Development of BES Search: A Primary and Secondary School Level Pedagogical English Example Sentence Corpus and Search Tool

NISHIGAKI Chikako[1]＊, MIZUMOTO Atsushi[2] and AKASEGAWA Shiro[3]

[1]Faculty of Education, Chiba University
[2]Faculty of Foreign Language Studies, Kansai University
[3]Lago NLP

入門・初級レベル英語コーパスの構築と検索ツールの開発

西垣知佳子[1]＊・水本　篤[2]・赤瀬川史朗[3]

[1]千葉大学・教育学部
[2]関西大学・外国語学部
[3]Lago NLP

　Using data-driven learning（DDL）in foreign language instruction has become increasingly prevalent in recent years. However, few DDL applications target elementary, middle, and high school students at introductory and beginner levels. One of the reasons for this shortage is the absence of corpora with appropriate difficulty tailored for these learners. In this paper, we introduce the development and release of the Basic English Sentence（BES）Search, which is a combination tool of level-appropriate English sentences for elementary to high school levels and a search mechanism. This search tool is an application integrating a corpus exclusively extracted from copyright-free English sentences from the precursor corpus called the Pedagogical Sentence Corpus（PSC）（Nishigaki & Akasegawa, 2020）and a uniquely crafted search function. The BES Search allows users to customize searches for English example sentences according to the length of sentences, grammatical components, lemmas, and phrases, ensuring alignment with learning objectives. Additionally, BES Search is designed with consideration for users who may not be familiar with corpus studies, facilitating accessibility. Sentences extracted through the BES Search can be used in lesson preparation, DDL instruction, and the development of DDL applications without the concerns of copyright infringement, enabling targeted English sentence searches and utilization for specific objectives.

　近年，外国語学習において，DDLの活用が盛んになっている。しかし，小学生から中学・高校生までの入門・初級レベルを対象とした　DDLの活用事例は非常に少ない。その原因のひとつとして，当該学習者のレベルに適切な難易度のコーパスが存在しないことがあげられる。そこで初等・中等教育用に調整された英文を集めたコーパスと検索ツールを組み合わせたBES（Basic English Sentence）Searchを開発し，公開した。本検索ツールは，我々が開発した既存のPedagogical Sentence Corpus（PSC）Search（Nishigaki & Akasegawa, 2020）から，著作権フリーの英文のみを抽出して構築したコーパスと，独自開発の検索ツールを組み合わせたアプリで，無料で，登録不要で使用できる。BES　Searchは，英文の長さ，文法項目，レマ，語句の検索条件を設定することにより，学習ターゲットに合わせて，英語例文を検索でき，コーパスに関する知識のないユーザが利用できるように配慮されている。BES Searchによって抽出された英文は，授業準備，DDL指導，DDLアプリケーションの開発に著作権を気にすることなく，目的にあった英文を検索し，利用できる。

キーワード：introductory and beginner-level learners（入門・初級レベル学習者），
　　　　　　independent sentence corpus（用例コーパス），pedagogical corpus（教育用コーパス），DDL

1.　Background

　Digitization has revolutionized the compiling and analyzing of electronic corpora, distinguishing sharply from manual text processing in earlier times（McEnery & Wilson, 2001）. As corpora became more promi-nent in linguistics, their applications expanded beyond research to adopt language education. Within English Language Teaching（ELT）, corpora have significantly enriched language learning for intermediate and advanced students by showcasing authentic language use（Conrad, 2000）. One prominent approach is data-driven learning（DDL）, where learners engage with a corpus directly, employing tools like concordancers

＊連絡先著者：西垣知佳子　gaki@faculty.chiba-u.jp

to explore and derive vocabulary and grammatical rules (Johns, 1991).

Meta-analyses have underscored DDL's superior efficacy compared to other instructional methods (Boulton & Cobb, 2017; Mizumoto & Chujo, 2015). Typically, DDL employs corpora such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), which are rich repositories of authentic language from adult native speakers.

