# The Contextualized Representation of Collocation

**Daohuan Liu**
Huazhong University of
Science and Technology
liudh@hust.edu.cn

**Xuri Tang**
Huazhong University of
Science and Technology
xrtang@hust.edu.cn

## Abstract

Collocate list and collocation network are two widely used representation methods of collocations, but they have significant weaknesses in representing contextual information. To solve this problem, we propose a new representation method, namely the contextualized representation of collocate (CRC), which highlights the importance of the position of the collocates and pins a collocate as the interaction of two dimensions: association strength and co-occurrence position. With a full image of all the collocates surrounding the node word, CRC carries the contextual information and makes the representation more informative and intuitive. Through three case studies, i.e., synonym distinction, image analysis, and efficiency in lexical use, we demonstrate the advantages of CRC in practical applications. CRC is also a new quantitative tool to measure lexical usage pattern similarities for corpus-based research. It can provide a new representation framework for language researchers and learners.

## 1 Introduction

Collocation is an important concept in the fields of linguistics and computational linguistics (Firth, 1957; Halliday and Hasan, 1976; Sinclair, 1991), which can be widely used in language teaching, discourse analysis and other fields. Currently, there are two widely used representation methods of collocation, namely collocate list and collocation network. However, they are both flawed.

Collocate list takes the list of collocate words as the main form and generally provides the correlation strength, co-occurrence frequency, etc. between the node word and the collocate word[1]. Sometimes a collocate list may also include information such as the total frequency of collocates, the frequency of appearing on the left and right sides, etc. Table 1 shows the collocation list of the node word *importance* in a small news corpus.

| Collocate | PMI | Co-occur Frequency |
|---|---|---|
| attach | 11.216 | 29 |
| underscore | 9.223 | 2 |
| emphasize | 8.821 | 4 |
| stress | 8.811 | 19 |
| aware | 8.528 | 3 |
| awareness | 8.386 | 2 |
| great | 7.555 | 28 |

Table 1: Sample Collocate List of *importance* as a node. Pointwise Mutual Information is adopted to measure the association strength between two words (`measure=PMI`); Only collocates with 2 or more co-occurrences are considered (`min_freq=2`); Only collocates with an association strength greater than 7.5 are displayed (`thresh=7.5`).

[1]We follow the names by Sinclair (1991), and call the focal word in the collocation a "node word" (Node), and call the word appearing in the other position in the collocation a "collocate word" (Collocate).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

The expression capability is very limited through collocate lists, as they could neither present the interaction between collocates nor be visually friendly to readers. However, connectivity is an important feature of collocation knowledge (Phillips, 1985). In order to improve these weaknesses, Brezina et al. (2015) implemented the representation method of collocation graph and network[2] (see Figure 1). In a collocation graph, the collocates are scattered around and connected to the central word (node). The closer a collocate is linked to the node, the stronger it is associated with it. Compared to collocate list, collocation graph improves the visualization and enables the interaction of multiple collocations through node connection and graph extension. Brezina (2018) also demonstrates the possible applications of collocation networks with cases including discourse analysis, language learning, and conceptual metaphor research.

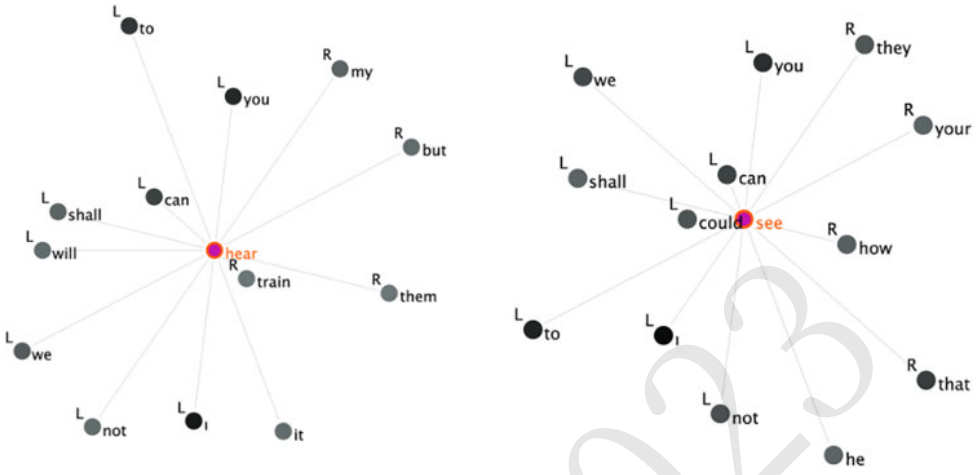



Figure 1: Sample Collocation Graphs of *hear* and *see* for image comparison, based on the corpora of World War I poems (Taner and Hakan, 2021).

