

Introducing Collocations to Japanese Learners: A Task-Based, Corpus-Informed Approach

Lars Martinson

ELS Program at Kindai University

Abstract

This paper presents a proposal for introducing collocational concepts to upper-intermediate Japanese learners. It is structured around the description and discussion of: 1) fixed expressions, 2) binomials, 3) semi-restricted collocations, 4) polysemy, and 5) semantic prosody. A complementary five-stage pedagogic sequence is proposed using multi-word items taken from a New York Times article (Brody, 2017) and supplementary linguistic data gathered from the Corpus of Contemporary American English (Davies, 2008). These sources are used in order to promote reflection about broad lexical tendencies found in authentic contexts. Most of the activities proposed are task-based, and many of them provide students with the opportunity to investigate collocational relationships using tools at their immediate disposal (e.g., commercial search engines and smartphones). This approach is intended to foster learner autonomy by encouraging students to both notice and investigate collocational relationships inside (and hopefully outside) of the classroom.

Keywords: collocations, corpus linguistics, task-based learning

Collocations are combinations of words that tend to appear in close proximity to each other more frequently than can be accounted for by chance (Carter, 1998, p. 51-53). This paper describes collocations identified in an authentic text and offers a proposal for how they could be presented to a class of Japanese university students at the upper-intermediate level. Particular attention will be given to how these examples might be used to help students develop an awareness of the constraints of word combinations and to motivate them to notice and investigate collocations autonomously.

Methodology

Text Selection

The text explored here for pedagogic use is the first eight paragraphs of a New York Times article entitled “Hooked on Our Smartphones” (Brody, 2017). This text was chosen based on its salience to Japanese young adults, the vast majority of whom (96%) own smartphones (Taylor & Silver, 2019, p. 13). The article is critical of frequent smartphone use, arguing in a somewhat dismissive manner that it is symptomatic of misplaced priorities. This perspective may run counter to the students’ beliefs and it is hoped would engage them by providing the opportunity to consolidate and/or reevaluate their attitudes regarding this ubiquitous technology.

Identification of Collocations

Word combinations in the text that seemed to represent collocations were first identified based on intuition. These were then checked against the Corpus of Contemporary American English (Davies, 2008), also known as COCA, to ensure that their collocational

relationships were significant. Those that held up to corpus evidence were then considered for their salience in the service of introducing students to important collocational concepts. Using these procedures, a total of thirteen collocations were chosen for presentation.

Organization

Moon (1997, p. 47-48) describes the difficulty in categorizing multi-word items, arguing that truly discrete lexical categories are a rarity and that some overlap is inevitable. In consideration of this, there has been no attempt to categorize the identified collocations canonically. Instead, they have been grouped opportunistically based on how they might be exploited as a part of a pedagogic sequence.

Five categories have been designated for this purpose and have been organized here in the same order as they would be presented to students. The first three categories (fixed expressions, binomials, and semi-restricted collocations) are collocation types adapted from those described by Carter (1998, p. 70-71) and Alexander (1984). The final two (polysemy and semantic prosody) are characteristics commonly associated with collocations, as described by Carter (1998, p. 12-13) and Partington (2004) respectively.

Each collocation category will be discussed in three parts, starting with 1) a *description* of both the category and the specific instances identified in the text, 2) a proposal for its *presentation* to students, and 3) a *rationale* for both its place in the pedagogic sequence and for the techniques suggested.

Disclaimer

The proposed five-stage pedagogic presentation described below has not been formally tested and as such no claims about its efficacy can be made. It has been provided primarily to serve as a tangible example of how one might introduce collocations to students and as such is intended to be illustrative rather than prescriptive.

Fixed Expressions

Description

Fixed expressions as described by Carter (1998, p. 65-68) are multi-word phrases that possess the following three qualities:

- A) *syntactic irregularity*, or a failure to conform to conventional grammar rules (e.g., the phrase *the more the merrier*),
- B) *semantic opacity*, or having a meaning that cannot be logically construed from the expression's component parts (e.g., *kick the bucket*), and
- C) a *resistance to modification*, be that via inflection, syntactic alternation, or substitution of the expression's component parts.