However, while corpora are invaluable for more advanced learners, their authentic features, such as long sentences, vocabulary for specific purposes, and intricate grammar structure, can overwhelm beginners (see Figure 1). Hence, DDL's use for beginners remains limited (Crosthwaite, 2020). The efficacy of DDL depends on the appropriateness of the chosen corpus to the learner's proficiency level (Chujo & Oghigian, 2012; Sinclair, 2005). The gap between the corpus and the learner's proficiency level underscores the need for beginner-tailored corpora that align with Krashen's i+1 principles (Krashen, 1982), ensuring that linguistic complexity matches the learner's proficiency.

SCoRE (Sentence Corpus of Remedial English (Chujo et al., 2015) is an illustrative example of a tailored beginner-level corpus. Designed specifically for remedial-level university students, SCoRE is comprised of independent sentences, each meticulously crafted to suit ESL learners by considering factors such as sentence length, vocabulary, grammar, and cultural context. SCoRE was the inspiration for an educational English example sentence corpus for primary and secondary school levels, called the Pedagogical Sentence Corpus (PSC) Search. However, because the sentences contained in the PSC were protected by copyright, a new corpus (and search tool) was created called Basic English Sentence (BES) Search. In this paper, we provide an overview of the development processes of PSC and BES and detail the proficiency level and unique functions and features of BES Search.

2. The Development of BES Search

2.1 PSC (Pedagogical Sentence Corpus) Search

PSC Search (Nishigaki & Akasegawa, 2020) combined a corpus of beginner-level English sentences with a functional search application. It empowered users to locate English sentences by specifying a grammatical structure, phrase, or part of speech. It incorporated a corpus of isolated English sentences, tailored to align with the linguistic development of Japanese students from elementary through high school (spanning the third to 12th grade) with the acknowledgment that formal English education commences in the third grade in Japan. It comprised three distinct linguistic data types, as shown in Table 1. The first type included English sentences derived from government-authorized textbooks used in elementary, junior, and senior high schools across Japan, China, Korea, and Taiwan (Nishigaki et al., 2011) in addition to commercial English texts for elementary students and U.S. textbooks for reading and language arts. The second type was amassed from extensive reading texts, like those found in the Oxford Reading Tree and Penguin Readers. Lastly, the third type of data was collected from child-friendly websites, such as *Tatoeba*, *Science News for Students*, and *News in Levels*. An overview of the compiled database is illustrated in Table 1.

PSC Search has English sentences for a wide range of users, from elementary students to introductory college levels. However, due to copyright issues, PSC Search cannot be used in its original form. Therefore, BES Search was developed by compiling a corpus from copyright-free English sentences extracted from PSC Search and equipping it with search software. Details regarding the language data incorporated into BES Search are shown in Table 2. Among the source

Figure 1 *Screenshot of COCA Search Results for Green*

Table 1 *PSC Search Sources*

Source Data	No. of Words	No. of Sentences
1. Government-authorized textbooks	736,703	88,404
2. Books for extensive reading	300,339	41,753
3. Children's websites	22,990,173	2,009,209
Total	24,027,215	2,139,366

Table 2 BES Search Corpus Sources

Source Data	No. of Words	No. of Sentences
1. SCoRE	136,914	15,628
2. hDDL	6,942	1,227
3. eDDL	22,061	4,364
4. *Tatoeba*	10,578,938	1,315,915
Total	10,744,855	1,337,134

data in Table 2, SCoRE is a DDL application developed for university students, as was explained in the previous section. hDDL and eDDL are DDL applications developed by the authors (Nishigaki et al., 2020; Nishigaki & Akasegawa, 2023), with copyright-free corpora. *Tatoeba* is a free collection of example sentences with translations for foreign language learners.

2.2 Indexing Corpus Files for Using the Search Engine

BES Search employs the Blacklab Query Tool (http://inl.github.io/BlackLab/) as its search engine to query corpus data efficiently. The Blacklab Query Tool is a search engine specializing in corpus searches and can utilize corpus query language (CQL), a standard query language for corpus searches. To utilize this search engine, it is first necessary to create an index file. The general procedure for this is as follows:

1. The collected English sentences were segmented and the number of words in each sentence and its source were added (see Figure 2; the number of words appears in the far left column and the source

appears in the third column).
2. The Stanford Log-linear Part-Of-Speech Tagger was used to assign part-of-speech tags.
3. The file was converted into an XML format compatible with the Blacklab Query Tool (SketchEngine format).