However, these two traditional representation methods both have many critical flaws. The fundamental problem is that they neglect the natural language as a kind of sequence data. Collocate list regards collocation as a simple juxtaposition of tokens, and collocation graph regards each word as a free discrete data point in the space. Nevertheless, the context information is not only related to the semantics of the collocates surrounding a node but also related to the order and the position of the words.

First of all, they only tell the semantic relations but ignore the syntactic relations between nodes and collocates. The semantic association between nods and collocates is direct and clear. Firth's (1957) defines collocation as a container for semantic associations between the two words; and the meaning of a word comes not only from itself, but also from other words that co-occur with it. The association scores shown in both collocate list and collocation graph are an evaluation of co-occurrence, in other sense, a reflection of the semantic relationship. Nevertheless, Tang (2018) addresses that the syntactic association between the node and the collocate is also an essential part, acknowledging collocation as a complex of syntactic and semantic knowledge. He also sorts out the key role of syntactic relations in semantic theories and their applications through related studies (Katz and Fodor, 1963; Petruck, 1996). The syntactic nature of collocations is mainly reflected in the fact that collocations have direction and span. For instance, the two semantically related words *student* and *diligent* generally do not appear as "student diligent" in actual language use, which is syntactically incorrect in most cases; while "diligent student" or "student is diligent" is much more common and intuitively correct application. This shows that the collocation knowledge is actually an overall model that is restricted both by semantic relations and grammatical relations.

Secondly, they fail to reflect the relative position (relative distance) between nodes and collocates. Qu (2008) pointed out that the two components in a collocation tend to have fixed positions, i.e., one word

[2]A collocation network is a connected network of multiple collocation graphs. The two terms "collocation network" and "collocation graph" are used interchangeably in this paper, referring to the same representation method.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

always appears on the left or right side of the other word. For example, among the collocates in Table 1, *attach* mostly appears on the left side of the node *importance*. In the case that the positions of these two words are reversed, the syntactic relationship between the two words should also change. According to the data samples in the corpus, *importance* mostly acts as the object in *attach-importance* collocations, while *importance* often serves as the subject in *importance-attach* collocations.

Finally, these two representations cannot reflect the freedom of choice of collocates, which would restrict the usage of collocation in practice. This is not conducive to group collocates and find patterns with respect to semantic or syntactic relations. For example, if we want to find other predicates to substitute *attach* for the node *importance*, it is hard to tell from a raw collocate list or collocation graph. In order to satisfy this requirement, an extra screening operation such as Part-of-Speech (POS) tagging or syntactic analysis is required.

To overcome the above-mentioned shortcomings, we propose a new representation and visualization method to describe collocation, called Contextual Representation of Collocations (CRC), which makes improvements in syntactic representation and visualization abilities. Key features of CRC include:

- Applying conflated linear representation with respect to the nature of language;
- Foregrounding positional information of the collocates;
- Using spatial and visual symbols to indicate strength.

We will include three case studies of CRC respectively applied to synonym distinction, image analysis, and language teaching. These practical applications should reveal the advantages of this approach. In terms of knowledge representation, it can present more detailed grammatical information; in terms of knowledge application, it can achieve an accurate comparison of collocation distributions. CRC can provide a new representation framework for language researchers and language learners, as well as facilitate language research and teaching.

## 2 Contextualized Representation of Collocation (CRC)

While retaining the dimension of association strength, CRC promotes the relative position of collocates to another major feature dimension, so that each collocate can present those two important attributes at the same time. Therefore, the essence of the CRC is a two-dimensional scatter plot.

Figure 2 is an instance of CRC, which is based on the same data source as Table 1. The key features of this visualization will be explained in detail.

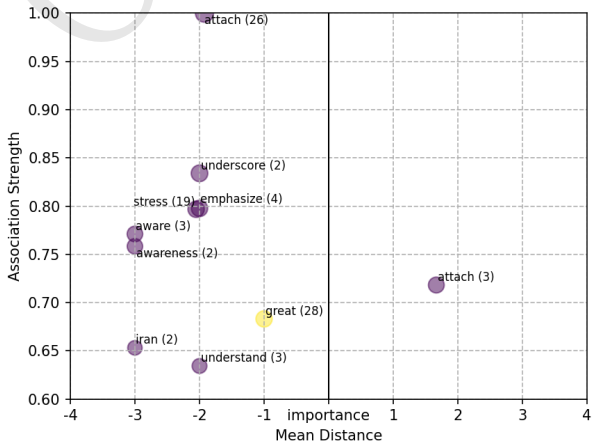



Figure 2: CRC of node *importance* (`measure=PMI`, `min_freq=2`, `thresh=0.6`).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

### 2.1 Conflated Linear Representation

Compared with the network structure of collocation graph, CRC follows the linear characteristics of natural language and uses conflated linear representation in expressing the relations of all the collocate-node pairs. In Figure 2, collocation relations are described as parallel horizontal lines, and compressed in a certain space range. This parallel and linear presentation helps to visually compare the commonalities and differences of different collocations, which is the basis of all the other advantages of CRC.

