KWISC —Dependency Visualization for Understanding Context—

Yusuke Arai, Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama

***Abstract*—We propose a dependency visualization method for understanding context, KWISC. KWISC displays dependency between *bunsetsus* in Japanese sentences. In particular, it splits sentences into bunsetsus and displays them hierarchically according to the depth of dependency, to understand context. In addition, KWISC is able to align two keywords respectively, while KWIC aligns one keyword. This helps to find collocations among words. Furthermore, KWISC is able to expand and collapse bunsetsus. It can shorten distances of collocating words since it shows only the main structure of sentences by collapsing the bunsetsus. Therefore, collocating words are easy to fit on the screen, and horizontal eye movement is decreased when we look for collocations. As an evaluation experiment, we have collected pairs consisting of a bunsetsu that includes a keyword and another bunsetsu on which it depends from 207,802 sentences in the EDR corpus, and have measured the distances between them in KWISC. As a result, it is confirmed that KWISC makes the distances shorter than the conventional methods, and we have shown that KWISC makes it easy to find collocations.**

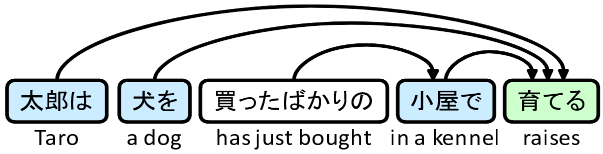

Fig. 1. Dependency in Japanese language

a_j: 犬 は次郎が育てる
b_j: 息子のために初めて 犬 を育てる
c_j: 太郎は 犬 を買ったばかりの小屋で育てる

a_e: Jiro raises a dog.
b_e: (He) raises a dog for his son for the first time.
c_e: Taro raises a dog in a kennel that he has just bought.

Fig. 2. Example of KWIC visualization

I. INTRODUCTION

WHEN we make a sentence by referring to examples, it is important to understand not only the meaning of words in sentences but also their context.

One of the methods for understanding context is KWIC[4] (KeyWord In Context), which shows sentences with keywords aligned in the center. It makes it easy to find collocations between a keyword and words around it by sorting the sentences by the strings before or after it, which leads to understanding the context of keywords. However, it helps to understand context derived only from word proximity while context is determined by various information. From this information, we consider not only word proximity but dependency between words as context.

Here, Figure I shows an example of dependency in Japanese. The Japanese sentence “太郎は犬を買ったばかりの小屋で育てる” (“Taro raises a dog in a kennel that he has just bought”) consists of five *bunsetsus*, which are linguistic units in a Japanese sentence and roughly correspond to basic phrases in English. A bunsetsu consists of a word alone, or a word with some particles or other suffixes. In Figure I, the last bunsetsu is a predicate consisting of a verb “育てる” (“raises”) alone. The first bunsetsu “太郎は” (“Taro” — a person’s name) is a subject, the second “犬を” (“a dog”) is an object of the verb, and the fourth “小屋で” (“in a kennel”) indicates a location. These three bunsetsus depend on the last bunsetsu “育てる” (“raises”) and they include a particle “は (*ha*),” “を (*wo*)” and “で (*de*),” respectively. Such particles play important roles in Japanese grammar. The third bunsetsu “買ったばかりの” (“has just bought”) is a modifier and depends on the fourth bunsetsu “小屋で” (“in a kennel”). As you see, such dependency information are useful for understanding context.

Now we will point out two problems with KWIC when we try to find collocations. First, although KWIC can help find collocations with proximity, it cannot help find collocations with dependency. Therefore, we sometimes find a useless collocation between a keyword and the word located around it and but not depending on it. In fact, Figure 2 shows an example in which KWIC shows sentences including a collocation “犬–育てる” (“dog–raising”). In the third sentence c_j , same as Figure I, since “買った” (“bought”) is located around a keyword “犬” (“dog”), it may lead to misunderstanding that “太郎は犬を買った” (“Taro bought a dog”), while “犬” (“dog”) actually depends on “育てる” (“raise”), not “買った” (“bought”).

Secondly, it is not easy to find collocations with the words not around a keyword using KWIC. Insertions and modifications around the keyword make the collocating words be located at various positions of sentences, which may cause us to miss collocations. For example, the word “育てる” (“raise”) in Figure 2 is located at various positions of sentences. In addition, if a sentence is longer than screen width, some words run off the screen and it is even more difficult to find collocations with such outside words.