Carter (1998, p. 68-70) acknowledges that fixed expressions possess these qualities in varying degrees, and that they might best be considered on a cline of fixity. The following two examples taken from the New York Times article are not exemplars for fixed expressions according to the criteria described above, but they would fall squarely on the fixed end of a cline:

- **No doubt** you've seen the following scenarios... (Brody, 2017, para. 2)
This usage of *no doubt* functions as what Swan (2005, p. 20) categorizes as an adverb of certainty. It is somewhat semantically opaque in that it merely means that something is probable or likely but not certain; an expression such as *there is*

no doubt that... can be used instead to express complete confidence (Swan, 2005, p. 353).

- **Don't get me wrong.** (Brody, 2017, para. 5)
This phrase isn't entirely fixed in that it allows for some modification through substitution and inflection (e.g. *don't get us wrong* or *you got me all wrong.*). That said, it is partially restricted in that it doesn't allow substitutions of its final word (e.g., **please get me right*), and when used as in the text, *don't get me wrong* functions as a relatively fixed discourse marker.

Presentation

Introductory tasks are designed to center around processing the gist of the text for meaning. This includes a discourse-level task that addresses *don't get me wrong*. Students should first analyze each paragraph in the text to determine its main topic and whether the author's appraisal of that topic is positive or negative. This could be used as a springboard to discuss how *don't get me wrong* functions to introduce discourse that qualifies or softens statements made immediately prior.

Following these exercises, *no doubt* should be explored. Students should be provided with a randomly ordered list of adverbs of certainty with which they are already familiar (e.g., *probably*, *maybe*, *definitely*, and *probably not*) and arrange them on a cline. They should then be instructed to make predictions about where *no doubt* should be inserted based on its usage in the text. This cline could then be used to contrast *no doubt* with *there is no doubt* to emphasize that the meaning of *no doubt* is not completely logical and that it might best be considered as an unevaluated, atomic vocabulary item.

Rationale

Fixed expressions can serve as an introduction to collocations in accordance with the principle that language should be processed for meaning ahead of form (see Lightbown & Spada, 2013, p. 182-194). Fixed expressions can be evaluated with minimal grammatical analysis (Boers & Lindstromberg, 2012) and since *no doubt* and *don't get me wrong* function to introduce longer stretches of discourse, they might best be exploited in context as a part of introductory, meaning-centric tasks.

Binomials

Description

Binomials are semantically related word pairs that tend to collocate in a specific order (Malkiel, 1959, p. 113). For example, *facts and figures* is far more common in authentic contexts such as the COCA than *figures and facts*. Binomials can be either free or fixed depending on how rigid their syntactic tendency is. Those that are completely fixed are called irreversible binomials (Malkiel, 1959, p. 113) (e.g., *spic and span*, which cannot be rendered as *span and spic*). The following examples of three binomials and one trinomial taken from the New York Times article are freer in that they would still be meaningful in the opposite order, but COCA indicates that they have a strong preference for a particular order:

- **men, women and children** (Brody, 2017, para. 1)
- **time and effort** (Brody, 2017, para. 5)
- **facts, figures** (Brody, 2017, para. 5)
- **friends, acquaintances** (Brody, 2017, para. 6)

