# Corpora: From theoretical linguistics to language teaching

*By* Ikmi Nur Oktavianti

WORD COUNT 5960

TIME SUBMITTED 06-JUL-2021 12:32PM

PAPER ID 74183404

ISBN 978-623-6071-02-1


# Corpora: From theoretical linguistics to language teaching

Ikmi Nur Oktavianti

Universitas Ahmad Dahlan, Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Bantul, Daerah Istimewa Yogyakarta 55191, Indonesia
ikmi.oktavianti@pbi.uad.ac.id

ARTICLE INFO

**Article history**
Received 07 December 2019
Revised 11 March 2020
Accepted 21 August 2020
Available Online 15 January 2021

**Keywords**
corpus
big data
language teaching

ABSTRACT

Corpus has gained its popularity in linguistics over the last five decades, from the computerized storage of English language in Survey of English Usage in 1959 to the ongoing development of Corpus of Contemporary American English. Because of the huge size of actual language data compiled in corpora, many linguists and language teachers working with English language have benefited from them in linguistic research and teaching practice. Up to now, there are innumerable English online corpora recording data from various genres, modes, and regions as well as corpus tools to analyze self-compiled corpus. The massive development of corpora, however, has not been widely discussed among English language researchers and practitioners in Indonesia, let alone in English language teaching. Although linguistics and language teaching are two inseparable and firmly related fields, corpus as a concept and product of linguistics seems ignored or even avoided. This paper then aims to review the nature of corpus and how it is used to assist linguistic analysis. More importantly, this paper discusses another possible application of corpus, e.g., the use of corpus in teaching language. Considering the nature and the benefits of using corpora, it is then important to promote the use of corpus to enhance English language teaching and learning, either directly in the classrooms or indirectly in materials development.



## 1. Introduction

Corpus (pl. *corpora*) derives from Latin word literally means 'body' (*Oxford Dictionary of English*, 2014). According to *Oxford Dictionary of English* (2014), the term *corpus* begins to refer to collection of texts since the early 18th century. In linguistics, corpus is defined as a collection of texts stored digitally to serve as the assistance of language studies (Lüdeling & Kytö, 2009; McEnery & Hardie, 2012; McEnery & Wilson, 2001; O'Keeffe & McCarthy, 2012; O'Keeffe, Mccarthy, & Carter, 2007; Timmis, 2015). A corpus is simply a huge collection of texts we can study further; it is not a theory of language, but it does affect our way of thinking about language and language teaching (McCarthy, 2004).

Corpus is widely applied in linguistic analysis, as it is able to provide empirical evidence to the description of language structure and language use. For decades, its popularity in linguistics is sharply inclining along with the building of new corpora for specific purposes. Whereas language researchers have experienced the advantage of using corpora to improve the description of language, the use of corpora in language teaching, especially in EFL classroom, is still uncommon. In Indonesia, the existence of corpora seems ignored based on the minimum discussion in this topic. One reason is that corpora and learning how to use corpora is seldom part of teacher training courses. Consequently, teachers lack the skills needed to use corpora as native speaker consultants (Granath, 2009).

It seems quite surprising that the use of corpus is not widely discussed in Indonesia. Considering that English status in Indonesia is still a foreign language (EFL), thus using corpus that recording actual use of English is basically crucial in the teaching practice. Moreover, following the program from Ministry of Research, Technology, and Higher Education to embrace *Revolution 4.0* we have to be well-prepared by focusing on *literacy*. Using corpora, language researchers, teachers, and learners, are trained to get used to technology and *big data*. It means that there are two types of literacies covered by using corpus: *technology literacy* and *data literacy*. This paper, hence, attempts to promote the application of corpus, prominently in English language teaching in Indonesia.

## 2. The Nature of Corpus

The term corpus principally refers to at least three things: (1) the large number of data either available online or self-compiled, (2) the tools to analyze big data, and (3) the analysis done in dealing with big data (e.g., frequency, concordance, colocation). One of corpus strengths lies in its ability to store a large amount of data as it is computerized. It enables language researchers to collect both statistically good quantitative data and various qualitative data due to its size as well as enormous data sources (Mair, 2013; O'Keeffe & McCarthy, 2012). This idea is supported by McEnery & Hardie (2012: 27-28) who claim that corpus use in language research might assist to answer various questions because it comprises huge language data. Besides, it is assisted by relevant tools to analyze the data, provided with advanced search enabling accurate search results along with the necessary information (McEnery & Hardie, 2012: 27—28). According to Burkette & Kretzschmar Jr. (2018: 208) language experience of each language user is limited as opposed to the vast number of language users with various background. The concept of big data, thus, is of significant consideration because it gives another insight on providing more empirical language data in bigger size to solve language related problem more comprehensively.