For these problems, Akiyama et al.[1] proposed a method to show the main structures of parsed sentences and developed a tool using the method: TextImi[5]. It extracts simple sentences (i.e., sentences containing at most one subject and

Manuscript received June 15, 2009.
Y. Arai is with Graduate School of Information Science, Nagoya University, Chikusa-ku, Nagoya 464-8603, JAPAN.
M. Hagiwara is with Graduate School of Information Science, Nagoya University, Chikusa-ku, Nagoya 464-8603, JAPAN.
Y. Ogawa is with Graduate School of Information Science, Nagoya University, Chikusa-ku, Nagoya 464-8603, JAPAN (e-mail: yasuhiro@is.nagoya-u.ac.jp).
K. Toyama is with Graduate School of Information Science, Nagoya University, Chikusa-ku, Nagoya 464-8603, JAPAN (e-mail: toyama@is.nagoya-u.ac.jp).

978-1-4244-4139-6/09/$25.00 ©2009 IEEE

は (*ha*)	が (*ga*)	を (*wo*)	に (*ni*)	で (*de*)	Others	Predicate	Sentence
犬は a dog	次郎が Jiro					育てる raises	犬は次郎が育てる Jiro raises a dog.
		犬を a dog	ために for		初めて for the first time	育てる raises	息子のために初めて犬を育てる (He) raises a dog for his son for the first time.
太郎は Taro		犬を a dog		小屋で in a kennel		育てる raises	太郎は犬を買ったばかりの小屋で育てる Taro raises a dog in a kennel that he has just bought.

English translations are for purposes of illustration only.

Fig. 3. Example of the TextImi visualization

one predicate) from Japanese sentences, and shows their main structure. Here, the main structure of a sentence is defined as follows:

> The *main structure of a sentence* is its last bunsetsu and bunsetsus that depend on it.

In Figure I, for example, the main structure of the sentence consists of the last bunsetsu “育てる” (“raises”) and the three bunsetsus “太郎は” (“Taro”), “犬を” (“a dog”) and “小屋で” (“in a kennel”) that depend on it. The third bunsetsu “買ったばかりの” (“has just bought”) is excluded since it is not the last bunsetsu and it does not depend on the last bunsetsu.

Figure 3 shows an example of the same sentences as in Figure 2 displayed by TextImi. TextImi aligns the bunsetsus of the main structure of sentences in fixed columns in the [は (*ha*)], [が (*ga*)], [を (*wo*)], [に (*ni*)], [で (*de*)], [Others], [Predicate] order.

Here, last bunsetsus are aligned in the [Predicate] column. Bunsetsus including particles “は” “が” “を” “に” “で”[1] are aligned in the corresponding columns for [は], [が], [を], [に], [で]. Remaining bunsetsus are joined by a delimiter “|”, and aligned in the [Others] column. It only shows bunsetsus of the main structure of a sentence and removes insertions and modifications. In addition, it aligns bunsetsus with the same particle, which makes it easy to find collocations with the words not around a keyword.

However, it aligns bunsetsus in a fixed column, which may change the word order of a sentence. In the second sentence in Figure 3, “ために” (“for”) and “初めて” (“for the first time”) move rightward and cross over “犬を” (“dog”), and “犬を” moves away from “育てる” (“raise”). Since Japanese is a free word order language, visualizing this with TextImi does not very reduce the readability of the sentences but cannot help find collocations with proximity. That is, changing word order may lose proximity information in TextImi.

As we have mentioned, KWIC has two problems for Japanese, that is, it cannot help find collocations with dependency and it cannot help find ones with the words not around a keyword. TextImi overcomes the problems but causes another problem. Therefore, we propose a dependency visualization method for understanding context: KWISC (KeyWord In Structured Context), and solve the two problems by taking a different approach from that of TextImi.

[1]The roles for these particles are — “は”: subjective, “が”: subjective, “を”: objective, “に”: illative, inessive, “で”: locative, instructive.

