Corpus Search in Life Science Dictionary (LSD) as a Tool for Writing Scientific Papers

Takeshi Kawamoto[1,5], Nobuyuki Fujita[2,5],
Hiroshi Ohtake[3,5], and Shuji Kaneko[4,5]
[1]*Hiroshima University,* [2]*National Institute of Technology and Evaluation,*
[3]*Fukui Prefectural University,* [4]*Kyoto University,* [5]*The Life Science Dictionary Project*

Writing academic papers in English is an essential skill even for non-native English speakers, including Japanese researchers. The corpus in the Life Science Dictionary (WebLSD Corpus) is a powerful tool that can extract useful information about correct or typical usages of English words from the LSD corpus. For example, *expression*, which usually means "gene expression" in the life science field, is used as an uncountable noun, which can be clarified by the WebLSD Corpus. Since this usage is specific to life sciences, it is necessary to check its usage in scientific papers. *Associate* can be used as both a transitive verb and an intransitive verb, and the meaning of *associate* differs depending on its usage pattern. *Insight* is usually used as both an uncountable noun and a plural form of a countable noun, but not in the singular form. These unexpected usages can be verified by using the WebLSD Corpus. To find critical information, we need to select the proper search options in the WebLSD Corpus. In this article, we show how to search for verbs, nouns, and irregular verbs using the WebLSD Corpus to enable researchers to find a variety of useful information on discipline-specific English without seeking help from native English speakers.

1. Introduction

Most Japanese researchers have to write scientific papers in English to be recognized in the scientific community, despite being non-native English speakers. In order to help these researchers, we launched an online dictionary service called Life Science Dictionary (WebLSD, https://lsd-project.jp/), which includes the LSD Corpus (WebLSD Corpus) (Kaneko et al., 2003). The WebLSD Corpus is an excellent tool, enabling researchers to find information about typical usages of English in scientific papers.

English and Japanese grammar is fundamentally different, especially in terms of verbs. Japanese follows a functional structure, while English follows a constituent structure; Japanese verb placement is verb-final, while English is verb-medial; Japanese is a pronoun-dropping language, but English is not; Japanese verbs are constructed differently with tense contained within the word's structure, which is not the case with English (Dalrymple, 2001).

Traditional dictionaries allow users to look up the correct form of the verb, and while discipline-specific dictionaries such as medical dictionaries do exist, they are of limited use to Japanese non-native English speaking researchers. Such dictionaries do not address other issues, such as verb placement, or whether pronouns should be used and which pronouns should be used. The grammar used in scientific writing is also different from general use, employing passive forms, and is very lexically dense, so even dictionaries that offer extensive lists of examples are of limited use to Japanese academics writing scientific papers. A final consideration is that styles change, which means that Japanese scientists require access to a resource that reflects the contemporary usage of words (Crystal & Davy, 1973; Halliday &

NU Ideas Volume 6. 2017. Nagoya University Institute of Liberal Arts & Sciences
Special issue: *Proceedings of the Third International Symposium on Academic Writing and Critical Thinking*

Martin, 1993). Non-native speakers tend to focus on acquiring information about the meaning of words without considering usage patterns or collocations, resulting in the misuse of words (Kawamoto et al. 2004). Sometimes, a word used in scientific papers such as *expression* has a distinctive meaning specific to the research field. More than 95% of usages of *expression* in the corpus mean "gene expression," while about 2% of occurrences refer to "facial expression(s)." Although the meaning of words used in scientific papers is very strict, some words can have multiple meanings. For example, the word *associated* has two meanings, "related" and "bound," as discussed in Section 3.

For the reasons stated above, the WebLSD Corpus was constructed (Kaneko et al., 2003) to analyze collocation patterns in scientific papers (Kawamoto et al., 2004; Kawamoto et al., 2005; Ohtake et al., 2006, 2007, 2008). It presents words in context, allowing students, researchers, and translators in the field of life science to find examples of how words are currently being used in the life sciences, complete with grammatical context. Users can ascertain not only in what form the verb should be used, but also with what pronouns, where in the sentence it should be placed, and so forth.