Regarding the nature of corpus there are three important points to note, namely big data, representativeness, and authenticity. Big data refers to a large collection of data purports to provide more accurate basis for many purposes. Hurwitz, et al. (2013: 10) defines big data as any kind of data source that has three shared characteristics, namely extremely large size of data, high velocity of data, and wide variety of data. Big data is one of the results of technology advancement of which computerized storage is possible and commonly used. Using big data, one can inductively interpret the data aiming to prove the existing hypothesis or to generalize the patterns. In other words, in big data culture, one is equipped with the ability to read and analyze data accordingly. A corpus can consist of million words (e.g., COCA, BNC, among others) or even billion words (e.g., i-Web, GloWbE, and others) that meet the criterion of extremely large size of data. Imagine if we as researchers collect the data manually, it is nearly possible to reach that amount. Moreover, a corpus can be accessed quickly (online or offline), and comprises language data from various texts, be they spoken or written. Those two points meet the other criteria of big data; they are high velocity and wide variety. In other words, corpus is indeed big data.

Besides considering the size, a corpus should be representative enough to enable the users generalize the data accurately (Biber, 1993). Representativeness should be taken into account since there are diverse language modes as well as text genres or types. Thus, the design of corpus should concern the target population and sampling method. Biber (1993) states that to fulfill representativeness criterion, a particular corpus should consider range of text types in a language and range of linguistic distribution of a language. One of the well-designed corpora is LOB corpus whose target population are all published texts printed in 1961 in the United States and United Kingdom, with 15 text categories (along with subgenre distinctions). The corpus was also compiled using sampling frames, enabling probabilistic, and random sampling of the population (Johansson, Leech, & Goodluck, 1978). However, the concept of representativeness is debatable. Leech (2007) argues that Biber's concept of representativeness is actually the concept of heterogeneity. It means that language varieties are represented in the corpus in terms of their heterogeneity, instead of focusing to the proportion of the use in the textual universe (Leech, 2007). Apart from the debate, but actually they share the same goal, that is to extrapolate the data outside corpus which later can be used to generalize language use.

## 3. Corpus in Linguistics

Linguistics has benefited from the emergence of corpora several decades ago. In its early emergence, corpus has offered an alternative insight to language analysis that previously relied on native speaker's judgment (McEnery & Wilson, 2001). As McEnery & Hardie (2012) state that today it is almost difficult to find linguistic studies that do not use corpus. This section, therefore, discusses the application of corpora in several aspects of linguistic analyses, covering the structure, meaning, use, and beyond.

Mostly known, corpus is used to assist language researchers in describing the pattern and meaning of a linguistic unit or construction. In doing so, the corpus is advantageous compared to manual methods. Composing a large size of data, the patterns can be more various and thus lead to a more comprehensive result. A study has previously been carried out by Biber et al. (2004: 63) studying the English nominalization suffixes by consulting Lancaster Corpus. It is figured that –tion/-sion is the most frequent one in all registers (academic prose, fiction, and speech) as shown by Table 1.

**Table 1.** Usage percentage of nominalization suffixes (Biber et al., 2004: 63)

| Suffixes | Academic prose | Fiction | Speech |
|---|---|---|---|
| -tion/-sion | 68% | 51% | 56% |
| -ment | 15% | 21% | 24% |
| -ity | 15% | 15% | 15% |
| -ness | 2% | 13% | 5% |

Based on corpus investigation we can then interpret accurately that the most productive nominalization suffix in English is *–tion/-sion.* The result of this interpretation can prove the existing hypothesis about suffix productivity as well as support the qualitative analysis on suffix description.