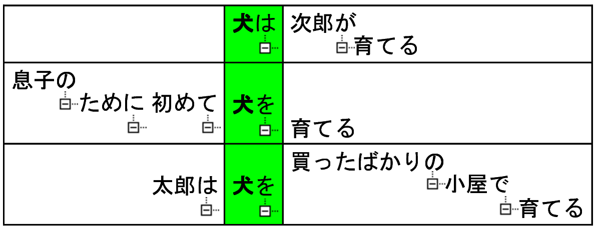

Fig. 4. Dependency visualization of KWISC

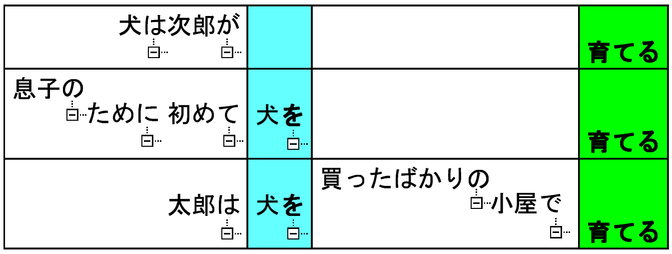

Fig. 5. Specifying a particle “を (*wo*)” as the secondary keyword against Figure 4

II. DEPENDENCY VISUALIZATION FOR UNDERSTANDING CONTEXT BY KWISC

In this section, we describe the proposed method, KWISC, which solves the two problems with KWIC by displaying dependency, aligning by two keywords, and expanding and collapsing bunsetsus in Japanese sentences. It does not change word order so that it does not cause the problem in TextImi.

A. Dependency Visualization

In order to solve the first problem of KWIC, KWISC displays dependency between bunsetsus. Figure 4 shows an example of dependency visualization by KWISC. It splits sentences into bunsetsus and displays them hierarchically according to the depth of dependency, i.e., the deeper the bunsetsus are in a parse tree, the higher level at which they are displayed. This helps to understand the dependency relationship between bunsetsus. In the third sentence, for example, you can easily understand that the third bunsetsu “買ったばかりの” (“has just bought”) depends on the fourth bunsetsu “小屋で” (“in a kennel”) and the three bunsetsus “太郎は” (“Taro”), “犬を” (“a dog”) and “小屋で” depend on the last bunsetsu “育てる” (“raises”).

In addition, KWISC shows sentences with keywords aligned in the center similarly to KWIC. However, it splits sentences into bunsetsus, and aligns bunsetsus including keywords. Figure 4 shows the same sentences as Figures 2 and 3, where the keyword is “犬” (“dog”) so that the bunsetsus “犬を” that includes the keyword are aligned.

B. Specifying Secondary Keyword

In order to solve the second problem with KWIC, KWISC enables the alignment of two keywords respectively without changing the word order, while KWIC uses only one keyword. Figure 5 shows the same sentences as Figure 4 in KWISC, with the keywords “を (*wo*)” (= objective particle) and “育てる” (“raise”). First, KWISC searches sentences by one keyword, i.e., the primary keyword. It aligns two keywords in the sentences respectively by specifying another keyword (secondary keyword). In Figure 5, the primary keyword “育てる” (“raise”) and the secondary keyword particle “を” are aligned. In Japanese grammar, such particles play important roles and the particle “を” indicates an object of a verb. So specifying “を” helps to find collocations between verbs and their objects. Specifying other particles also helps to find collocations with verbs and this advantage of KWISC is similar to TextImi. In addition, KWISC can specify words other than particles and it helps to find collocations between any two words so that it solves the second problem.

C. Expanding and Collapsing Bunsetsus

Using the method mentioned in Section II-B, KWISC aligns primary keywords and secondary keywords, respectively. However, it may align secondary keywords farther from primary keywords since it aligns bunsetsus without changing word order. Figure 5 shows such a case where the modifying bunsetsus “買ったばかりの” (“has just bought”) and “小屋で” (“in a kennel”) in the third sentence pull “犬を” (“dog”) and “育てる” (“raise”), while those are apart in the second sentence.

In order to alleviate this problem, we designed KWISC so that it can expand and collapse bunsetsus like the directory tree in Microsoft Windows Explorer. This enables showing and hiding each bunsetsu one by one, so that we can hide insertions and/or modifications between collocating words. Figure 6 shows an example of expanding and collapsing bunsetsus, where ⊞ buttons to expand are shown instead of collapsed bunsetsus. Figure 7 shows an example of collapsing all bunsetsus against Figure 5.

The initial display of KWISC shows only the main structure of a sentence as same as TextImi. Thus the lower figure in Figure 6 is the initial display when we search the sentence by the keyword “犬” (“dog”).