In our CRC implementation, we arrange the longitudinal spatial distribution of all collocate-node pairs according to their association score[3]. For the convenience of drawing and reading, the scores are normalized to [0,1]. The closer to 1, the higher the collocation strength.

### 2.2 Positional Information

In Figure 2, the horizontal dimension represents grammatical relations in terms of direction and distance. In this instance, the distance of each collocate is the average of the distances of all the collocate-node pair occurrences. `Distance=0` represents the position of the node word (*importance*).

Foregrounding positional information is the core contribution of CRC, as well as the key feature that distinguishes CRC from the other two representation methods. It is easy to understand that if the positional information is removed from Figure 2, all the collocate points will appear on the same vertical line, hence, it will degenerate into a simple visualization of the collocate list. Instead, if the Cartesian coordinate system is transformed into a polar coordinate system, it then becomes a collocation graph with fixed node positions.

Positional Information reflects the order of words, and the order of words further reflects the syntactic relationship. This can benefit CRC users with plenty of straightforward linguistic knowledge. Taking Figure 2 as an example, we could observe at least the following facts:

- The word *attach* tends to appear on the left of *importance* (`frequency=26`, `strength=1.00`) rather than on the right (`frequency=3`, `strength=0.72`), which might imply the *attach-importance* collocation is more likely to be used in active voice instead of passive voice.
- The word *importance* has a strong right-leaning tendency (Wang et al., 2007), which means it expects to be modified by a modifier prior to it.
- Collocates like *attach*, *underscore*, *emphasize*, *stress*, and *understand* might all play similar grammatical roles in the relationship with the node importance, because they all appear in the -2 position.
- People tend not to say "attach importance" but to use "attach great importance", which can be reckoned from their positions (*attach*=-2, *great*=-1, *importance*=0). This shows that CRC could also be used to recognize continuous word clusters and phrase patterns, making CRC more prospective in the application of analyzing and teaching. And it is capturing common contexts that the node is often used in. And this is also the reason why this representation of collocation is termed "contextualized".

### 2.3 Visualization Strategy

In addition to its advantage in context modelling of the node, as a representation method, CRC could be easily visualized with many visualization strategies. It can combine many spatial methods and visual symbols to expand the expression and presentation of collocation knowledge. As mentioned before, in addition to implementing the CRC in the plane Cartesian coordinate system, it can also be realized in the polar coordinate system; the size, color, and grayscale (transparency) of the data points in the figure can all be useful tools to group or describe collocates.

In general, compared with existing collocation representation methods, i.e., collocate list and collocate network, CRC can intuitively present richer context information and provide more convenience for researchers who use collocation analysis. In the following sections, we will apply CRC in three case studies of recent years to demonstrate the superiority of the new representation method.

[3]We use various measuring algorithms to calculate association scores and pick the intuitively best one from all the results. The adopted measuring method for each case is described in the captions of the figures.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

## 3 Case Study 1: CRC in Synonym Distinction

Many researchers utilize collocation analysis to distinguish synonyms. Liu has studied the usage differences of many synonym sets in English with the COCA corpus, such as *Actually*, *Genuinely*, *Really*, *Truly* (2012) and *Chief*, *Main*, *Major*, *Primary*, *Principal* (2010) using a behavioral profile approach. He also analyzed the learners' misuse of three synonym groups of *circumstance*, *demand*, and *significant* by comparing the use of these words in a second-language learner corpus (2018). Xiong and Liu (2022) compared the usage patterns and semantic differences of the two synonyms, *absolutely* and *utterly*, with the help of collocation lists and the Key-Word-In-Context (KWIC) function provided by the COCA corpus. Their conclusions are largely based on random sampling and qualitative analysis. Obviously, the above research methods take a large manual workload in observation and statistics. The use of CRC can not only carry out descriptive conclusions easily and clearly but also provide quantitative measures to differentiate synonyms.

In this case study, we use CRC to restudy the differences among the synonyms *Actually*, *Really*, *Truly*, and *Genuinely*, and try to verify some findings reported by Liu and Espino (2012). We choose a smaller corpus, BROWN, instead of COCA as the data source of this case study for its availability. The frequencies of each adverb in the BROWN corpus are shown in Table 2. It can be found that these four adverbs have a similar frequency proportion although the total data amount of BROWN is much smaller than COCA. Since *Genuinely* is not found in BROWN, we only study the other three adverbs.

| Corpus | #actually | #really | #truly | #genuinely |
|---|---|---|---|---|
| COCA | 105,039 | 263,087 | 20,504 | 3,065 |
| BROWN | 166 | 275 | 57 | 0 |

Table 2: Frequencies of *Actually*, *Really*, *Truly*, and *Genuinely* respectively in COCA and BROWN.