Corpus can also be used to study language variation as corpus compiles language data from wide variety of texts. As an example, language variation study using corpus in identifying the use of reduced modal verbs (e.g., *gonna, wanna, gotta*) has been conducted by Oktavianti (2018). Oktavianti finds out that the use of reduced modals in fiction differs from academic and news texts. The use of reduced modals is dominantly found in fiction compared to the other ones. The least frequent usage is found in academic text. Another example of using corpus to describe language variation is done by Oktavianti (2019) who studies the use of modal verbs in speech and writing by consulting to spoken and written subcorpus of COCA. Based on the corpus investigation, the distinct characteristics of spoken and written language might result in different choice of linguistic units. As in the study, *be going to* that belongs to quasi-modal is highly frequent in speech rather than in writing, as well as *have to.* In writing, the equivalent units that are more frequent are *will* and *must.*

In relation to language use, it is not impossible to use corpus to depict the dynamic of use. It means that we as researchers compare language use in different periods or a particular time span. By consulting a corpus, this can be accomplished quite easily, let alone there are many types of corpora relevant to this need. As an example, Oktavianti (2019) investigates the use of quasi-modals in Early Modern English using EEBO corpus. In that study, it is evident that quasi-modals are increasingly used in the period. Oktavianti (2019) also studies the changes of modal verbs from Early Modern English to Present-day English by using two corpora, ARCHER to provide data from Early Modern English and COCA to provide Present-day English language data. After investigating those two corpora, it is evident that some modals, especially the reduced ones, are sharply increasing. On the contrary, some modals (including core modal *shall* and obsolete modal *dare*) are dramatically declining.

Not limited to structure and use of language, corpus investigation might also assist the analysis of language to reveal something beyond the language. It means that in this kind of study, language serves as a means of revealing social facts behind the use of the language. This study is known as Critical Discourse Analysis and is compatible with corpus. Baker et al. (2013) analyze the representation of the word 'muslim' in British Press from 1998—2009. This kind of study is possible if we use corpus, in this case is a self-compiled corpus and analyzed using corpus

perspective and corpus to Baker, Gabrielatos, & McEnery (2013) map the categories of the collocates of *muslim* into ethnic/national identity, characterizing/differentiating attributes, culture, conflict, religion group/organization. This kind of classification is not possible without the use of large amount of data. Other studies on Critical Discourse Analysis can be seen in Zahra, Tanvir, & Khoula (2017) and Makamani & Mutasa (2017).

Despite many benefits a corpus offer to linguistic analysis, some researchers prefer to work with small size of data rather than a large corpus. This is related to the depth of analysis the researchers want to focus on. With small amount of data, language researchers can relate the analysis with social context, things that seem to be impossible if they have large amount of data (McEnery & Hardie, 2012). This argument, however, is not disputable since corpus data basically can support the analysis of social context with small subset of data (KhosraviNik, 2009). In other words, corpus study tests the hypothesis by providing empirical evidence with bigger number of data. Another point to take into account is the fact that corpus can provide us with data of many types but it cannot inform us why a certain pattern is used (Sinclair, 1991). It is supported by Cook (1998) who argues that a corpus cannot tell us the process of language in the mind of a speakers or the intention behind an utterance. It means that, as researchers, we have a role to interpret the data a corpus can provide.

## 4. Using Corpus in Language Teaching: Insights For ELT Practitioners

Corpora have influenced English language teaching (henceforth, ELT), particularly in ESL and EFL contexts for years as they can help the design of syllabus, materials, grammars, textbooks, and activities in the classrooms (Conrad, 2000; Jones & Waller, 2015; McEnery & Xiao, 2013; O'Keeffe et al., 2007; Timmis, 2017). Corpus and language teaching get together since the COBUILD project run by John Sinclair to provide English language learners with better dictionaries and teaching materials that present 'real' English and used in actual communicative situations (Römer, 2010: 20). More specifically, corpus was said to revolutionize language teaching. This idea was disputed by Conrad (2000) by stating that corpus is not to revolutionize language teaching. Instead, it might help language teaching and provide another insight. Conrad (2000) claims that it is irrelevant to teach students structures that are never used by native speakers. Based on her corpus investigation, progressive aspect accounts for only small portion and function in conversation, but it is considered important in beginning conversation textbooks.

Several previous studies have shown that there is a gap between real practice of language use and the language used in textbooks. Holmes (1988) figures that epistemic modality in ESL textbooks are not in accordance with epistemic modality in corpus data. Moreover, Carter (1998) compares *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE) with dialogues from textbooks and finds out that the dialogues lack the primary spoken language features like discourse markers, hedges, ellipsis, etc. Furthermore, Gilmore (2004) finds out that the dialogues in seven course books published between 1981—1997 differ from the authentic interaction in a corpus in terms of lexical density, false start, repetitions, hesitation devices, and so on. Another researcher, Römer (2004), compares the use of nine central modal verbs as listed in Quirk, Greenbaum, Leech, & Svartvik (1985) in a corpus and in an EFL textbook in German. She describes that the use of central modal verbs in textbook differs from those in actual use of English. Many studies have proven that the language designed in textbooks is still based on intuitions on how we use language, rather than the actual use of the language. Meanwhile, the idea that scripted dialogues are not effective to improve conversation skills has been revisited in years (Timmis, 2015). It is evident that there is a gap between actual language use and materials designed for learners.