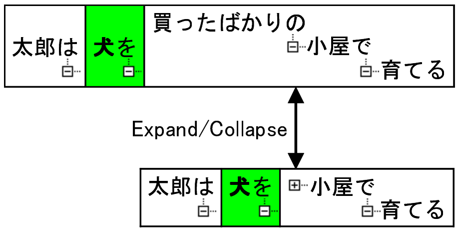

Fig. 6. Expanding and collapsing bunsetsus

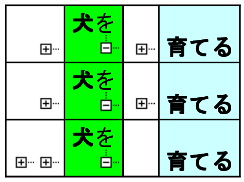

Fig. 7. Collapsing all bunsetsus against Figure 5

III. EXPERIMENTS

KWISC can shorten distances of collocating words since it only shows the main structure of a sentence, and can collapse these bunsetsus. Shortening this distance may be correlated with the ease of finding collocations, since collocating words are easy to fit on the screen, and the horizontal eye movement of users viewing this screen is decreased when we look for collocations. In this section, we evaluate how easy it is to find collocations with KWISC. Ideally, a subject experiment should be done to measure the ratio of successfully found collocations and the time taken to find them. However, it is not easy because of the time and the problem of choosing appropriate subjects.

Instead, the experiments measure the distances between two bunsetsus in KWISC on the screen, and we compare the result with related studies.

A. Experimental Settings

In this section, we explain the common settings for our two experiments. We have collected pairs consisting of the bunsetsus that satisfy specific conditions and other bunsetsus on which they depend from the corpus, and have measured distances between such bunsetsus.

The corpus consists of 207,802 Japanese sentences extracted from the EDR corpus[2]. Dependency information to collect pairs is taken from the analysis results of CaboCha[3], which is a Japanese dependency/syntactic parser based on machine learning. Although the EDR corpus included dependency information, we have adopted the analysis results of CaboCha, since it is a common free software in practical use.

Note that, to measure distances between bunsetsus on the screen is not easy since there are so many pairs, and the distances may vary depending on the environment such as size and resolution of the screen. Therefore, we measured the number of characters between two bunsetsus as distance in the experiments. We assume that non-proportional fonts are used, the width of ASCII characters is 1, and the width of Japanese characters is 2. Also, the width of ⊞ buttons in KWISC is assumed to be 2.

B. Comparison with KWIC

In the first experiment, we compare KWIC, which shows sentences as they are, with KWISC, which only shows the

TABLE I
INTER-BUNSETSU DISTANCE RESULT FOR KWIC AND KWISC

	Average	Standard deviation
KWIC	24.4	22.8
KWISC	9.9	9.8

main structures of sentences, and measure how much KWISC can shorten inter-bunsetsu distance.

1) Experimental Method: In this experiment, we measured the distance between depending and depended bunsetsus and we refer to the distance as *inter-bunsetsu distance*. We assumed a situation where we would search for some noun, and then find collocations of bunsetsus including the noun and other bunsetsus on which they depend from the result.

Therefore, we have collected 745,837 pairs consisting of a bunsetsu including a noun and another bunsetsu on which it depends. Since pairs of adjacent bunsetsus are not possible to be shortened any further, we removed such pairs and measured the inter-bunsetsu distances for the remaining 321,286 pairs. In addition, the "shortening ratio" of inter-bunsetsu distance using KWISC is measured, and is calculated by the following:

$$\text{Shortening ratio} = 1 - \frac{\text{inter-bunsetsu distance in KWISC}}{\text{inter-bunsetsu distance in KWIC}}.$$

2) Results and Discussion: First, Table I shows the average and standard deviation of the inter-bunsetsu distances. These are both less than half for KWISC than for KWIC. From this result, we conclude that KWISC makes it easier to find collocations.

Secondly, Figure 8 shows the distribution of the shortening ratio of inter-bunsetsu distance by KWISC. If only the main structures of sentences are found between collocating words, the shortening ratio is 0 since the inter-bunsetsu distance is the same as KWIC's. Although we excluded such pairs, the inter-bunsetsu distance of the remaining 253,266 pairs (78.8%) was shortened, meaning that KWISC is effective. The average of the shortened ratio is 0.51 and a large portion of the shortening ratio is distributed around 0.6–0.9. Therefore, KWISC can shorten distances of collocating words that have insertions and modifications more than 50%.