3. Proficiency Level of BES Search

This section clarifies the proficiency level of sentences within the BES Search, focusing on sentence length and vocabulary level. Sentence length significantly correlates with its complexity; shorter sentences typically possess simpler structures, while longer ones introduce complexity (Yano, 1994).

3.1 Sentence Length

In order to evaluate the appropriateness of the sentences in the corpus, first, the BES sentences were examined by sentence length. Every sentence was classified based on the number of words it contained, and the proportion was calculated. The results are present-

0005	A dog attacks the boy.	http://www.newsinlevels.com/products/cat-saves-a-boy-level-1
	Cat saves a boy - level 1	
0005	A dog barks at me.	HTR:houghton-mifflin-english-gr1-3
0005	A dog barks at strangers.	TTB:4451449
0005	A dog bit her leg.	TTB:1047232
0005	A dog bit his leg.	TTB:4728052
0005	A dog chased a cat.	TTB:945575
0005	A dog gets into canal.	http://www.newsinlevels.com/products/dog-in-a-canal-level-1
	Dog in a canal – level 1	
0005	A dog has four legs.	TTB:239195
0005	A dog has four paws	TTB:6392737
0005	A dog is barking now.	TTB:4237245
0005	A dog is following us.	TTB:5222028
0005	A dog is in danger.	http://www.newsinlevels.com/products/firefighters-save-a-dog-level-1
NIL	Firefighters save a dog - level 1	
0005	A dog or a cat?	AST:24-one-world1f
0005	A dog ran after it.	HTR:houghton-mifflin-english-gr1-3

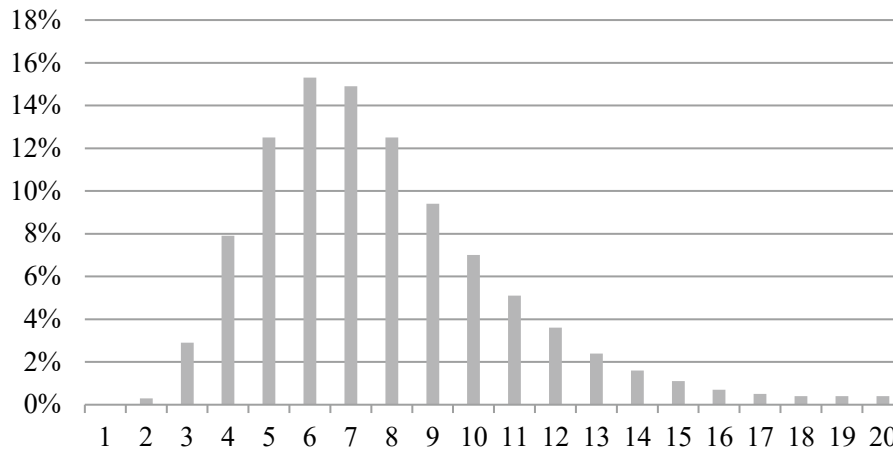Figure 2　Example of a Corpus Index File

Figure 3　*Sentence Length of BES Search*

Table 3　*Transition of Sentence Length*

Grade	The first part of the first lesson	The last part of the last lesson
7th	3.25	5.60
8th	6.21	8.07
9th	7.75	12.90
10th	9.33	9.31
11th	10.45	14.75

ed in Figure 3. Sentences ranged from one-word imperatives like "jump" to those with 20 or more words that concatenate multiple sentences using conjunctions such as "and," "that," and "when."

Next, we examined the length of English sentences used in school textbooks (see Table 3). The textbooks used in junior high school (grades 7 through 9) are at about the same level, regardless of the specific textbook chosen. Therefore, we focused on the most commonly used textbooks in Japan. On the other hand, the level of textbooks used in high school (grades 10 through 12) varies from advanced to elementary, depending on the textbook. Consequently, we selected one intermediate-level textbook for this study. We scrutinized the initial and final reading sections of these texts' first and last lessons, detailing the results in Table 3. It is noteworthy that 12th graders, due to an ongoing revision in the Course of Study, still employ an older edition, which was excluded from our analysis. Our focus remained on the textbook used in a class titled *English communication* for first and second-year senior high school students, in line with the new Course of Study. Table 3 suggests that junior high school students frequently encounter sentences of 13 words or fewer in their textbooks. Similarly, BES Search offers a substantial number of these concise English sentences, making it a suitable resource for this age group.