2. Search options in the WebLSD Corpus

The WebLSD Corpus has various search options as shown in Figure 1. Initial settings for the search options are shown in boldfaced type. Combinations of certain search options provide different types of information. The first line in the option panel is the selection for search types. "Sort by left" means that the sentences obtained in the search result are sorted in alphabetical order according to the left word of the query. "Statistics" provides the frequency of words collocating with the query. The second line concerns the selection of whether the result window will stay in the current window or open a new window. The third line is for selecting the number of sentences shown in the result window. The fourth line is the selection of the number of characters per line shown in the result. The fifth line is the option for case sensitivity; "insensitive to case" means that the search does not distinguish between upper case letters and lower case letters. The sixth line is the option for inflection of the query word. The last line provides the option to sort the sentences given in the results according to the alphabetical order of the query word or the word to the right word, when choosing "sort by right." This article provides information about the different kinds of data the WebLSD Corpus can provide and how to operate the WebLSD Corpus by selecting search options.

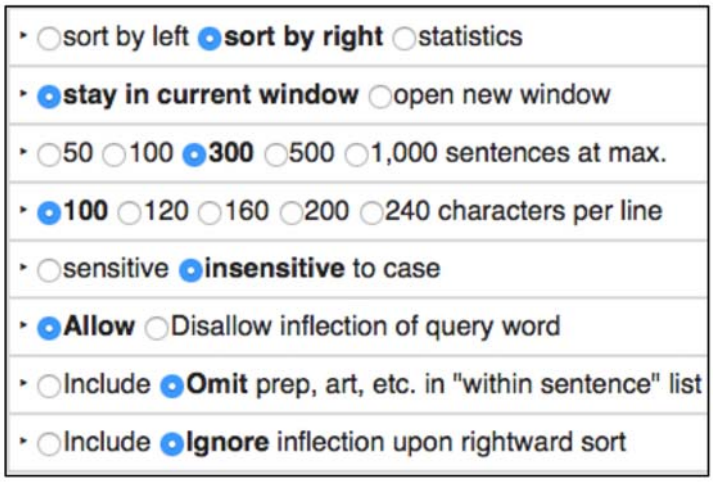

Figure 1. Search options. Boldfaced types in these options show initial settings.

3. Searching for verbs

To use English verbs properly, one has to learn the difference between a transitive verb and an intransitive verb, the construction of passive forms, and combinations of verbs and prepositions. For example, *associate* is one of the most frequently used verbs in scientific papers. Figure 2 shows the search result for *associate* sorted by left. The option was changed from the initial setting of 300 sentences to 50 for the reader's convenience. The search result, consisting of 50 sentences, is easy to look at. By focusing on two boxed areas, the two typical usages of *associate* are estimated. The first is "noun-associated noun," such as *HPV-associated tumors* or *ribosome-associated proteins*. *HPV-associated tumors* refers to tumors that are related to HPV. However, *ribosome-associated proteins* means that the proteins are bound to the ribosome. Thus, *associated* has two meanings: "related" and "bound."