We retrieve the free collocates of *actually*, *really*, and *truly* from the corpus and generate three CRC instances (Figure 3a, Figure 3b and Figure 3c). Based on the frequency ratio of these three words, we set the thresholds of association strength to 0.3, 0.5 and 0.1 respectively for better visualization effects.

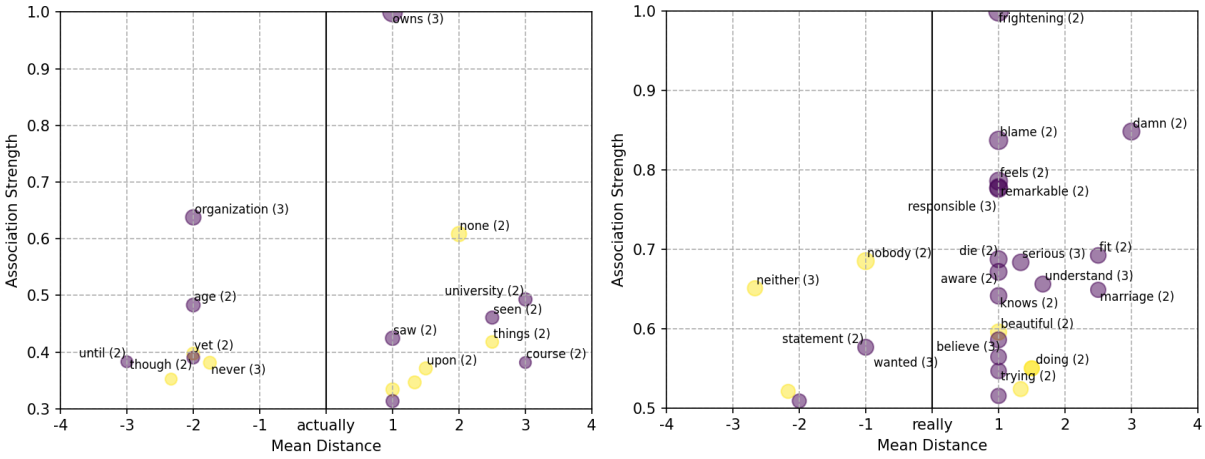
The usage pattern of *really* (Figure 3b) is significantly different from the other two words. There are much more highly associated collocates to the right of *really* (especially at `position=1`) than *actually* and *truly*, suggesting that *really* is more often used as verb and adjective modifiers. From Figure 3a, it is hard to find clear usage patterns of *actually* from the distribution of collocates on both sides. However, when compared to *truly* (Figure 3c), it could be observed that *actually* is surrounded by more content words and has many adversative words on the left side (e.g. *never*, *yet*, *though*, *until*), indicating that it may be used more as a disjunct. Figure 3c presents few meaningful collocates, but we can also reckon that *truly* is prone to occur as an adjective modifier from the limited samples. Moreover, from its potential context (*will/be* + *truly* + *fine/great*, etc.) we can infer that it is prone to be used for attitude emphasis and enhancement.

The above inferences are entirely based on CRC figures and are basically consistent with the main findings of Liu and Espino (2012) (see Figure 4) who adopted rigorous and systematic statistical methods (Hierarchical configural frequency analysis, HCFA). This demonstrates that CRC is a fast and handy tool in certain lexical studies. Of course, it would also be encouraged that researchers use other corpus approaches such as concordance and HCFA as auxiliary verification methods.

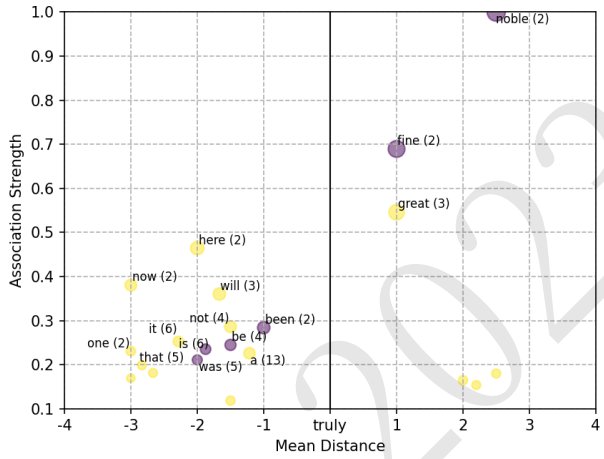
The network structure in Figure 4 is artificially constructed through subjective analysis. Yet CRC can make this relationship network more objective and accurate by quantifying the differences between the usage patterns of these words.

CRC preserves precise location and strength information of the surrounding words, thereby it is a collocate distribution of the node. This allows us to compute the degree of difference between distributions, namely the distance between CRCs. The smaller the distance is, the more similar the usage patterns of the two words. To compute the distance between two distributions, we are free to apply any suitable distance algorithm. Here we use a simple processing flow:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.