Once the BES corpus was modified with original

(noncopyrighted) sentences and index files were created, the PSC Search functions were applied with some modifications. Once these were completed, the BES Search became a corpus and tool that (a) allows users to search for educationally written English sentences, adjusted for educational use at beginner and elementary levels, by specifying grammatical structures and parts of speech; and (b) presents short, complete English sentences, not fragments of longer English sentences (called concordance lines), as often happens in authentic, more advanced level corpora. Although it was primarily envisaged as a tool for Japanese instructors of English, learners can also use it, and because it has both Japanese and English versions, it can be used by international teachers and students. It is also available without cost or the need for registration. BES Search can be accessed at https://bessearch.ddl-study. org/.

3.2　Vocabulary Level

We next examined the lexical difficulty of BES Search, employing the New Word Level Checker (NWLC; Mizumoto, 2022) for this purpose. The analysis was conducted using the New JACET List of Basic Words (New JACET8000; JACET, 2016). The results are displayed in Table 4, where "Frequency" refers to the total number of words, and "Type" indicates the count of unique words, both categorized by level.

We categorized proper nouns and numbers—represented as "PropNoun_Num" in Table 4—as "known words." We then assessed the cumulative vocabulary coverage rate. According to the current Course of Study guidelines in Japan, high school graduates should have mastery over 4,000 to 5,000 words. Our findings indicate that the cumulative coverage rate reached approximately 98% (97.75%) at Word Level 5.

According to the Course of Study guidelines, junior high school graduates should be familiar with 2,200 to 2,500 words, broken down as 600+1,600 or 700+1,800,

respectively. Our analysis of Word Levels 2 and 3 indicates that the cumulative coverage rate is approximately 95% (94.37%) or higher (96.17%). These findings suggest that the vocabulary level in BES Search is highly suitable for both junior and senior high school students. Additionally, a review of the "Type" column

shows that nearly all the words from levels L1 to L8 of the New JACET8000 list are included. This indicates that BES Search offers a high level of learning efficiency, as users are exposed to almost the entire range of words in the New JACET8000 list.

Figure 4 graphically illustrates the data from Table 4 showing the vocabulary coverage of each JACET8000 level in BES Search. The graph confirms that the 2,000 to 3,000-word range has more than 95% coverage, and the 5,000-word level shows nearly 98% coverage.

The left column in Figure 5 lists the ten most frequent words in BES Search. Interestingly, "the," which is usually the most common word in standard English texts, only ranks 7th in BES Search. This pattern is in line with trends sometimes observed in elementary and junior high school English textbooks, where "the" is not always the top word. This suggests that BES Search focuses on sentences that are easy to understand and tailored for educational use, rather than featuring complex, natural English. In other words, it offers examples ideal for beginner-level DDL, making it
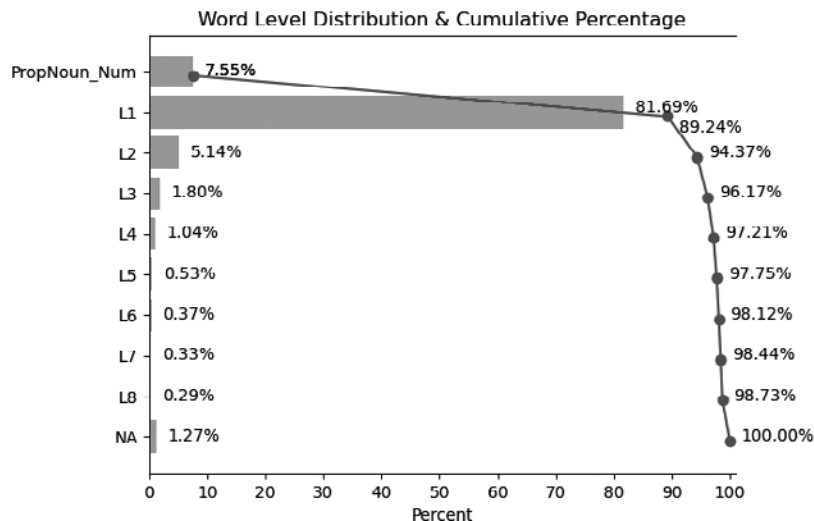
Table 4　Results of Word Level Analysis