In teaching practice, sometime teachers or ELT practitioners would propose questions like (a) how many words do my students need to learn? (b) does this grammar point is necessary for my students? (c) is this structure in accordance with actual language use?, (d) which words should I use along with this structure?, and so forth. According to Timmis (2015: 2), these questions can be answered by consulting to corpora. Corpora can inform us the most frequent words, the collocation of the words (the profile of the words), the most frequent grammatical structure, and the concordances of a particular linguistic unit. Such information is advantageous to language teaching and learning. As in grammar teaching, it is necessary to prioritize specific grammatical constructions in the teaching materials. As Conrad (2000) claims that valuable and plausible decisions about grammar materials are crucial since teachers cannot teach everything in ESL or EFL grammar class.

By applying corpus, the decisions can be based on large size of empirical evidence on language use to decide what should be taught at each level. It is prominent to highlight that corpus is beneficial in language teaching due to its nature. Cook (1998: 57) proposes insights in relation to the nature of language from corpus investigation. *First,* actual language use is not a matter of using grammatical rules in combination with lexical units, but more a question of collocation. *Second,* some utterance are grammatically possible but do not quite often occur in real practice. *Third,* grammar should be taught altogether with lexical unit concept because they are mutually dependent. These three aspects are possible if teachers consult a corpus. In other words, language teaching should not only consider the methods to teach but also the language input of the learners.

### 4.1. Direct Application of Corpus in ELT

The most widely known direct application of corpora is through the implementation of *Data-Driven Learning* (DDL). Data-driven learning enables the study of large amount of databases of texts (corpora) using software program called concordancer (some onlince corpora are equipped with the concordancing feature) serving as the tool to identify patterns of language data compiled in the corpora (Hadley, 2002; Leńko-Szymańska & Boulton, 2015; Oktavianti, 2015; Timmis, 2015). Using corpus with DDL is useful specifically when working with vocabulary (e.g., synonymous words or expressions) and grammatical patterns as corpus can provide large amount of evidence learners can interpret (Szudarski, 2017; Timmis, 2015).

It will be helpful to look at the example of DDL task to describe the basic principle of DDL. The example below shows the concordances of two synonymous words *persuade* and *convince* in COCA. Learners can be exposed the data and analyze the patterns and, at last, conclude the difference of *persuade* and *convince.*

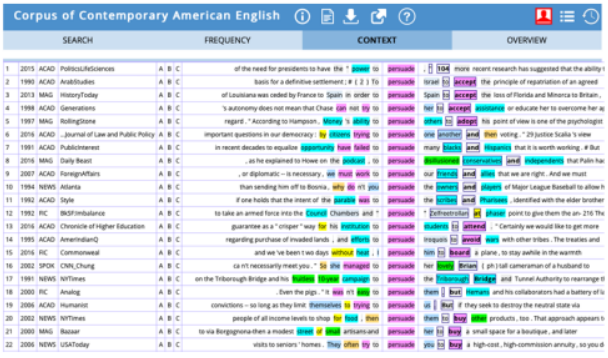


**Fig. 1.** COCA concordance: *persuade*


**Fig. 2.** COCA concordance: *convince*

After observing the concordance lines, teachers then can ask which verb is used more often with the pattern *verb + object + infinitive?*

In teaching grammar, DDL is helpful to recognize the common element precedes or follows a certain unit, as in the example below. Corpus, however, does not prove that a certain pattern is wrong. Instead, it shows which one is more frequently used (thus, normally accepted among native speakers) and which one is infrequently used. As a direct application of corpus—and basically belong to CALL—there are some benefits of DDL. O'Keeffee et al. (2007) claim that by using DLL learners get hands-on experience of using a corpus through guided task or through materials based on corpus evidence, e.g., concordance lines on handouts. Learners can learn how to identify patterns and inductively generalize the patterns. They will learn a lot on how to deal with data (*data literacy*). In other words, learners are trained to implement discovery learning in the classroom by inductively interpreting the data. However, it is a little surprising that DDL does not seem to be used widely and have a wide impact on language teaching. According to Timmis (2015: 142), the word *data* is not an inspiring word for most people and *driven* is probably negatively perceived as no one wants to feel 'driven'. Actually if DDL is used appropriately, it can be of importance for some learners to raise their awareness on language.