Finally, Figure 9 shows relations between inter-bunsetsu distance in KWIC and the shortening ratio by KWISC. Although the shortening ratio is less than 0.5 for an inter-bunsetsu distance of 20 in KWIC, the longer the distance is, the higher the shortening ratio becomes. Therefore, it is confirmed that the longer the inter-bunsetsu distance is in KWIC, the higher, and thus more effective, the shortening ratio is in KWISC.

C. Comparison with TextImi

In the second experiment, we compare TextImi, which may change the word order, with KWISC, which does not change it, and investigate the relationship between word order change and inter-bunsetsu distances.

1) Experimental Method: In this experiment, distances between bunsetsus including any particle of "は (*ha*)" "が (*ga*)" "を (*wo*)" "に (*ni*)" "で (*de*)" and other bunsetsus on which they depend are measured. Note that, since bunsetsus with such a particle usually depend on predicates in Japanese, we will refer to the depended bunsetsus only as *predicates* in the following.

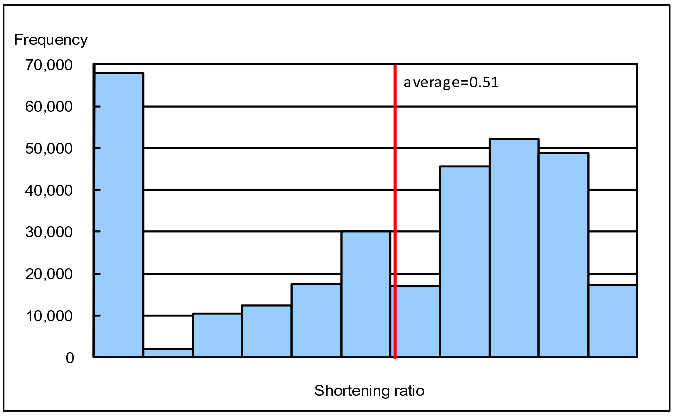

Fig. 8. Distribution of shortening ratio of inter-bunsetsu distances

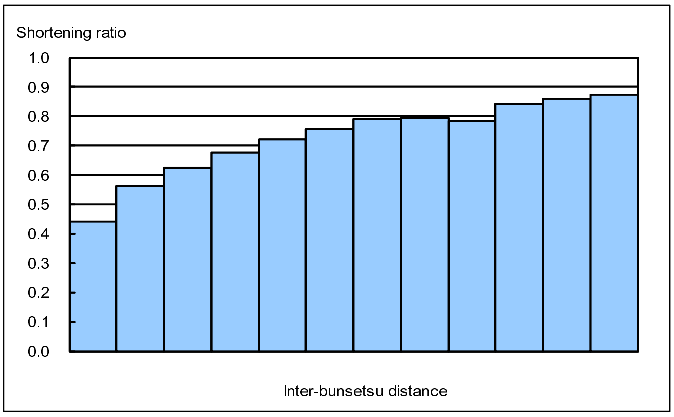

Fig. 9. Relation between inter-bunsetsu distances and the average of shortening ratio

We assumed that TextImi searches for some predicate, while KWISC specifies any particle of "は" "が" "を" "に" "で" as a secondary keyword after searching by the predicate. We have collected 511,982 pairs consisting of bunsetsus including any particle of "は" "が" "を" "に" "で" and other bunsetsus on which they depend, except for the pairs occurring only once in the corpus since collocations are usually found through more than one sentence. Table II shows the breakdown for each particle and we measured inter-bunsetsu distances of each pair.

2) Results and Discussion: Table III shows the average of inter-bunsetsu distance for each particle. KWISC shows shorter inter-bunsetsu distances than TextImi except for the particle "で (*de*)" (= locative or instructive particle). The cause for this is word order change. TextImi aligns all bunsetsus except for ones including any of the particles "は" "が" "を" "に" "で" to the column [Others]. In Figure 10, this Japanese sentence means "(Someone) drinks a large amount of alcohol until heavily drunk" where the subject is omitted, "つぶれるまで" ("until heavily drunk") and "ぐびぐび" ("(drink) a large amount"), which are located before "酒を" ("alcohol"), are aligned in the [Others] column, so the distance between "酒

TABLE II
STATISTICS OF BUNSETSU PAIRS USED IN THE EXPERIMENTS

	は (*ha*)	が (*ga*)	を (*wo*)	に (*ni*)	で (*de*)	Total
Number of pairs	108,681	102,179	145,307	109,099	35,217	500,483
Type of predicates	5,095	4,615	5,128	4,471	2,893	8,677