(a) *actually*. (b) *really*.




(c) *truly*.

Figure 3: CRCs of node *actually*, *really* and *truly* (`measure=PMI`, `min_freq=2`, `thresh` differs to accommodate to the number of data points for a better visualization).

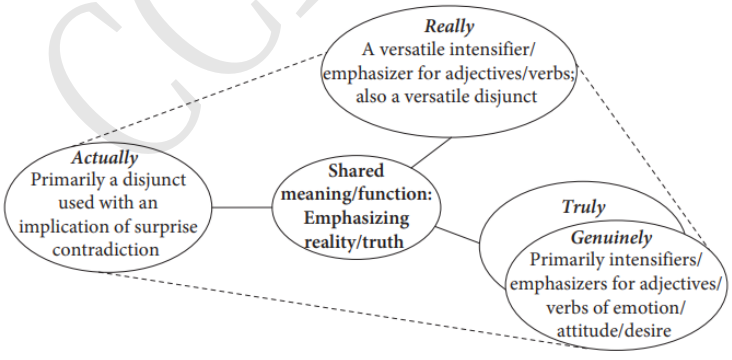



Figure 4: The internal semantic structure of the four synonymous adverbs by Liu and Espino (2012).

(1) For a common collocation word, calculate the Euclidean distance between the coordinates of the word on the two graphs.

(2) For unique collocations, they are not included in the distance calculation.

(3) Finally, average the Euclidean distance between all points to obtain the comprehensive distance.

Using the above algorithm, we obtained three distances (Figure 5). Our results show slight divergence

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

from Figure 4, as Liu and Espino considers *really* more similar to *truly* but CRC distance indicates that *really* is closer to *actually* (D(*really*, *actually*)=0.1641) rather than *truly* (D(*really*, *truly*)=0.1853). This difference might be due to BROWN's insufficient amount of data. It is also interesting to see CRC applied to COCA and verify if the semantic structure in Figure 4 is accurate.

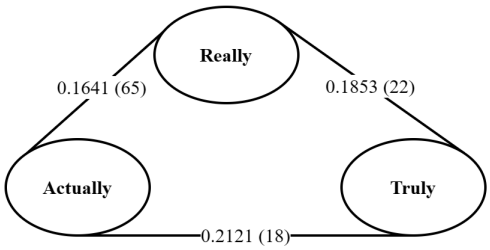



Figure 5: The collocation distribution distance of *actually*, *really*, and *truly* in the BROWN corpus. The number of valid collocation words actually involved in the calculation is indicated in brackets (`measure=PMI`, `min_freq=2`).

## 4 Case Study 2: CRC in Image Analysis

As an analytical tool, collocation is also widely used in other fields besides linguistic research. In the area of journalism and information communication, collocations are also used in assistance to discourse analysis, image analysis, and sentiment analysis (Koteyko et al., 2013). For example, Pan and Hei (2017) inspected the verb collocates on the right side of *we* in the interpreting corpus of press conferences, so as to analyze the image construction strategy of the government. In the field of digital humanities, collocation could also facilitate the style analysis of writing and author (Vickers, 2012; Wijitsopon, 2013; Taner and Hakan, 2021), as well as the image analysis of characters in the literary works. We select an image analysis task in a literary work and use CRC as an analyzing tool for research and discussion.

We pick the classic fiction "Lord of the Flies" (Golding, 1954) as our research subject, because the characters in this novel have distinctive characteristics and the language is simple and straightforward. Its characters and images have been heatedly discussed through book reviews and literary interpretations (Oldsey and Weintraub, 1963; Spitz, 1970), yet CRC may reveal the character-building methods from a corpus perspective.

This fiction narrates the story of a group of teenagers surviving on a desert island. The cooperation and confrontation among those children are interpreted as the epitome of human antagonism and political game. We select three main characters: Ralph, Jack and Simon as research objectives. Their right-side verb collocates are retrieved and described through CRC (Figure 6a, Figure 6b and Figure 6c). Before extracting collocations, the text is lower-cased and POS tagged. Different thresholds of association strength are applied in order to show a similar amount of collocates.