Word Level	Type	Frequency	%	Cumulative
PropNoun_Num	13345	823862	7.55	7.55
L1	1002	8915915	81.69	89.24
L2	1000	560658	5.14	94.37
L3	1000	196461	1.8	96.17
L4	1000	113329	1.04	97.21
L5	1000	58096	0.53	97.75
L6	997	40467	0.37	98.12
L7	991	35863	0.33	98.44
L8	979	31528	0.29	98.73
NA	22128	138224	1.27	100.0
ALL	43442	10914403	100	100

Figure 4　Visual Representation of Vocabulary Coverage Rate

	Word	Frequency
1	be	588581
2	to	392061
3	tom	388934
4	I	379140
5	you	374670
6	do	343783
7	the	331346
8	not	291767
9	that	271715
10	a	229117

	Word	Frequency	Lemma Word Form(s)	Level
976	giraffe	763	giraffes (419), giraffe (344)	NA
977	bored	763	bored (763)	NA
1038	o'clock	705	o'clock (705)	NA
1291	separatist	515	separatists (372), separatist (143)	NA
1359	cooky	473	cookies (473)	NA
1487	annoyed	415	annoyed (415)	NA
1630	ex	363	ex (354), exes (9)	NA
1647	amused	357	amused (357)	NA
1661	hijab	353	hijab (353)	NA
1794	terrified	310	terrified (310)	NA

Figure 5　Most Frequent Words and Top Words Outside New JACET8000

suitable for classroom use.

The right side of Figure 5 lists the top 10 words not included in the New JACET8000 list. While these words are absent from New JACET8000, they encompass everyday, simple terms like "giraffe," "bored," and "o'clock," as well as theme-specific words such as "hijab." This indicates that BES Search also includes a variety of vocabulary that extends beyond lists like the New JACET8000. These additional words, which range from everyday terms to theme-specific vocabulary, offer a broader linguistic exposure. This diversity can be particularly useful for students, as it introduces them to words they are likely to encounter in everyday conversations or in specific contexts, thereby enriching their overall language comprehension and usage skills.

4.　BES Search Tool Features and Functions

4.1　Features of BES Search

BES search tool is characterized by its capacity to enable efficient and exhaustive searches for examples containing specific grammatical items, using the CQL language described in section 2.2, not only at the level of surface forms such as phrases, but also by combining lemmas, parts of speech, and word count in the search expressions. The features include:

Phrase Search: Users can specify words or phrases contained in the sentences they wish to explore. For example, to search for English sentences in the form of "I am from....," each word should be enclosed in straight double quotations like "I" "am" "from."

Lemma Search: This feature can be utilized to search for verbs in various forms, plural nouns, and comparative and superlative adjectives collectively.

For instance, specifying [lemma="go"] would encompass "go" and its varied forms（goes, went, gone, going）.

Part of Speech Search: For example, if [lemma="talk" & pos="N."] is specified, it designates the noun "talk" and its variations, while [lemma="talk" & pos="V."] specifies the verb "talk" and its variations.

Word Count Specific Search: Users can specify the number of words intervening between the searched words. For example, specifying [lemma="go"] "to" []{0,3} "school" will search for examples where 0-3 words are interposed between "go to" and "school."

Moreover, users can specify the number of words in English sentences to expedite efficient example generation during the search. Additionally, regarding the result display, it is arranged to present shorter English sentences first. Given that the length of English sentences（word count）is a crucial factor in adjusting difficulty（Yano et al., 1994）, arranging result displays with fewer words first serves as a metric when extracting less difficult English sentences and proves to be an effective display method for English example generation.