は (*ha*)	が (*ga*)	を (*wo*)	に (*ni*)	で (*de*)	Others	Predicate	Sentence
		酒を			つぶれるまで | ぐびぐび	飲む	つぶれるまでぐびぐび酒を飲む

Fig. 10. Aligning bunsetsus to [Others]

は (*ha*)	が (*ga*)	を (*wo*)	に (*ni*)	で (*de*)	Others	Predicate	Sentence
電話は	太郎が					使う	電話は太郎が使う
次郎は		ネットを				使う	次郎はネットを使う
お金は			大胆に			使う	お金は大胆に使う
辞書は				中学で		使う	辞書は中学で使う
言葉は					正しく	使う	言葉は正しく使う

Fig. 11. Examples of empty cells

は (*ha*)	が (*ga*)	を (*wo*)	に (*ni*)	で (*de*)	Others	Predicate	Sentence
太郎は		犬を		庭で、	初めて	育てる	庭で、太郎は犬を初めて育てる

Fig. 12. Example of TextImi where a bunsetsu with "で (*de*)" is at the head of a sentence

TABLE III
INTER-BUNSETSU DISTANCES FOR PARTICLES IN TEXTIMI AND KWISC

	は (*ha*)	が (*ga*)	を (*wo*)	に (*ni*)	で (*de*)	Average
TextImi	45.1	32.3	27.9	19.1	14.5	29.2
KWISC	29.3	16.2	9.8	14.2	18.5	17.6

Fig. 13. Example of KWISC for the same sentence as Figure 12

Fig. 14. Showing sentences with farther collocating keywords

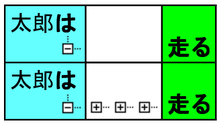

Fig. 15. Collapsing bunsetsus against Figure 14

を" and its predicate "飲む"("drink") gets longer. In addition, KWISC aligns bunsetsus to two columns, while TextImi aligns bunsetsus to seven columns of [は], [が], [を], [に], [で], [Others], [Predicate] in this order. Therefore, in TextImi, inter-bunsetsu distance may get longer than the original sentences and KWISC since there are more empty cells as shown in Figure 11.

Also, KWISC shows longer inter-bunsetsu distances than TextImi for the particle "で (*de*)" (= locative or instructive particle). This is because there are some sentences where the particle "で" is located before particles of "は" "が" "を" or "に" in the main structure. TextImi aligns bunsetsus to fixed columns in the order of [は], [が], [を], [に], [で], [Others], [Predicate]. Therefore, bunsetsus including particles of "は" "が" "を" or "に" which are located between bunsetsus including the particle "で" and the predicate move toward the beginning of the sentence, and inter-bunsetsu distance shortens.

Figures 12 and 13 show examples for a Japanese sentence "庭で、太郎は犬を初めて育てる" ("In the garden, Taro raises a dog for the first time.") in TextImi and KWISC, respectively. TextImi moves "太郎は" ("Taro" — a person's name) and "犬を" ("a dog"), which occur after "庭で、" ("in the garden"), toward the beginning of the sentence. As a result, "庭で、" is closer to "育てる" ("raise"), and their distance is 6 since there are three Japanese characters "初めて" between them in Figure 12. On the other hand, their distance is 16 in KWISC since there are eight Japanese characters "太郎は犬を初めて" between them in Figure 13. We counted the number of sentences that particles appear in the precise order "は" "が" "を" "に" "で", and found that the proportion of such sentences is 90.1%. From this, KWISC is not likely to lengthen distances compared to TextImi.

Finally, the average distances for the particle "は (*ha*)" (= subjective particle) in KWISC is longer than other particles. This is for the following reasons. When specifying the secondary keyword, KWISC aligns it based on the sentence that has the longest distance from the first keyword. Figure 14