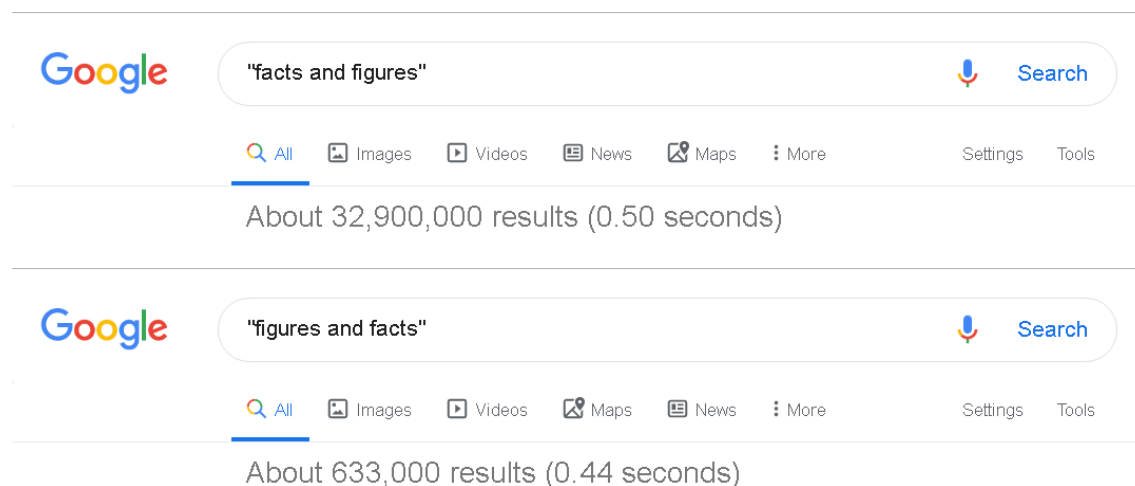
Presentation

The concept of binomials could be introduced by relating them to functionally similar compound words in the students' first language of Japanese (see *Rationale* below). Following this, students should set aside the text and not refer to it for the following tasks. They should be provided with a worksheet containing a jumbled list of lexical words comprised of the constituent parts of the binomials from the text. Students should group related words together and then organize each word pair in the order that seems most appropriate. This determination could be based either on the students' recollection of the text or their intuition.

Students should then test their predictions by using their smartphones to conduct internet searches for both syntactic possibilities of each word pair (e.g., *facts and figures* and then *figures and facts*). Search terms should be put in between quotes so that the results are restricted to only the exact phrase as entered (Figure 1). Students should then record the number of results returned for each combination on their worksheet. In the case of the binomials listed above, the preferred syntactic order offers at least twice as many search results as the non-preferred order, providing a strong indication of which variation is more common. Finally, students should be introduced to the "Me First" principle, which states that binomials tend to be ordered in accordance to their relevance to the "prototypical speaker" (Cooper & Ross, 1975, p. 64-67). Following this principle, *friends* precedes *acquaintances* because the former is more closely related to the prototypical speaker than the latter.

Figure 1

Comparison of quantity of search results for facts and figures (top) and figures and facts (bottom).



Rationale

Binomials were chosen to come second in the pedagogic sequence (following fixed expressions) because they are fairly straightforward and have numerous equivalences to compound words formed by two kanji characters in Japanese. Binomials often have the same syntax as their Japanese counterparts, such as in *husband and wife* (*fū-fu* 夫婦) and *before and after* (*zen-go* 前後). There are also contrasting examples in which the word order is reversed. *Black and white* is rendered as *white [and] black* in Japanese (*shiro-*

kuro 白黒), and the term *supply and demand* occurs as *demand [and] supply* (*ju-kyū* 需給). With an abundance of both complementary and contrastive examples, kanji compound words might help students to appreciate the importance of word order in natural language use.

The search engine task is intended to encourage learner autonomy by providing students with tools to conduct informal research on the words and phrases that they might encounter on their own. Granted, commercial search engines do not offer the same granular search options as fully-featured corpus software suites. But for something as straightforward as binomials, search engine results are consistent with the syntactic tendencies reflected in corpus data. Additionally, search engines are easy to access and use, which may encourage students to experiment with them as a linguistic research tool on their own. This might serve to facilitate the eventual migration to more sophisticated (if less user-friendly) language research tools in the future.

Finally, the “Me First” principle was selected for presentation as this has been found to positively influence Japanese university students’ ability to make accurate predictions about binomial word order (Morita et al., 2014).

Semi-restricted collocations

Description

Semi-restricted collocations offer more possibilities for modification than fixed expressions in that they are often syntactically regular and relatively semantically transparent. But as their name suggests, semi-restricted collocations are constrained in the sorts of words that they tend to “associate” with. The following two examples from the New York Times article illustrate this type of collocational relationship:

- **avid** reader (Brody, 2017, para. 1)
Avid tends to modify nouns related to leisure activities; thus *avid reader* or *avid golfer* are common, whereas *avid worker* is not.
- **transforming** modern society (Brody, 2017, para. 4)
Transform collocates with nouns that represent things that are complex and of great consequence. Among the three most common collocates in COCA are *lives*, *society*, and *economy*.

Presentation

The following procedure illustrates how *avid* might be pedagogically exploited, though *transform* could be similarly explored. Students should work in pairs and, with the aid of a worksheet, make binary judgements about *avid* based on its use in the text (e.g., whether it is positive or negative, whether it describes behavior or appearance, and whether it modifies people or things). Following this exercise, students should be provided with a number of dictionary definitions with the headword removed. These should be comprised of definitions for difficult words from the text with *avid* being among them. Students should be instructed to predict which is the correct definition for *avid*, after which the correct answer should be revealed.