4.2　Functions of BES Search

Figure 6 illustrates the interface of the example search tool. The interface is divided into two panels: the left panel, where search conditions are set, and the right panel, where results are displayed. The specific functionalities are described below.

4.2.1　Search Conditions

In Figure 6, marker ❶ indicates where search conditions are set. The tool specifies search queries using
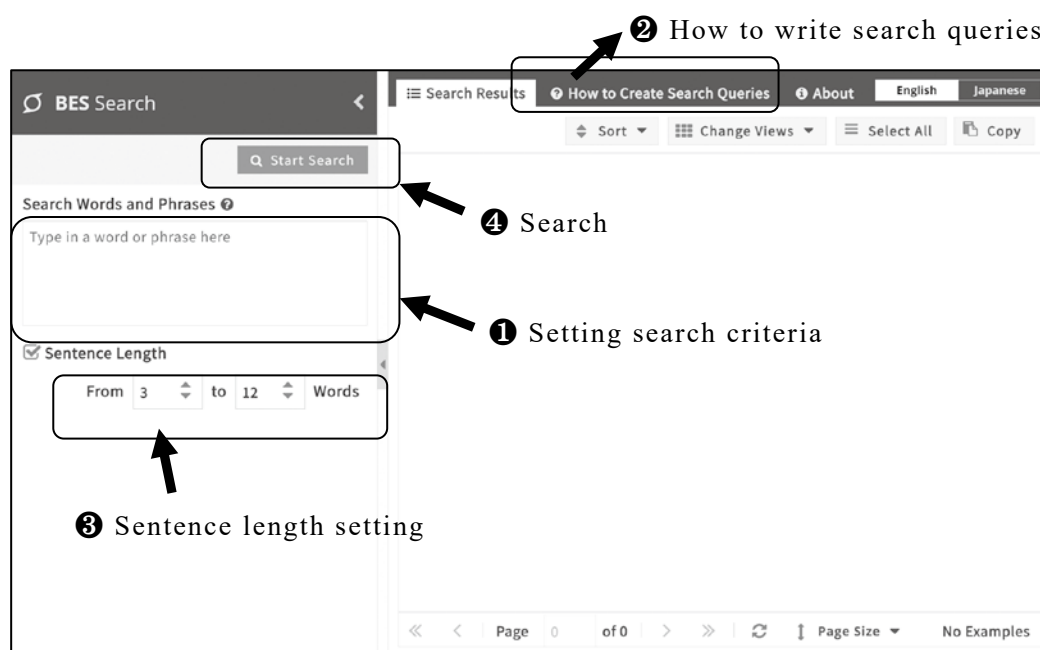

Figure 6　*Functions of BES Search*

CQL. For instance, to search for infinitives or present perfect verb forms, one would use:

For infinitives (e.g., "want to do"): [lemma="want"] "to" [pos="VB"]

For present perfect form: "have|has" [pos="RB"]{0,1} [pos="VBN"]

Notably, as initiating a search with CQL can be challenging, convenience is considered by providing a guide to writing search expressions at marker ❷ (see Figure 6), enabling quick reference.

4.2.2 Sentence Length

At marker ❸ in Figure 6, the tool allows users to set sentence length by specifying the minimum and maximum word counts in a sentence, such as "3 words minimum, 7 words maximum." Providing beginners with shorter example sentences, which contain fewer words, allows for easier comprehension of sentence structure. Thus, restricting sentence length during a search can efficiently facilitate material creation for beginner and introductory levels. For example, when learners study sentences of the form "I like…," if one searches with ❸ set to the default "1 to 20 words," 3,004 sentences are found. If one desires the simplest possible English sentences that do not require the use of articles, such as "I like soccer," "I like red," or "I like pizza," setting the search to "3 words to 3 words" narrows down the search results to 243 sentences, making it easier to find desired example sentences to show learners. This ability to specify English sentence length can be considered a considerate and innovative feature suitable for beginner and introductory English education levels, and it is a practical functionality not commonly found in general search tools.

After setting the search conditions and defining the sentence length, pressing the search button at marker ❹ will display the search results in the right panel.