### 4.2. Indirect Use of Corpus in ELT

Despite the direct application, corpus can also be used indirectly to enhance language teaching and learning, including in materials design and language testing (McEnery & Xiao, 2013). In designing and developing materials, consulting to a corpus is definitely crucial since corpus can inform the naturally-occurring patterns. In real practice speakers use language as a series of chunks rather than a series of independent words (Hunston, 2009). Statistical information given by a corpus can be used to identify semantic sequences that in turn identify 'what is often said' in communicative situation. In designing and developing materials, corpus use has three orientations: (1) corpus-informed, (2) corpus-based for patterns, and (3) corpus-based for contextualized use (Gablasova, 2018). Each might emphasize different aspect of learning and has its own strength. In a corpus-informed approach, the teacher analyzes the linguistic features in a corpus and modify to the needs of the teaching and the students. The teacher will then consider what aspects to focus on and what to teach first. Meanwhile, in corpus-based for patterns, the teachers will take the examples from the corpus and modify them to demonstrate patterns. In a corpus-based approach for contextualized use, teachers use the whole part of the corpus to expose students to the context of use. In many textbooks (or course books), corpus-informed is more preferable since it can be adjusted to the learning objectives and outcomes.

Corpora have been used to design learner dictionaries like *Cambridge Dictionary of American English* using 100-million-word sample of the *Cambridge International Corpus.* The dictionary consists of more than 40,000 words taken from the corpus so the learner can find commonly used patterns of English. The examples used in the dictionary are authentic taken from the corpus and the definitions of words are based on how they actually used. From dictionary, the corpus is later used to design a course book called *Touchstone.* The authors have spent several years studying the corpus, investigating the most useful grammar and vocabulary for learners from basic to intermediate level, and examining how people communicate in real contexts, especially in conversation (McCarthy, 2004: 3). The authors of the course book claim that using corpus is extremely helpful, e.g., in answering questions that are difficult to rely on intuition. For instance, when do we say *he isn't working* and *he's not working.* The Cambridge International Corpus used by the course book shows that when people use nouns they prefer the former and when they use pronoun, they prefer the latter. This indicates that by consulting to corpus, materials can be designed and developed as close to everyday grammar as possible. M McCarthy, McCarten, & Sandiford (2014) argue that *Touchstone* was designed and developed to meet the criteria of successful course, such as (1) it is interaction-based, (2) it personalizes the learning experience, (3) it promotes active and inductive learning, (4) it encourages students to be independent learners, and it recognize the importance of review and recycling. Point number 1, 3, and 4 are the strengths offered when using a corpus.

In designing grammar materials, in EFL context it is necessary to present grammar items in the best way. An example written by McCarthy (2004) on the use of *must.* Investigating the use of *must* in *The Cambridge International Corpus*, it is found out that on average only 5 percent of all its uses are related to the expression of obligation. The major use of *must* is in predictive statements. This

sort of information helps textbook writers to set priorities and consider which aspect should be taught. In addition, by using a corpus, the teaching of grammar is not in isolation, it will become more integrated with the teaching of vocabulary since grammar and vocabulary are actually connected (Conrad, 2000). For example, in applying grammar rule in conversation, people do not simply say *yes* or *no* to everything. Based on *The Cambridge International Corpus,* there are three most frequent expressions as responses in the spoken corpus, such as *oh that's great, oh that's interesting,* and *that's amazing.* Textbook writers then can use these expressions in their materials.

Not only to teach grammar, might corpus also be considered in the preparation of academic writing materials. Römer (2012) gives an example of corpus-based study in a pedagogical corpus, *Michigan Corpus of Upper-level Student Papers* (MICUSP), to show the use of phraseological items in academic writing, like *in addition to, would be * to.* This study identifies that the use of *it would be * to* is commonly appeared in the sentence-initial and text-final positions but it is not evenly distributed across the paragraph. Meanwhile, the n-gram unit, *in addition,* occurs more frequently at the beginning of sentences and paragraphs but does not show a preference for any particular position in a text. It means that it is safe to use *in addition* in any paragraph of an academic writing, but preferably in paragraph-initial position. This information is important for materials writers, teachers, and learners to achieve successful academic writing course because it is based on the real usage of English in academic context.