Having established a general sense of the meaning of *avid* both through context and a conventional dictionary definition, students should be provided with a fill-in-the-blank (cloze) exercise comprised of a simple sentence such as “He is an avid _____”. This should be accompanied by a word bank containing randomly ordered words of three types: A) those that would be grammatically incorrect if inserted into the blank, such as

reading or *work*, B) those that are grammatically correct but collocationally atypical, such as *worker* or *dentist*, and C) those that are both grammatically correct and collocationally typical, such as *reader* or *golfer*. Students should work in pairs to determine which words could be appropriately inserted into the blank. They should then come together as a class and form a provisional consensus. The instructor should provide explicit correction only if the students selected any words that were grammatically incorrect.

Students should then be provided with a corpus-based word frequency list for [*avid* + noun]. Based on this data, they should be tasked with generalizing about the types of words that keep company with *avid* and finally be given an opportunity to revise their guesses from the previous cloze activity.

Rationale

Willis (1990) notes that no pedagogic grammar could ever capture how language actually works, and that it is important for students to develop learning strategies independent of grammatical description. In a similar vein, Sinclair (2004, p. 133) argues that decontextualized dictionary definitions are incapable of capturing contextualized word meaning, stating that “many, if not most, meanings require the presence of more than one word for their normal realization”. *Avid* can serve as an illustration for these two observations. *Avid worker* is grammatically acceptable, and if one were to define *avid* as “extremely eager or interested” as the Cambridge Dictionary (Cambridge University Press, n.d.-b) does, a language learner could be forgiven for overextending its usage and assuming that *avid worker* was appropriate.

The pedagogic techniques described above are intended to help students gain an understanding that while dictionary definitions and traditional grammatical descriptions can be a valuable resource, they do not provide a complete picture and should be supplemented with other strategies in order to develop a more nuanced sense of lexical relationships and word meanings.

Polysemy

Description

Polysemy is the phenomenon of a word or phrase having more than one meaning. One of the ways this can occur is through what Carter (1998, p. 12) refers to as “figurative extension”, by which a word that describes something concrete is used metaphorically to conceptualize something more abstract (also see Lakoff & Johnson, 1980). An example is *to grasp*, the original meaning of which (to seize and hold tightly) has been extended to mean “fully understand”, as in *grasp an idea*. The following three polysemous words found in the text have had their meanings extended in a similar fashion:

- **busy** streets (Brody, 2017, para. 2)
Students are almost certainly familiar with *busy* in the sense of having a lot to do. They may be less familiar with it in the sense of full of activity as it is used in the text. This sense of the word frequently collocates in COCA with nouns related to transportation infrastructure (e.g., *road*, *intersection*, and *airport*).
- **avert** countless embarrassing errors (Brody, 2017, para. 5)
Avert was first used to mean to “turn (something) away” (Harper, n.d.-a), and COCA shows that this usage is still common, particularly in fiction (e.g., *he averted his eyes*). In the text, *avert* is used to mean “to prevent something bad from happening” (Cambridge University Press, n.d.-a). This represents a

figurative extension as preventing something might be thought of as metaphorically “turning away” from it. This sense of the word typically collocates in COCA with abstract nouns such as *war*, *disaster*, and *crisis*, and COCA shows that this usage is common in periodicals, newspapers, and academic contexts.

- **bombarded** by bells, buzzes and chimes (Brody, 2017, para. 7)
The verb *bombard* was originally used to mean “to fire heavy guns” (Harper, n.d.-b), and this sense collocates with words such as *missiles*. Its use in the text represents a metaphorical extension of this concept, meaning “to inundate with”. In this capacity it is used almost exclusively with undesired forms of communication, collocating in COCA with words such as *questions*, *images*, and *messages*. This sense of the word is in fact more common in COCA than its original meaning in general contexts.

Presentation

Students should be shown the top Google image search result for *busy*, which is an illustration of a woman with eight arms frantically doing office work. The instructor should elicit descriptions of the image and ask students to guess what the search term was. The correct answer should then be revealed after which students should be asked to locate *busy* in the text. They should then be asked to do an image search for *street*, which mostly returns photos of empty roads. The instructor should then show both the images for *busy* and *street* and ask students to imagine what the results for *busy street* will look like before asking them to verify by searching for the term themselves.