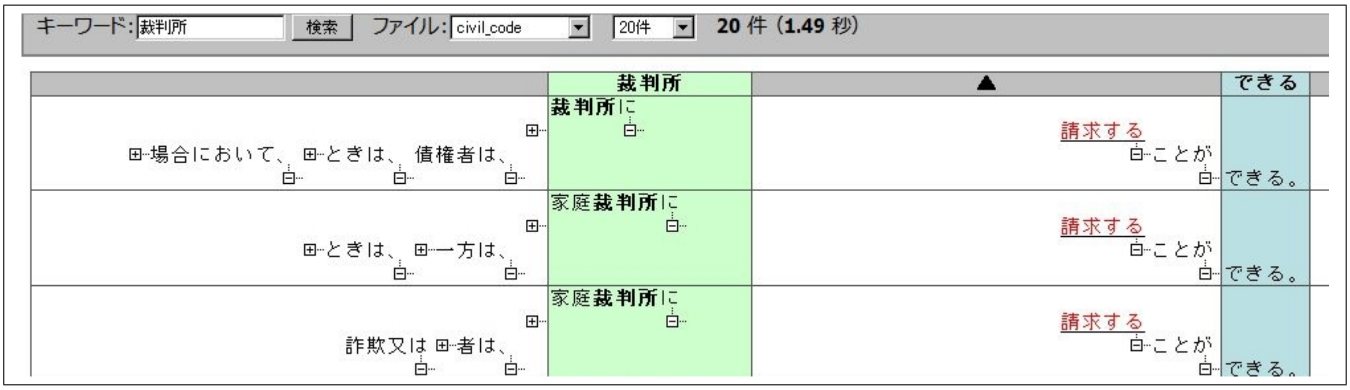

Fig. 16. Behavior of KWISC as a web application

shows an example of aligning based on the sentence with the longest inter-bunsetsu distance. In addition, the bunsetsu including the particle "は (*ha*)" is usually at the beginning of a Japanese sentence so that the distance from a predicate becomes longer than other particles.

Even in that case, KWISC can shorten such distances by collapsing each bunsetsu. Figure 15 shows an example of collapsing every bunsetsu in Figure 14. With this feature, KWISC makes it easy to find collocations.

IV. IMPLEMENTATION OF KWISC

We implemented the proposed method KWISC as a web application. The system assumes the search keyword to be a string not spanning two or more bunsetsus, and the searched target is a parsed Japanese corpus. Note that dependency should satisfy the following features that Japanese sentences usually have:

1) Each bunsetsu depends on only one bunsetsu except for the last bunsetsus. (The last bunsetsu does not depend on any bunsetsu.)
2) The depending bunsetsu always precedes the one on which it depends.
3) Dependency relations do not cross.

For now, we use a corpus parsed by CaboCha[3], which satisfies these features. Figure 16 shows an example of the system behavior.

V. CONCLUSIONS

We proposed a dependency visualization method for understanding context, KWISC, and solved the two problems with KWIC for Japanese. KWISC displays dependency between bunsetsus. In particular, it splits sentences into bunsetsus and displays them hierarchically according to the depth of dependency, and it helps to understand dependency.

In addition, KWISC is able to align two keywords respectively, while KWIC aligns one keyword. This helps to find collocations. Furthermore, KWISC is able to expand and collapse bunsetsus like a directory tree and it shortens distances between collocating words. Therefore, collocating words are easy to fit on the screen, and the horizontal eye movement is decreased when looking for collocations.

As an evaluation experiment, we have collected pairs consisting of the bunsetsus including a keyword and other bunsetsus on which they depend in the EDR corpus, and have measured the distances between them in KWISC. As a result, it was confirmed that KWISC makes the distances shorter than the conventional methods, and we have shown that KWISC makes it easy to find collocations.

Future tasks include supporting languages other than Japanese. For this purpose, it needs to support forward dependency to preceding bunsetsus, dependency to more than one bunsetsu, and crossing of dependency, all of which are rarely seen in Japanese. In addition, we need to precisely estimate the ease of finding collocations by a subject experiment that measures the ratio and the time to find collocations using KWISC.

REFERENCES

[1] Y. Akiyama, M. Fukaya, H. Ohiwa, and M. Tateno, "Extending Kwic Concordance by standardization of sentence pattern," *IPSJ SIG Notes*, Vol.2006, No.94, 2006, pp.105–112.
[2] Japan Electronic Dictionary Research Institute, *EDR Electronic Dictionary*, 1995.
[3] T. Kudo and Y. Matsumoto, "Japanese Dependency Analysis using Cascaded Chunking," *Proc. of the International Conference on Computational Linguistics*, 2002, pp.1–7.
[4] H.P. Luhn, "Key-Word-In-Context Index for Technical Literature (KWIC Index)," *American Documentation*, Volume 11, No. 4, 1969, pp.288–295.
[5] T. Nakano, *Development of the Meaning Chunk Extraction Tool TextImi and Its Significance for Web-based Social Survey*, Policy and Governance Working Paper Series, No.143, 2008.