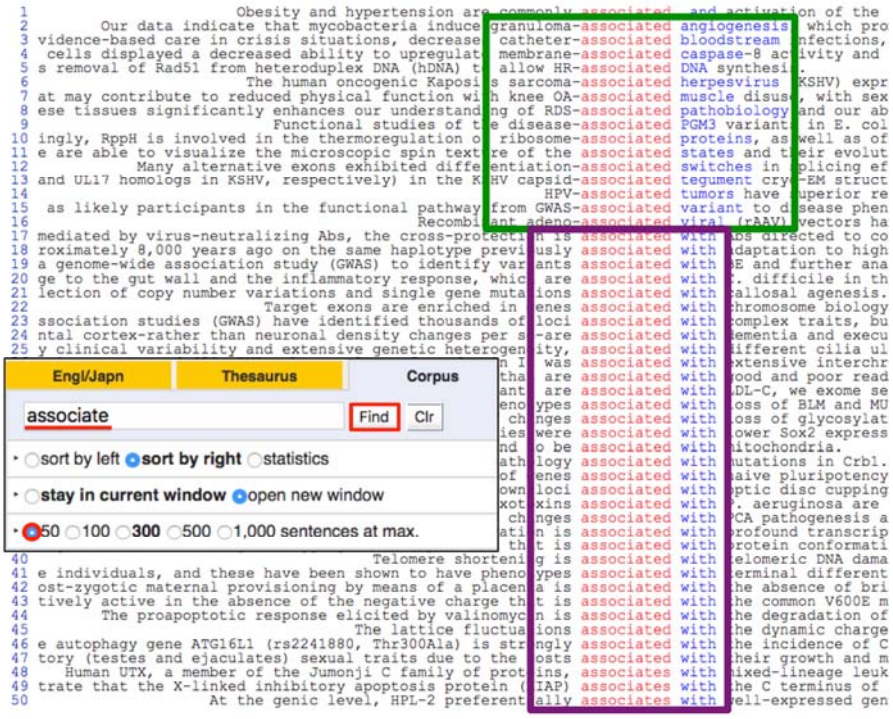

Figure 2. Search results.

Second, Figure 2 shows *associated* or *associates*, and indicates that the preposition *with* should probably be used after *associate*, because more than 60% of the occurrences of *associate* collocate with the preposition *with*. This rule can be confirmed by the "statistics" section in which 635 out of 1,000 occurrences of *associate* collocate with the preposition *with*.

As suggested by the data in Figure 2, *associate* can be used as both a transitive verb and an intransitive verb. The existence of *associates with* as a present tense verb with a preposition indicates that *associate* is an intransitive verb. In contrast, the search of *associated with* shows that in more than 99% of occurrences, *associated* is used in the passive form of a transitive verb such as *be associated with*.

In scientific papers, the usage of words should be simple and specific to avoid confusion. In the case of *associate*, however, different usages exist that suggest different meanings. What is the difference between the meaning of *be associated with* and *associate with*? The "statistics" for *be associated with* provides the answer in the form of a collocated word list, which shows that *expression*, *variants*, or *mutations* are associated with *increased risk* or

higher expression (Figure 3). These results suggest that *be associated with* has the same meaning as *be related to*.

2nd left		1st left		1st right		2nd right	
and	25	and	66	a	95	of	44
of	24	that	48	the	85	in	35
in	24	to	43	increased	53	and	29
these	22	have	33	an	36	risk	25
that	21	which	28	decreased	19	increased	23
found	10	has	25	reduced	17	to	11
this	10	this	19	higher	16	expression	10
the	10	may	16	poor	14	development	9
known	9	expression	13	lower	11	reduction	8
1	8	it	10	more	10	levels	7
has	7	cells	10	human	9	cancer	7
have	7	variants	7	enhanced	9	morbidity	7
2	6	mutations	7	disease	8	higher	7

Figure 3. Search results for "*be associated with*."

An input option of "(*associate with*|*associates with*)" enables us to search two query words, *associate with* and *associates with*, at the same time. As shown in Figure 4, *protein* or *proteins* (*physically*) associate(s) with *membranes*, *chromatin*, *proteins*, or *ribosomes*. These results indicate that *associate with* means "bind to." However, sometimes the search is associated with *increased risk*, and in this case, *associate with* means "relate to." In summary, *be associated with* means "be related to." On the other hand, *associate with* usually means "bind to," but sometimes means "relate to."

2nd left		1st left		1st right		2nd right	
that	88	to	95	the	202	and	81
and	46	that	81	a	36	in	23
protein	20	and	54	and	23	of	20
did	14	also	38	increased	13	1	11
of	12	physically	29	multiple	9	at	11
proteins	12	not	22	membranes	9	rna	10
ability	11	can	22	chromatin	8	proteins	9
factor	11	which	22	an	8	to	8
found	10	it	15	one	6	risk	8
which	10	preferentially	14	ribosomes	6	or	7
these	9	protein	13	its	6	a	7
where	9	directly	11	these	6	protein	7
not	9	strongly	9	other	6	membranes	7

Figure 4. Search results for "*associate(s) with*."

4. Searching nouns

For Japanese researchers, discriminating between a countable noun and an uncountable noun is a serious problem, because many nouns used in scientific papers refer to abstract concepts. For example, we need to know whether the term *expression* is used as a countable or uncountable noun. *Expression* in the life sciences field usually means "gene expression," which is different from the general usage of *expression* in other fields. To answer this

question, we can calculate the ratio of the plural *expressions* to the singular *expression* by searching *expression* using the WebLSD Corpus.