### 4.3. Effectiveness of Using Corpus

To support the claim of this paper on the vital role of corpus in language teaching, it is then essential to mention some studies on the effectiveness of corpus-based teaching. One of the studies was done by Lewandowska (2014) that studies the DDL to enhance learner autonomy. The study found that corpus is quite successful in helping students learn autonomously and most of them had positive feelings about the lessons. Another study done by Almutairi (2016) who examining the effectiveness of corpus-based approach in teaching personal statement writing. She has proved that using corpus is very helpful especially for non-native speakers of English. Like Almutairi, Akıncı & Yıldız (2017) find that the use of corpus is effective in teaching English, especially in teaching *verb + noun* collocation to advanced ELT students.

### 4.4. Some Consideration of Using Corpus in Language Teaching

Apart from the benefits, a corpus can offer in language teaching and learning, there are some points to be taken into account. Hunston (2002) describes four main limitations of corpora as the followings:

1. Corpora inform us whether something has occurred and whether it is frequent but it cannot inform us what is possible in a language.
2. Corpora can never be a real representative of a language since language use is extremely complex.
3. Corpora provide us a number of evidence of language use but they do not provide us with interpretations.
4. Corpora cannot capture language use in the whole context (e.g., visual, spatial, or social contexts).

To complete Hunston's statements on limitations of corpora, below is the considerations of using corpora stated by (Flowerdew, 2009). Flowerdew mentions that corpus linguistics techniques encourage an inductive approach of text in which concordance lines are analyzed atomistically or in other words they are separated from the whole context. In addition, corpus data are decontextualized and for this reason may not be directly transferable to students. In terms of student's capacity, some students might not be appropriate to learn using corpus due to their inability to use inductive approach or discovery learning. Flowerdew (2009) also states that it is also quite challenging to determine the corpus to use because there are different types of corpora.

These considerations, however, do not mean to avoid the use of corpora in language teaching. Although it is worth noting that corpus is not everything, but it can be used to assist language teaching and learning in a way that it offers some benefits other media or method cannot; that is related to the content (linguistic aspect) of the teaching and learning practice.

## References

Akıncı, M., & Yıldız, S. (2017). Effectiveness of corpus consultation in teaching verb+noun collocations to advanced ELT students. *Eurasian Journal of Applied Linguistics*, *3*(1), 91–108. https://doi.org/10.32601/ejal.461036

Almutairi, N. D. (2016). The effectiveness of corpus- based approach to language description in creating corpus-based exercises to teach writing personal statements. *English Language Teaching, 9*(7), 103. https://doi.org/10.5539/elt.v9n7p103

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching muslims: A corpus driven analysis of representations around the word "muslim" in the british press 1998-2009. *Applied Linguistics*, *34*(3), 255–278. https://doi.org/10.1093/applin/ams048

Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, *8*(4), 15.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371-405.

Burkette, A., & Kretzschmar Jr., W. A. (2018*). Exploring linguistic science: Language use, complexity, and interaction* (1st ed.). https://doi.org/10.1017/9781108344326

Carter, R. (1998). Orders of reality: CANCODE, communication and culture. ELT Journal, 52, 43–56.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, *34*(3), 548. https://doi.org/10.2307/3587743

Cook, G. (1998). The uses of reality: A reply to Ronald Carter. ELT Journal, 52(1), 57–63. https://doi.org/10.1093/elt/52.1.57

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International journal of corpus linguistics*, *14*(3), 393–417. https://doi.org/10.1075/ijcl.14.3.05flo

Gablasova, D. (2018). *Developing corpus-based materials. Online course presented at the corpus linguistics: method, analysis, interpretation, Lancaster University*. Retrieved from https://www.futurelearn.com/courses/corpus-linguistics

Gilmore, A. (2004). A comparison pf textbook and authentic interactions. *ELT journal*, *58*(4), 363–374.

Granath, S. (2009). Who benefits from learning how to use corpora? In K. Aijmer (Ed.), *Corpora and language teaching*. Amsterdam: John Benjamins Publishing Company.

Hadley, G. (2002). Sensing the winds of change: An introduction to data-driven learning. *RELC Journal*, *33*(2), 99–124.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied linguistics*, *9*(1), 21–44.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, Susan. (2009). The usefulness of corpus-based descriptions of English for learners. in K. Aijmer (Ed.), *Corpora and language teaching*. Amsterdam: John Benjamins publishing company.

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data for Dummies*. Hoboken, New Jersey: John Wiley & Sons Inc.

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo-Bergen corpus of British English for use with digital computers*. Department of English, University of Oslo.