4.2.3 Sorting

Figure 7 displays the results of a search for "enjoy + verb in -ing form." The search results in the right panel are shown in order of sentence length, starting with the shortest. In Figure 7, sentences consisting of three words are lined up.

Upon pressing ❺ in the right panel of Figure 7, users can sort the list of displayed English sentences by "Keyword Left," "Keyword Right," "Keyword," and "Appearance Order." Selecting "Keyword Left" allows the user to sort and display alphabetically based on the phrase appearing to the left of the searched keyword, while "Keyword Right" sorts based on the phrase to the right. Sorting by "Keyword" arranges the sentences according to the phrase specified in the search, and "Appearance Order" returns the list to the initial order displayed right after searching.

4.2.4 Changing Views

By utilizing ❻ in Figure 7, users can choose how English sentences are displayed on the terminal screen, selecting between "Sentence View" and "KWIC View." Figure 7 demonstrates the KWIC View, where the keyword is placed in the center of the displayed sentence. Selecting Sentence View will present the English sentence left-aligned, adhering to conventional English display formats.
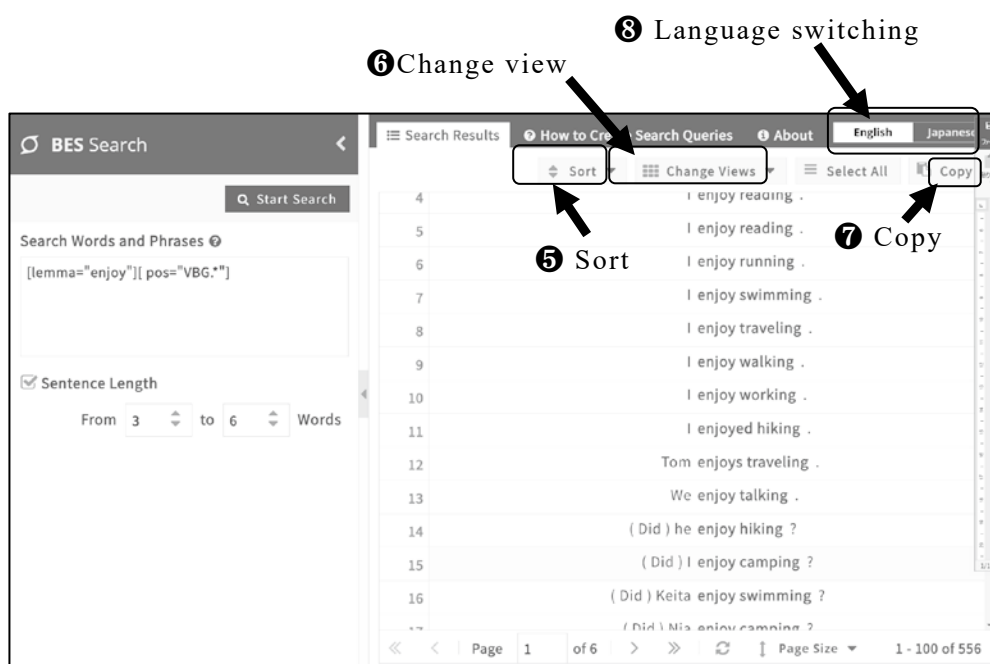

Figure 7 *Screenshot of BES Search Result Screen*

4.2.5　Copy

Pressing ❼ in the right panel of Figure 7 allows users to copy the English sentences from the search results. Using the copy functionality facilitates the efficient creation of educational materials.

4.2.6　Switching Languages

The language used can be switched between Japanese and English by pressing the ❽ button.

5.　Conclusion and Implications

Corpora are pivotal for linguistic research and offer tremendous utility in educational contexts. Nonetheless, while valuable, existing corpora such as the PSC Search contain English sentences for beginners and elementary-level learners have copyright restrictions that preclude their public release and unrestricted usage. Consequently, this study set out to compile the BES Search, which is accessible and free for anyone to use, and has publicized it under the friendly name of BES Search. This paper reported on developing and releasing a use-case search tool designed to search through English sentences tailored for beginners and elementary-level learners.