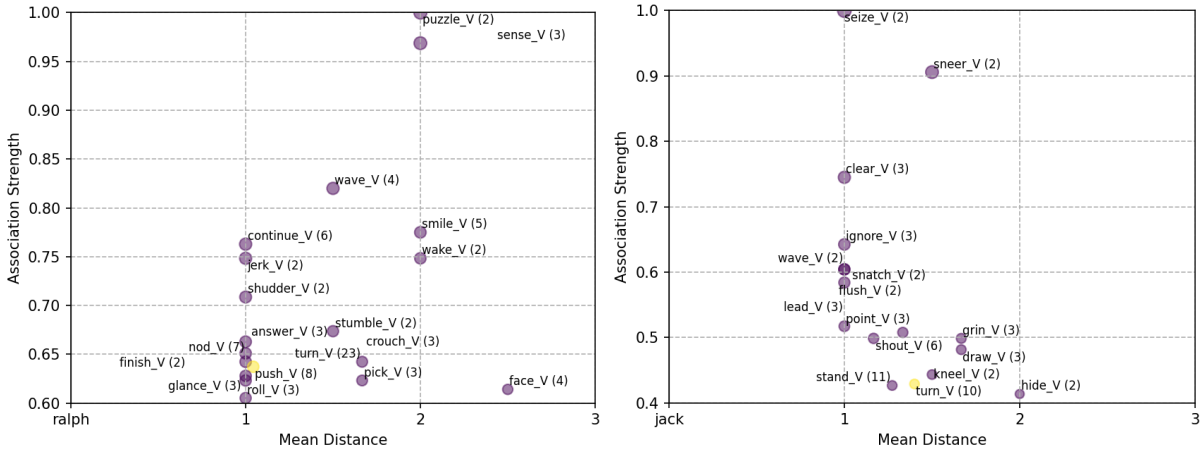
By comparing the three CRC figures, it can be found that when describing different characters, different verbs are used to shape the image of the characters. From the sentiment of verbs used to describe the character, we can infer the author's tendency in the image-building of each character.

Ralph is the main positive leader in the fiction, representing civilization and democracy before and after World War II. Figure 6a shows many clues in favor of that description. Verbs such as *puzzle*, *sense*, and *shudder* place Ralph on the opposite side of chaos and violence; others like *answer* and *nod* depict Ralph actively affirming and responding to others' opinions. These behaviors together shape Ralph as a "democratic man, the symbol of consent" (Spitz, 1970).

Jack is the representative of brutality and power. His unique behaviors include *seize*, *snatch*, *clear*, and *ignore* (Figure 6b), which show Jack's tendency to command and enforce, in consistent with the evaluation by Spitz (1970): "Jack then, is authoritarian man ... like Hitler and Mussolini".

Simon is regarded as "the Christ-figure, the voice of revelation" (Spitz, 1970). From his unique behaviors *lower*, *walk*, *speak*, *feel*, etc. (Figure 6c), readers envisage a sanctified image with calm,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.




(a) *ralph*. (b) *jack*.

(c) *simon*.

Figure 6: CRCs of node *ralph*, *jack* and *simon* (`measure=PMI`, `min_freq=2`, `thresh` differs to accommodate to the number of data points for a better visualization).

humility, detachment, and transcendence.

Apart from unique behaviors, similar behaviors are also described with verbs with different semantic polarities. For example, the author uses *smile* for Ralph but *sneer* and *grin* for Jack to express laugh; this further consolidates the contrasting images of the two characters.

## 5 Case Study 3: CRC and Efficiency in Lexical Use

Collocations could also play a role in second language acquisition and language teaching. The mastery of collocation is considered to be the decisive factor for the naturalness of a language learner's expression (Oktavianti and Sarage, 2021). One of the cases of collocation network illustrated by Brezina (2018) is the analysis and evaluation of different-level second language learners' expression. The selected corpus is Trinity Lancaster Corpus (TLC) of spoken L2 English (Gablasova et al., 2017), a transcribed corpus of English interview responses. Brezina divided the corpus into three sub-groups according to the speakers' language proficiency levels: Pre-intermediate (B1), Intermediate (B2) and Advanced and Proficiency (C1/C2). The most common collocates of the three verbs *make*, *take* and *do* used by students at these levels are shown with collocation graphs. Figure 7 shows the situation of *make*, from which we can observe a rise in the richness of collocates as the proficiency level lifts.

However, Brezina pointed out that no such "clear relationship between increasing proficiency and a higher number of collocates" was found on *take* and *do* (p.73). This implies that the increase in