Jones, C., & Waller, D. (2015). *Corpus linguistics for grammar: A guide for research*. London ; New York: Routledge, Taylor & Francis group.

KhosraviNik, M. (2009). The representation of refugess, asylum seekers and immigrants in British newspapers during the Balkan conflict (1999) and the British general election (2005). *Discourse and Society*, *20*(4), 477–498.

Leech, G. (2007). New resources, or just better old ones? The holy grail of representativeness. *In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), Corpus linguistics and the web*. https://doi.org/10.1163/9789401203791_009

Leńko-Szymańska, A., & Boulton, A. (Eds.). (2015). *Multiple affordances of language corpora for data-driven learning*. https://doi.org/10.1075/scl.69

Lewandowska, A. (2014). Using corpus-based classroom activities to enhance learner autonomy. *Konińskie Studia Językowe*, *2*(3), 237–255.

Lüdeling, A., & Kytö, M. (Eds.). (2009). *Corpus linguistics: An international handbook*. Berlin ; New York: Walter de Gruyter.

Mair, C. (2013). Using "small" corpora to document ongoing grammatical change. In M. Krug & J. Schlueter (Eds.), *Research methods in language variation and change.* Cambridge: Cambridge University Press.

Makamani, R., & Mutasa, D. E. (2017). A corpus-based critical discourse analysis (CDA) of the linguistic encoding of HIV and AIDS discourse by the Kwayedza newspaper in Zimbabwe. *South African journal of African languages*, *37*(1), 85–98. https://doi.org/10.1080/02572117.2017.1316933

McCarthy, M, & O'Keeffe, A. (2004). Research in the teaching of speaking. *Annual review of applied linguistics*, *24*, 26–43.

McCarthy, M, McCarten, J., & Sandiford, H. (2014). *Touchstone*. Retrieved from https://www.cambridge.org/gb/cambridgeenglish/catalog/adult-courses/touchstone/methodology-and-research.

McCarthy, Michael. (2004). *Touchstone: From corpus to course book*. New York: Cambridge University press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics*. Cambridge: Cambridge University press.

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

McEnery, T., & Xiao, R. (2013). What corpora can offer in language teaching and learning. *In handbook of research in second language teaching and learning*. https://doi.org/10.4324/9780203836507.ch22

O'Keeffe, A., & McCarthy, M. (Eds.). (2012). *The routledge handbook of corpus linguistics*. Milton Park, Abingdon, Oxon ; New York: Routledge.

O'Keeffe, A., Mccarthy, M., & Carter, R. (2007). *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Oktavianti, I. N. (2015). Data-Driven Learning in the Classroom: The use of British National corpus in teaching vocabulary. The 62nd TEFLIN International conference. Presented at the The 62nd TEFLIN International Conference, Udayana University.

Oktavianti, I. N. (2018). The use of phonetically reduced modals in present-day English: A Corpus-Based Analysis. *English language teaching educational journal*, *1*(3), 134. https://doi.org/10.12928/eltej.v1i3.749

Oktavianti, I. N. (2019). Verba bantu modal bahasa Inggris: Karakteristik, pemakaian dan perubahan. Universitas Gadjah Mada, Yogyakarta.

Oxford dictionary of English [Computer Software]. (Version 2014). Retrieved from: https://support.apple.com/guide/dictionary/welcome/mac Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: longman.

Römer, U. (2004). A corpus-driven approach to modal auxiliaries and their didactics. In J. Sinclair (Ed.), *How to use corpora in language teaching*. Amsterdam: John Benjamins Publishing Company.

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M. C. Campoy, M. L. Gea-valor, & B. Belles-fortuno (Eds.), *Corpus-based approaches to English language teaching*. London: Continuum.

Römer, U. (2012). Corpora and teaching academic writing: Exploring the pedagogical potential of MICUSP. In input, proccess and product. *Developments in teaching and language corpora*, 70–82.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Szudarski, P. (2017). *Corpus linguistics for vocabulary: A guide for research* (1st ed.). https://doi.org/10.4324/9781315107769

Thornbury, S., & Slade, D. (2006). *Conversation: From description to pedagogy*. Cambridge: Cambridge University Press.

Timmis, I. (2015). *Corpus linguistics for ELT*: Research and practice. London; New York: Routledge, Taylor & Francis group.

Zahra, T., Tanvir, O., & Khoula, K. (2017). *A corpus-based critical discourse analysis of racial stereotyping in American newspapers*. Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018). Presented at the Asia Pacific Corpus Linguistics Conference 2018, Takamatsu, Japan.