The combination of "sort by right" and "include inflection upon rightward sort" provides a sorting of *expression* as the singular, and then *expressions* as the plural. Figure 5 shows part of the search results for *expression*, consisting of 1,000 sentences with the word *expression*. There are only eight usages of *expressions* out of 1,000 sentences (0.8%), indicating that in most cases, the singular form *expression* is used as an uncountable noun in the life sciences. The phrase *facial expressions* is used, but in an overwhelmingly small proportion.

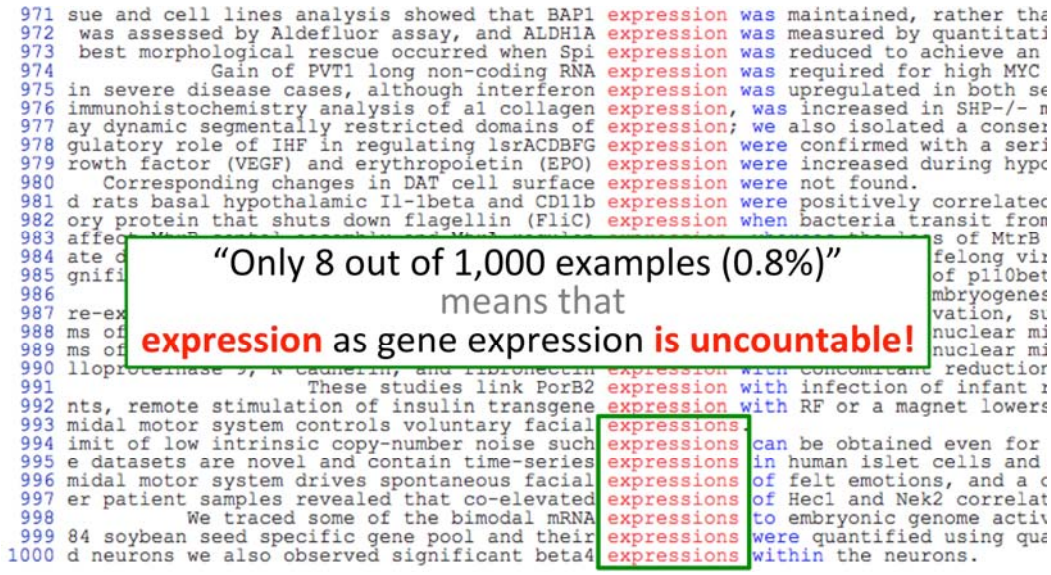

Figure 5. Search results for "*expression*."

To further confirm this conclusion, we examined articles using *expression* in the singular form. For this purpose, it is necessary to search a noun combined with a preposition, because an article corresponds to the last noun in the phrase. The phrase *expression in* is one of the frequently used combinations of *expression* and a preposition. We checked 50 sentences containing *expression in* one by one, by selecting the "disallow inflection of query word" option. As shown in Table 1, there is no indefinite article preceding *expression in*, although one definite article *the*, two pronouns *its*, and 19 gene symbols precede *expression in*. In addition, there are 28 "no article" examples. These results indicate that *expression* is used in the life science field as uncountable in most cases.

a/*an*	0
the	1
pronoun (*its*)	2
gene symbols (without article)	19
no article	28
Total	**50**

Table 1. Articles for "*expression in*." "*Expression*" is an uncountable noun.

5. Irregular usage

Irregular usages of English words are confusing to non-native English speakers. In fact, sometimes we find unexpected usages of words, including *insight*, *movement*, *analysis*, *enhancement*, and *alteration*. Given the particularly strong connection between *insight* and *into*, we examined the usage of *insight into* as a singular form by selecting the "disallow inflection of query word" option. The analysis of 50 sentences for articles containing the phrase *insight into* in its singular form shows that there is no indefinite article, and only one definite article preceding *insight into*. In addition, 49 out of 50 sentences are marked as "no article," suggesting that *insight* is an uncountable noun.

Figure 6. Search results for "insight into."

To confirm this tendency, we calculated the ratio of the plural *insights* to the singular *insight*. The search for *insight into*, by selecting the "allow inflection of query word" option, demonstrates that there are 29 *insights* out of 50 examples, compared with only 21 singular cases *insight* (Figure 6). This result clearly indicates that *insight* is more often a countable noun. Taken together, both the countable form, *provide insights into*, and the uncountable form without the article, *provide insight into*, can be used. However, the phrase *provide* ***an*** *insight into* does not seem to be a standard expression.