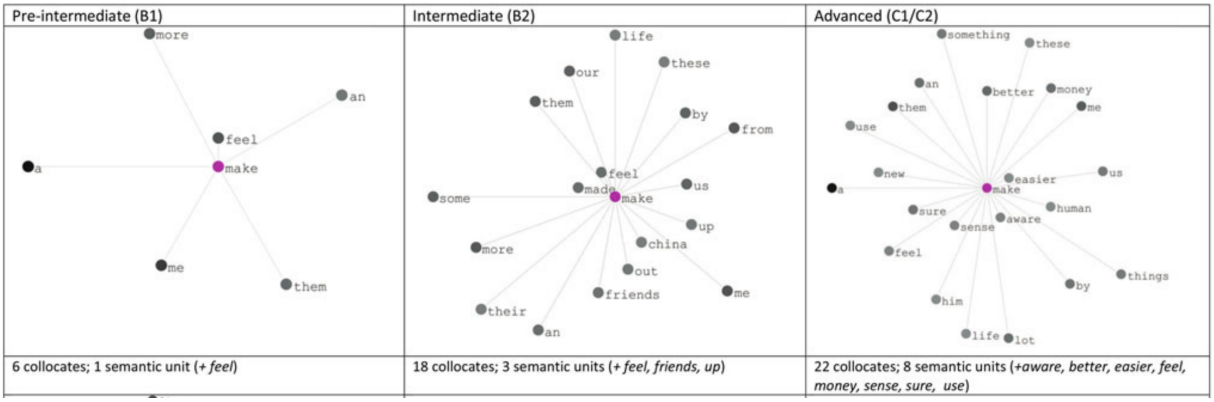



Figure 7: Collocation graphs of *make* for band B1, B2, and C1/C2 (Brezina, 2018).

collocate richness is not the decisive factor in measuring one's language proficiency. When talking about the language learner's communicative language competencies, *Common European framework of reference for languages: Learning, teaching, assessment* (CEFR) points out that a competitive language learner should not only "has a good command of a very broad lexical repertoire" (vocabulary range) but also masters "idiomatic expressions and colloquialisms" and use words "correct and appropriate" (vocabulary control) (2001)(pp.112,114). In other words, the collocation network alone cannot reveal the relationship between the group's collocation performance and their language competency, because collocation network cannot tell the above aspects, i.e., the naturalness and accuracy of the collocations.

While CRC with its quantitative ability (as used in Chapter 3) is a solution to the above problem. To examine whether the speakers' language competency truly matches their labeled level, we select the native speaker sub-group (NS) as a reference corpus, and respectively compute the CRC distances between NS with B1, B2, and C1/C2, so as to evaluate the usage patterns between different levels. Intuitively, we may expect the gap becomes smaller from B1 to C1/C2, because native speakers usually produce the most natural expressions.

The CRC distances of B1-NS, B2-NS and C1/C2-NS are shown in Table 3, including those of *take* and *do*. It can be seen that the speakers' usage pattern of *make* is approaching that of native speakers from B1 to C1/C2. However, statistics on the other two verbs do not present a similar trend; *do* even displays a totally opposite attitude, showing an increasing discrepancy from native speakers as "language level" rises. A possible explanation might be the insufficient data samples. For instance, only 30 common collocates are used to calculate the CRC distance between C1/C2 and NS, most of which are trivial words with low association strength such as *not*, *what*, and *want*.

| Distance | B1-NS | B2-NS | C1/C2-NS |
|---|---|---|---|
| make | 0.19 (48) | 0.1453 (101) | 0.1223 (93) |
| take | 0.2349 (32) | 0.1743 (49) | 0.2077 (45) |
| do | 0.0915 (116) | 0.1371 (66) | 0.1436 (30) |

Table 3: The collocation distribution distance of *make*, *take* and *do* in the three bands and NS corpus. The number of valid collocation words actually involved in the calculation is indicated in brackets (`measure=Log-Likelihood`, `min_freq=2`).

Nevertheless, the findings are basically in line with that of Brezina (2018) but in a more comprehensive and more precise manner. Besides, the distances on word pairs disclose the most misused collocates of the node, which might be helpful in language evaluation and grammar correction. To be specific, CRC could tell which collocates are most distantly distributed in the usage pattern of language learners and of native speakers, so as to improve the learners' worst-acquired collocation knowledge.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

## 6 Conclusions and Future Work

This paper re-examines two widely used representation methods of collocation, i.e., collocate list and collocation network. In view of their weakness in expressing contextual information, we propose a new representation method, namely the contextualized representation of collocation (CRC). CRC adopts conflated linear representation and highlights the importance of the position of the collocates. It pins a collocate as the interaction of two dimensions, i.e., association strength and co-occurrence position. With a full image of all the collocates surrounding the node word, CRC carries the contextual information and makes the representation much more informative and intuitive. We did three case studies to demonstrate the advantages of CRC in practical applications, covering synonym distinction, image analysis, and efficiency in lexical use. Besides, CRC provides a new quantitative tool to measure lexical usage pattern similarities for corpus-based research.

We believe that the potential power of CRC is far beyond the cases we have discussed. As an auxiliary corpus tool, it may also be used directly in teaching activities. The importance of corpus tools in language teaching is investigated by Boldarine and Rosa (2018), who used some of the searching functions provided by COCA, mainly the collocation tool, to improve students' phrasal integrity. Their survey shows that most students are happy to use corpus tools to test their language intuition, though the aids are less effective for students with poor performance. CRC can be used as a good visualized presentation tool for phrase, collocation and idiom studying, and should intuitively be more friendly to "weaker students" because it is much more easy-reading and more informative compared with collocate list and collocation network.