The BES Search tool enables efficient searching of example sentences that match the learner's level and learning objectives by allowing users to set parameters like sentence length, grammatical items, and phrases. Given that the difficulty of English sentences presented to beginners and elementary learners is significantly influenced by sentence length, this use-case search tool can extract and display examples based on sentence length, a unique feature not observed in other corpus search tools. By utilizing this use-case search tool, it is anticipated that the workload associated with creating DDL materials for beginners and elementary learners will be reduced, thereby streamlining lesson preparation.

BES Search tool opens avenues for further research into its actual applicability and efficacy in pedagogical contexts. Investigating its usage in real classroom scenarios, gathering feedback from educators, and subsequently iterating on the tool would be pivotal future steps. This endeavor to construct and disseminate a freely accessible corpus and search tool not only stands as a resource for educators and learners but also as a foundation for further language learning and teaching innovations.

Acknowledgments

References

Baker, P. (2006). *Using corpora in discourse analysis.* Continuum.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67,* 348–393.

British National Corpus (BNC). https://www.english-corpora.org/bnc/

Chujo, K., & Oghigian, K. (2012). DDL for EFL beginners: Recent gains and student views on the role of L1 and paper-based concordancing. In J. Thomas & A. Boulton, (Eds.), *Input, process and product: Developments in teaching and language corpora,* Masaryk University Press, 169–182.

Chujo, K., Oghigian, K., & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial EFL learners, In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven Learning.* John Benjamins, 109–128.

Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*(3), 548–560.

Crosthwaite, P. (2019). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners.* Routledge.

Japan Association of College English Teachers (JACET). (2016). *The new JACET list of 8000 basic words.* Kirihara Shoten.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), Classroom Concordancing. *English Language Research Journal, 4,* 1–16.

Krashen, S.D. (1982). *Principles and practice in second language acquisition.* Pergamon.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics.* Edinburgh University Press.

Mizumoto, A. (2022). New Word Level Checker no gaiyou [The overview of New Word Level Checker]. *Reports Vol. 12 of Japan Association for Language Education and Technology* (LET)*, Kansai Chapter, Methodology Special Interest Group* (SIG), 22–43. https://doi.org/10.31219/osf.io/whr9a

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom, English Corpus Studies, 22, 1–18.

Nishigaki, C., & Akasegawa, S. (2020). Sakurei sanshō-yō sōsukōpasu no kōchiku to yōrei kensaku tsūru no kaihatsu [Construction of source corpus for example reference and development of example search tool], *Bulletin of the Faculty of Education, Chiba University, 68,* 159–164.

Nishigaki, C., & Akasegawa, S.（2023）. Development and revision of DDL tools for secondary school students: What we can do to nurture autonomous corpus users? *English Corpus Studies, 30*, 131–149.

Nishigaki, T., Amano, K., Yoshimori, T., & Nakajo, K.（2011）. An attempt at developing data-driven English learning materials using concordance lines for middle and high school students. *Chiba University Faculty of Education Research Bulletin, 59*, 235–240.

Nishigaki, C., Hoshino, Y., Abe, T., Kamiya, N., Koyama Y., & Ishii Y., Shougakusei no tame no data kudou-gata eigo gakushuu shien saito no kaihatsu to koukai, [Development and publication of a data-driven English learning support site for elementary school students], *JES Journal, 20*, 367–382, 2020.

Sinclair, J.（2005）. Corpus and text—Basic principles. In M. Wynne（Ed.）, *Developing linguistic corpora: A guide to good practice*（pp. 1–16）. Oxbow Books.

Yano, Y., Long, M.H., & Ross, S.（1994）. The effects of simplified and elaborated texts on foreign language reading comprehension, *Language Learning, 44*（2）, 189–219.

〈Data Sources〉

News in Levels https://www.newsinlevels.com/（retrieved October 15, 2023）

Oxford Reading Tree Series, Oxford University Press.

Penguin Readers, Pearson.

Science News for Students https://www.sciencenewsforstudents.org/

Tatoeba https://tatoeba.org/jpn（retrieved October 15, 2023）