Students should then be provided with a list of word frequencies for nouns that collocate with *busy* taken from COCA and should be asked to organize the words into two categories. Category A would be for nouns related to “having a lot to do”, and Category B would be for everything else. After the class reaches a consensus about which words belong where, the students should be asked to find a pattern for the words in Category B and devise a provisional rule for its use.

The words *avert* and *bombard* could be explored using a similar procedure. First the original literal meaning of the word would be introduced, then it would be related to its metaphorical use in the text, and finally students would formulate hypotheses about how the extended meaning typically functions.

Rationale

Polysemy was chosen to come fourth in the pedagogic sequence because it expands upon concepts introduced earlier, as all of the examples chosen to illustrate polysemy are also semi-restricted collocations. Since the identified collocations have had their literal meanings extended through metaphor, concrete examples are available to present to students to assist them in conceptualizing one of the processes through which polysemy occurs.

Activities include those that require inductive reasoning, as studies have suggested a number of benefits to this approach (Felder & Henriques, 1995, p. 26). The use of activities involving students accessing search engines on their smartphones is again to encourage them to consider different ways of researching language using the tools at their immediate disposal.

Semantic prosody

Description

Semantic prosody is the phenomenon of a seemingly neutral word or phrase having a positive or negative collocational tendency in authentic contexts (Partington, 2004). An example noted by Sinclair (1987, in Partington, 2004) is the word *happen* which despite not being explicitly evaluative tends to be used in reference to negative events. The following two examples of semantic prosody were taken from the text:

- Mr. Miranda's observation **bodes ill** for the future... (Brody, 2017, para. 2)
Conventional definitions do not allude to a prosodic leaning for *bode*, with the Cambridge Dictionary defining it as "to be a sign of something good or bad for the future" (Cambridge University Press, n.d.-c). In authentic contexts it tends to be used in the negative sense. Unlike in the text, *bode well* is the most common collocational relationship in COCA, but it is usually preceded by a negative auxiliary verb (e.g., ...that *doesn't* bode well for...).
- ...she has **encountered** many people who have become "disconnected from what really matters... (Brody, 2017, para. 3)
One sense of *encounter* is defined in explicitly negative terms as "to experience something, especially something unpleasant" (Cambridge University Press, n.d.-d). *Encounter* in the sense that it is used in the text is defined in more neutral terms as "to meet someone unexpectedly" (Cambridge University Press, n.d.-d). An examination of corpus results for [*encounter* + *people*] suggests that this sense of the word also has a significant negative semantic prosodic trend, as the people described tend to be disadvantaged and/or negatively evaluated.

Presentation

The instructor should first divide the whiteboard into three sections labeled "positive", "neutral", and "negative". Students should be asked to predict which category *bode ill* belongs in based on its use in the text. Following this they should then be introduced to the expression *bode well* by examining concordance lines taken from a corpus. These should be used to look for patterns in how *bode well* is used. As it is almost always qualified with some form of [auxiliary verb + *not*], it should be clear that while *bode well* has an explicitly positive meaning, it is still usually used to express negative meanings in authentic contexts.

Encounter should be explored by giving students a decontextualized list of collocating nouns. They should categorize each noun based on whether it has a positive connotation, a negative connotation, or is neutral. Students should then be asked to look for patterns in the words contained in the neutral category, almost all of which describe different kinds of people.

Attention should then return to the New York Times article, with the students being asked if the author is positively or negatively evaluating the people who were encountered. While the word *people* itself is neutral, it is post-modified to reflect a negative evaluation. The instructor should then attach an index card containing this sentence from the text to the "negative" section of the whiteboard.

Each student should then be provided with two to three different concordance lines from COCA for [*encounter* + *people*]. Each of these should be separated on loose index cards that would include expanded context to help students evaluate them. Students

should be tasked with determining whether the people are described in a way that is positive, negative, or neutral. They should then discuss their provisional categorizations in pairs and finally attach their index cards to the board under the appropriate section, providing the class with a visual representation of the prosodic leaning for this usage of *encounter*.

Rationale

Bode well and *bode ill* were chosen to introduce semantic prosody because they are fixed expressions and their usage is fairly restricted, allowing for a straightforward illustration of the concept.

Semantic prosody itself was chosen to be presented last in the pedagogic sequence because an exploration of *encounter* benefits from an understanding of both semi-restricted collocations and polysemy. The proposed activities combine several of the techniques from earlier stages, including those designed to promote inductive reasoning. These are meaningfully expanded by introducing concordance lines and giving students an opportunity to work with authentic contextualized language data.