In addition to involving CRC in teaching, we can also extend or adapt its visualization to fit more needs and scenarios. For example, CRC is also suitable for visualizing constructions (Fillmore, 1988) and collostructions (Stefanowitsch and Rosa, 2003) because it follows the sequential nature of the language. For instance, a CRC-like visualization of the construction "It be ADJ that" could set the slot *ADJ* to position 0, and respectively fix *it*, *be*, *that* at the positions -2, -1, and 1. This requires the support of construction searching algorithms.

In summary, we hope that CRC can provide a new representation framework for language researchers and learners, and will lead them to address the importance of contextual information in research and learning. More applications of CRC in teaching and research are worthy of further empirical study in the future.

## References

Brezina, V., McEnery, T., and Wattam, S 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Brezina, V. 2018. Collocation graphs and networks: Selected applications. In *Lexical collocation analysis* (pp. 59-83). Springer, Cham.

Boldarine, A. C. and Rosa, R. G. 2018. Prepping a prep course: a corpus linguistics approach. *BELT-Brazilian English Language Teaching Journal*, 9(2), 379-394.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*, Cambridge University Press.

Fillmore, C. J. 1988. The mechanisms of "construction grammar". In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 14, pp. 35-55).

Firth, John R. 1957. *Modes of meaning. In: Papers in Linguisticss, 1934-1951*. Oxford: Oxford University Press.

Gablasova, D., Brezina, V., Mcenery, T. and Boyd, E. 2017. Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, 38(5), 613–637.

Golding, W. 1954. *Lord of the flies*. Faber & Faber.

Halliday, M. A. K. and Hasan. R 1976. *Cohesion in English*. London: Longman Group Ltd.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.

Katz, J. J. and Fodor, J. A. 1963. The Structure of a Semantic Theory. *Language*, 39(2), 170–210. https://doi.org/10.2307/411200.

Koteyko, N., Jaspal, R. and Nerlich, B. 2013. Climate change and 'climategate' in online reader comments: A mixed methods study. *The geographical journal*, 179(1), 74-86.

Liu, D. 2010. Is it a chief, main, major, primary, or principal concern?: A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics*, 15(1), 56-87.

Liu, D. and Espino, M. 2012. Actually, Genuinely, Really, and Truly: A corpus-based Behavioral Profile study of near-synonymous adverbs. *International Journal of Corpus Linguistics*, 17(2), 198-228.

Liu, D. 2018. A corpus study of Chinese EFL learners' use of circumstance, demand, and significant: An in-depth analysis of L2 vocabulary use and its implications. *Journal of Second Language Studies*, 1(2), 309-332.

Oktavianti, I. N. and Sarage, J. 2021. Collocates of 'great' and 'good' in the Corpus of Contemporary American English and Indonesian EFL textbooks. *Studies in English Language and Education*, 8(2), 457-478.

Oldsey, B. and Weintraub, S. 1963. Lord of the Flies: Beezlebub Revisited. *College English*, 25(2), 90–99. https://doi.org/10.2307/373397.

Pan, F. and Hei, Y. 2017. Government Image-building in Chinese-English Press Interpretation: A Case Study of the collocation of Personal Pronoun "we". *Foreign Languages and Their Teaching* (5), 8.

Petruck, M. R. 1996. Frame semantics. *Handbook of pragmatics*, 2.

Phillips, M. 1985. *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam, Netherlands: North-Holland.

Qu, W. 2008. *Research on Automatic Word Disambiguation of Modern Chinese*. Beijing: Science Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Spitz, D. 1970. Power and Authority: An Interpretation of Golding's "Lord of the Flies." *The Antioch Review*, 30(1), 21–33. https://doi.org/10.2307/4637248.

Stefanowitsch, A. and Gries, S. T. 2003. Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2), 209-243.

Tang, X. 2018. *Collocation and Predicate Semantics Computation*. Wuhan University Press.

Taner C. and Hakan C. 2021. A warring style: A corpus stylistic analysis of the First World War poetry. *Digital Scholarship in the Humanities*, fqab047, https://doi.org/10.1093/llc/fqab047.

Vickers, B. 2012. Identifying Shakespeare's additions to The Spanish Tragedy (1602): A new (er) approach. *Shakespeare*, 8(1), 13-43.

Wang, D., Zhang, D., Tu, X., Zheng, X. and Tong, Z. 2007. Collocation Extraction Based on Relative Conditional Entropy. *Journal of Beijing University of Posts and Telecommunications*, 30(6), 40.

Wijitsopon, R. 2013. A corpus-based study of the style in Jane Austen's novels. *Manusya: Journal of Humanities*, 16(1), 41-64.

Xiong, Y. H. and Liu, D. F. 2022. A Corpus-based Analysis of English Near-synonymous Adverbs: Absolutely, Utterly. *Journal of Literature and Art Studies*, 12(4), 359-365.



Proceedings of the 22nd China National Conference on Computational Linguistics, pages 836-846, Harbin, China, August 3 - 5, 2023.