6. Discussion

In this article, we showed how to use the WebLSD Corpus to obtain useful information about typical usages and phrases in scientific papers. Ensuring the most appropriate selection of search options is key to finding the necessary information. By looking at 50 sentences at the same time, we can better understand the typical usage of each word.

The LSD corpus, derived from the PubMed abstract database, consists of abstracts from around 150 qualified journals. Abstracts written by researchers in the USA and the UK were collected for the corpus. The size of the LSD corpus is approximately 100,000,000 words.

Information from the WebLSD Corpus is thus highly reliable as information for the life sciences field. We therefore believe that it is useful not only for non-native English speakers, but also for native speakers in terms of providing a suitable guide to usages of words and phrases in specific scientific fields.

Corpora such as the WebLSD Corpus are useful tools for non-native speakers. Researchers who are not native English speakers can find a variety of useful information in such corpora without seeking help from native English speakers. This is a prime example of data-driven learning (DDL), which was introduced by Tim Johns (1990).

The question of how to select certain combinations of verbs and prepositions, or nouns and prepositions is a common problem for Japanese researchers. The WebLSD Corpus provides frequencies of word combinations, which helps researchers to select the most appropriate words, including prepositions.

Although the usage of words in scientific papers should be simple and specific, different usages mark different meanings. By searching for a word as a transitive verb and as an intransitive verb separately, different meanings of the same word can be clarified.

There are many unusual words such as *insight*, *movement*, *analysis*, *enhancement*, and *alteration* found in the LSD corpus. For example, *analysis* can be used as both a countable noun and an uncountable noun.

In conclusion, the WebLSD Corpus is a powerful tool for writing scientific papers in English. To obtain information about specific words, the proper selection of a combination of search options is required.

References

Crystal, David, and Derek Davy. 1973. *Investigating English Style*. London: Routledge.

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. San Diego: Academic Press.

Halliday, M.A.K., and J.R. Martin. 1993. *Writing Science: Literacy And Discursive Power*. London: Routledge.

Johns, Tim. 1990. "From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning." *CALL Austria* 10, 14–34.

Kaneko, Shuji, Nobuyuki Fujita, Yoshihiro Ugawa, Takeshi Kawamoto, Hiroaki Takeuchi, Masataka Takekoshi, and Hiroshi Ohtake. 2003. "*Life Science Dictionary*: A versatile electronic database of medical and biological terms." *Proceedings of Asialex 2003*, 434-439. Tokyo: Asian Association of Lexicography.

Kawamoto, Takeshi, Shuji Kaneko, Brian Morren, and Hiroshi Ohtake. 2004. "Collocational analysis of life science in English (1) – Lists of common collocates of *possibility, probability, implication, involvement, absence, presence, evidence*." *Studia Humana et Naturalia* 38, 19-53.

Kawamoto, Takeshi, Nobuyuki Fujita, Shuji Kaneko, Brian Morren, and Hiroshi Ohtake. 2005. "Collocational analysis of life science in English (2) – Lists of common collocates of *carry, confer, contribute, detect, elucidate, give, know, obtain, raise, understand*." *Studia Humana et Naturalia* 39, 26-69.

Ohtake, Hiroshi, Nobuyuki Fujita, Shuji Kaneko, Brian Morren, and Takeshi Kawamoto. 2006. "Collocational analysis of life science in English (3) – Lists of common collocates of *addition, analysis, hypothesis, identification, level, production, risk*." *Studia Humana et Naturalia* 40, 23-59.

Ohtake, Hiroshi, Nobuyuki Fujita, Shuji Kaneko, Brian Morren, and Takeshi Kawamoto. 2007. "Collocational analysis of life science in English (4) – Lists of common

collocates of *affinity, aim, difference, growth, importance, knowledge, observation, understanding*." *Studia Humana et Naturalia* 41, 67-108.

Ohtake, Hiroshi, Nobuyuki Fujita, Shuji Kaneko, Brian Morren, and Takeshi Kawamoto. 2008. "Collocational analysis of life science in English (5) – Lists of common collocates of *act, action, activate, activation, active, activity*." *Studia Humana et Naturalia* 42, 26-69.