Summary

The pedagogic activities described above are intended to serve as a broad introduction to the importance of collocation in authentic language use. They were designed to foster learner autonomy through the promotion of inductive reasoning and to demonstrate to students some of the ways that their smartphones could be used as informal research tools. It is hoped that these activities would help students better understand not only the specific words and phrases studied in the lesson, but also serve as a springboard to empower them to notice and reflect on collocational relationships that they might come across on their own.

References

- Alexander, R. J. (1984). Fixed expressions in English: reference books and the teacher. *ELT Journal*, 38(2), 127-134. <https://doi.org/10.1093/elt/38.2.127>
- Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*, 32, 83-110. <https://doi.org/10.1017/S0267190512000050>
- Brody, J. E. (2017, January 9). Hooked on our smartphones. *New York Times*. Retrieved from <https://www.nytimes.com/2017/01/09/well/live/hooked-on-our-smartphones.html>
- Cambridge University Press. (n.d.-a). Avert. In *Cambridge Dictionary*. Cambridge University Press. Retrieved June 21, 2021 from <https://dictionary.cambridge.org/us/dictionary/english/avert>
- Cambridge University Press. (n.d.-b). Avid. In *Cambridge Dictionary*. Cambridge University Press. Retrieved June 21, 2021 from <https://dictionary.cambridge.org/dictionary/english/avid>
- Cambridge University Press. (n.d.-c). Bode. In *Cambridge Dictionary*. Cambridge University Press. Retrieved June 21, 2021 from <https://dictionary.cambridge.org/us/dictionary/english/bode>

- Cambridge University Press. (n.d.-d). Encounter. In *Cambridge Dictionary*. Cambridge University Press. Retrieved from June 21, 2021
<https://dictionary.cambridge.org/dictionary/english/encounter>
- Carter, R. (1998). *Vocabulary: Applied linguistic perspectives* (2nd ed.). Routledge.
- Cooper, W. E., & Ross, J.R. (1975) World order. In Grossman, R., San, L., & Vance, T. (Eds.), *Papers from the Parasession on Functionalism*, 63-111. Chicago Linguistic Society. Retrieved June 21, 2021 from <http://www-personal.umich.edu/~jlawler/haj/worldorder.pdf>
- Davies, M. (2008). The Corpus of Contemporary American English (COCA).
<https://www.english-corpora.org/coca/>.
- Felder, R. M., & Henriques, E. R. (1995). Learning and teaching styles in foreign and second language education. *Foreign Language Annals*, 28(1), 21-31.
<https://doi.org/10.1111/j.1944-9720.1995.tb00767.x>
- Harper, D. (n.d.-a). Avert. In *Online Etymology Dictionary*. Retrieved June 21, 2021 from <https://www.etymonline.com/word/avert>
- Harper, D. (n.d.-b). Bombard. In *Online Etymology Dictionary*. Retrieved June 21, 2021 from <https://www.etymonline.com/word/bombard>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago University Press.
- Lightbown, P., & Spada, N. (2013). *How languages are learned*. Oxford University Press.
- Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 8, 113-160.
[https://doi.org/10.1016/0024-3841\(59\)90018-X](https://doi.org/10.1016/0024-3841(59)90018-X)
- Moon, R. (1997). Vocabulary connections: multi-word items in English. In Schmitt, N., & McCarthy, M. (Eds.), *Vocabulary: Description, acquisition, and pedagogy*, (pp. 40-63). Cambridge University Press.
- Morita, M., Sakaue T., Matsuno K., & Murao R. (2014). Strategies used by Japanese learners of English to determine word order in binomials. *ARELE: Annual Review of English Language Education in Japan*, 25, 65-78.
https://doi.org/10.20581/arele.25.0_65
- Partington, A. (2004). “Utterly content in each other’s company”: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9(1), 131-156. <https://10.1075/ijcl.9.1.07par>
- Sinclair, J. M. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
- Swan, M. (2005). *Practical English usage* (3rd ed). Oxford University Press.
- Taylor, K., & Silver, L. (2019) Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally. Pew Research Center. Retrieved from https://www.pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf
- Widdowson, H. G. (1989). Knowledge of language and ability for use. *Applied Linguistics*, 10(2), 128–137. <https://doi.org/10.1093/applin/10.2.128>
- Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. Collins ELT.