# Corpora: From theoretical linguistics to language teaching

ORIGINALITY REPORT

14%
SIMILARITY INDEX

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | journal2.uad.ac.id<br>Internet | 69 words — 1% |
| 2 | "Corpora and Language Teaching", John Benjamins Publishing Company, 2009<br>Crossref | 64 words — 1% |
| 3 | npu.edu.ua<br>Internet | 64 words — 1% |
| 4 | uteroemer.de<br>Internet | 51 words — 1% |
| 5 | Rina Febriani Sarie, Bambang Widi Pratolo, Eko Purwanti. "Identity formation: An auto-ethnography of Indonesian student becomes a legitimate speaker and teacher of English", International Journal of Evaluation and Research in Education (IJERE), 2020<br>Crossref | 46 words — 1% |
| 6 | mafiadoc.com<br>Internet | 28 words — 1% |
| 7 | docplayer.net<br>Internet | 24 words — < 1% |
| 8 | D. BIBER. "Representativeness in Corpus Design", Literary and Linguistic Computing, 10/01/1993<br>Crossref | 23 words — < 1% |

| | | |
|---|---|---|
| 9 | studentsrepo.um.edu.my<br>Internet | 21 words — < 1% |
| 10 | epdf.pub<br>Internet | 19 words — < 1% |
| 11 | idoc.pub<br>Internet | 19 words — < 1% |
| 12 | eprints.lancs.ac.uk<br>Internet | 18 words — < 1% |
| 13 | grad.litu.tu.ac.th<br>Internet | 17 words — < 1% |
| 14 | www.modlinguistics.com<br>Internet | 17 words — < 1% |
| 15 | Lynne Flowerdew. "Applying corpus linguistics to pedagogy", International Journal of Corpus Linguistics, 2009<br>Crossref | 15 words — < 1% |
| 16 | docshare.tips<br>Internet | 14 words — < 1% |
| 17 | www.lancs.ac.uk<br>Internet | 11 words — < 1% |
| 18 | Soyeon Moon, Sun-Young Oh. "Unlearning overgenerated be through data-driven learning in the secondary EFL classroom", ReCALL, 2017<br>Crossref | 10 words — < 1% |
| 19 | Alireza Jalilifar, Khodayar Mehrabi, Seyyed Reza Mousavinia. "The Effect of Concordance Enriched | 9 words — < 1% |

Instruction on the Vocabulary Learning and Retention of Iranian EFL Learners", Procedia - Social and Behavioral Sciences, 2014
Crossref

| | | |
|---|---|---|
| 20 | academic.oup.com<br>Internet | 9 words — < 1% |
| 21 | baallkale.files.wordpress.com<br>Internet | 9 words — < 1% |
| 22 | d-nb.info<br>Internet | 9 words — < 1% |
| 23 | etheses.bham.ac.uk<br>Internet | 9 words — < 1% |
| 24 | ojs.pnb.ac.id<br>Internet | 9 words — < 1% |
| 25 | othes.univie.ac.at<br>Internet | 9 words — < 1% |
| 26 | scholarworks.gsu.edu<br>Internet | 9 words — < 1% |
| 27 | Danica Salazar. "Lexical Bundles in Native and Non-native Scientific Writing", John Benjamins Publishing Company, 2014<br>Crossref | 8 words — < 1% |
| 28 | Y B Bhakti, I A D Astuti, R A Sumarni, D Sulisworo, M Toifur. "Implementation of ARCS models to improve teachers' ability in flipped classroom learning", Journal of Physics: Conference Series, 2021<br>Crossref | 8 words — < 1% |
| 29 | clok.uclan.ac.uk<br>Internet | 8 words — < 1% |

| | | |
|---|---|---|
| 30 | commerce3.derby.ac.uk<br>Internet | 8 words — < 1% |
| 31 | ijeltal.org<br>Internet | 8 words — < 1% |
| 32 | pure.qub.ac.uk<br>Internet | 8 words — < 1% |
| 33 | www.jbe-platform.com<br>Internet | 8 words — < 1% |
| 34 | Helmara F. R. de Moraes. "Use of Corpora in Teaching Speaking", Wiley, 2018<br>Crossref | 6 words — < 1% |

| | | | |
|---|---|---|---|
| EXCLUDE QUOTES | OFF | EXCLUDE MATCHES | OFF |
| EXCLUDE BIBLIOGRAPHY | ON | | |