-gloss* is a gloss which contains *both* image and Kanji information. For example, in case of a word 動物 (*animal*), which is not eligible for Kanji glossing for the above-described reasons, when coupled with a corresponding image, Kanji information may also be useful.

Consider the following scenario: learner encounters the word 動物 in the text. A Kanji gloss, asking *"What does `to move` + `thing, object` mean?"* is shown. Learner combines these two concepts into a single concept *"moving objects"*, and recalls some of the following (may vary from learner to learner): *cars, people, machines, airplanes*, etc. Some people may also recall *"animals"*, but the set of choices is too big. In this scenario, if the Kanji information would be coupled with a corresponding picture, then we could solve the above-described problem of backward-relations in Kanji glossing: learner would be able to narrow down the number of choices to a single concept *"animals"*. This phenomena is described in the Figures 4.4 and 4.5.

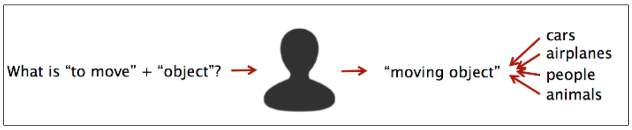

Fig. 4.4 Recalling process in a Kanji gloss

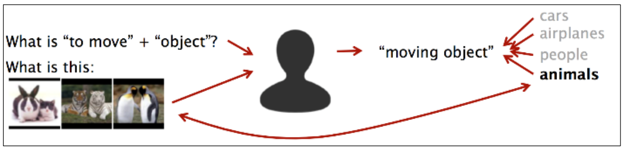

Fig 4.5 Recalling process in a multi-gloss

4.7 Dictionary

To provide translation assistance to the learner, a Japanese-English dictionary is required. For this purpose we use an open-source JMDict dictionary (http://www.edrdg.org/jmdict/j_jmdict.html). This dictionary is well-known for any Japanese language learner or Japanese language learning software developer. Such applications as "JED" or "imiwa?" for Android and iOS platforms respectively, are using this dictionary.

The dictionary comes in a raw XML format, which is unacceptable for our system due to very slow XML processing speeds compared to database engines. To solve this problem, we parsed the XML and imported all data into a MongoDB database. Furthermore, we mapped JMDict entry IDs to the Unidic entries and recompiled it, so that we could get translation information directly from the MeCab output.

4.8 Audio Guidance

Language learning applications, as well as most of the textbooks, are usually equipped with an audio guidance for vocabulary learning and listening practice. Usually, a pre-recorded voice of a real native speaker is used. This method can be implemented easily when the study materials are predefined, but it contradicts with our system's concept. To achieve the same goal, we use OpenJTalk (http://open-jtalk.sourceforge.net) - an open-source text-to-speech engine for Japanese language.

5. Discussion

While other researchers have already proved that using images has positive effect on learning efficiency, we have conducted a simple experiment to confirm this theory, and also introduced a solution for providing a "correct" image for each entity: by displaying several consequent images corresponding to each vocabulary entity. It is true, however, that this method strongly relies on external APIs and their algorithms (Google Image Search API, in our case), and therefore, can produce unexpected and uncontrollable results. This problem is addresses by many researchers in the image processing field, and numerous approaches for image ranking exist today. Also, ImageNet – the image database for WordNet synsets, could also serve our purpose. However, as a proof-of-concept, Google Image Search API showed acceptable results during our experiments. But as an improvement to this part of the system we consider replacing it with either ranking algorithms or an image database in the future.

Kanji decomposition algorithm is also a subject for discussion. We have implemented a simple algorithm for analyzing the semantic relatedness between each Kanji in the word and the word's meaning. We have also implemented a mechanism of a beforehand one-time analysis of all Japanese vocabulary and integration of its result (Kanji decomposition eligibility flag) into MeCab output. This allowed us to skip Kanji analysis procedure every time we display a gloss to the user. However, as the number of eligible words for this method (two- and three-character *kango* words) are finite, the necessity of an accurate Kanji analysis algorithm is arguable. Creating an algorithm capable of a very accurate Kanji analysis goes deep into natural language processing science and would consume a lot of time. To achieve our particular goal, the eligible words would be easier to separate by hand. There are approximately 50,000 two- and three-character *kango* words in Japanese language according to JMDict database, about half of which are relatively frequently used (words which have frequency factor above zero). Which means, that there are about 25,000 words which are the targets for our glossing method. Separating this number of words would be relatively easier and would produce much more accurate results, than creating a highly accurate separation algorithm.

We also have the Practice section partially implemented, but it needs further work. Current drawbacks of our system also include poor semantic relationship algorithms used in Kanji decomposition method, which is subject for a further research. We are also considering creating a tool for the manual separation of eligible words in our *Kanji decomposition* method.

After these major improvements will be made, we are planning to test our prototype system on real language learners and compare the study efficiency of our system compared to: a) traditional learning by using printed textbooks; b) other reading tutor applications.

6. Conclusion

In this study we are trying to analyze the impact of using different glossing methods, both separately and their combinations, coupled with the *learning through intensive reading* approach. For this purpose, we have created a prototype system which is capable of providing three types of glosses for reading foreign text: i) picture glosses; ii) Kanji glosses; iii) translation glosses. The current prototype uses all three types of glosses separately, but future work of this project also includes creating multi-glosses containing different codes (for example, both Kanji and image information).

We believe this learning system to be more efficient compared to traditional reading tutors and vocabulary learning methods. Our quick preliminary evaluation revealed the system to be more effective in terms of foreign text reading speed and vocabulary comprehension. As this is an ongoing research, however, this system has not yet been thoroughly tested.

Reference

1) Paivio A. (1971). Imagery and verbal processes. *New York: Holt, Rinehart, and Winston*.

2) Glenn M. (2008). The future of higher education: How technology is changing today's classrooms. *The Economist Intelligence Unit 2008*.

3) Nation P., Waring R. (1997). Vocabulary Size, Text Coverage and Word Lists. *Vocabulary: Description, Acquisition and* Pedagogy: 6-19. Cambridge: Cambridge University Press.

4) Nation P., Waring R. (2004). Second Language Reading And Incidental Vocabulary Learning. *Angles on the English Speaking World 4*: 97-110.

5) Pitts M., White H., Krashen S. (1989). Acquiring Second Language Vocabulary Through Reading. *Reading in a Foreign Language*, 5: 271-275.

6) Lomicka L. (1998). To Gloss Or Not To Gloss. *Language Learning & Technology Vol. 1, No. 2, January 1998*: 41-50.

7) Kilickaya F., Krajka J. (2010). Comparative Usefulness Of Online And Traditional Vocabulary Learning. *Turkish Online Journal of Educational Technology, 9(2)*: 55-63.

8) Folse S. (2004). Myths about Teaching and Learning Second Language Vocabulary: What Recent Research Says. *TESL Reporter 37, 2 (2004)*: 1-13.

9) Olson M., Carpenter K. (2011). Are Pictures Good for Learning New Vocabulary in a Foreign Language? Only If You Think They Are Not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol 38(1), Jan 2012*: 92-101.

10) Johnson C., Johnson D. (2014). Why Teach Vocabulary? *Anaxos, Inc.*

11) MeCab – Japanese Morphological Analyzer, https://code.google.com/p/mecab/

12) UniDic – Japanese Morphological Dictionary, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

13) JUMAN – Japanese Morphological Dictionary, http://www.ninjal.ac.jp/corpus_center/unidic/

14) The JMDict Project – Japanese-English Dictionary and Kanji Databases, http://www.edrdg.org/jmdict/j_jmdict.html

15) WordNet – A Lexical Database of English, http://wordnet.princeton.edu/

16) ImageNet – Image database for WordNet, http://www.image-net.org/