CORPUS STUDIES AND ELT

Semra ONAY TAŞ

Nazlı BAYKAL

1. Introduction

In 21st century, technology has become an indispensable teaching tool utilized for almost every different field of education. Especially for foreign language education, being a resource of motivation has increased the learner's attraction towards the foreign language courses. Besides being a motivation resource, it has been also a beneficial instrument in terms of bringing the learners together with the real life use of target language. In the past, being exposed to real language use was only restricted to going abroad where the target language was spoken or interacting with natives visiting our country. In that case, the duration of interaction was not satisfactory for acquisition. However, today the concept of corpus has the opportunity to bring the language in use samples to teachers and students. Electronic corpora and its attendant software have an increasing influence on language teaching in terms of language description and on the design of teaching materials.

Although corpora first came into view in 1987 following the publication of first corpus-based dictionary for learners - Collins Cobuild English Language Dictionary - (Gabrielatos, 2005), unfortunately, few language teachers have corpus awareness. Most of the teachers are not familiar with the nature of corpora, its relationship with language education and how to use it in their class. In addition, English language teaching programs at many universities have not incorporated corpus-based approach in methodology and material development course syllabi yet.

The frequently asked questions posed by teachers are related to what corpus and corpus linguistics is what it serves in language education and how to use it. Since the efficacy of English learning and teaching can be enhanced with the incorporation of data representing authentic English obtained from corpora (Phoocharoensil, 2012), Timmis (2015) states that language teachers should be familiar with corpus research. Therefore, the aim of this paper is to review corpus (linguistics) and its applications in language education. To do this, primarily, the word 'corpus' meaning body in Latin origin will be defined in the way we believe which is most comprehensive. Secondly, types of corpora (general, specialized, etc.) and their well-known examples will be mentioned. Finally, different types of use of corpora (direct and indirect) in English language education will be presented shortly. In addition to theoretical explanations, the examples of previous studies and future suggestions will be included.

2. Definition and Use of Corpora in General Terms

What is corpus? In a very simple way, Oxford learner's dictionary defines corpus (plural corpora) as a collection of written or spoken texts. In linguistics and in a broad way, Sinclair (2004), precursor of corpus linguistics identified a corpus as "collection of pieces of language text in electronic form, selected according to external criteria (examination of the communicative function of a text) to represent, as far as possible, a language or language variety as a source of data for linguistic research" (para.82). A recent short description was made by Friginal (2018) as systematically collected information of naturally occurring categories of texts. The word 'text' is a common key word in each definition of corpus. Texts refer to both written and spoken language (transcriptions of speech) in corpus linguistics (Timmy, 2015).

A corpus aims to represent a language or some part of a language. It shows us the language use in the real life. On the other hand, it does not answer all the questions about language use, yet the evidences it gives direct us towards a descriptive approach of a language (Jones & Waller, 2015).

A corpus is not only used by researchers but also by language teachers and learners for teaching and learning purposes besides academic researches. For example, using corpus software enables both teachers and students to analyse the most frequent words or language patterns taking place in a certain corpus data. Therefore, teachers and learners can visualise a picture about vocabulary or another language pattern that should be learned primarily (Fauzi, 2020). Corpus, therefore, makes the learners of English as a foreign language or English for a specific purpose get exposed to the language used by native speakers. Learners, teachers or researchers are not dependent on some natives' intuition around them. Thus, corpora are claimed to be more reliable than the native speakers' intuition which is a personal, independent and non-negotiable assessment of language pattern (Sinclair, 2004)

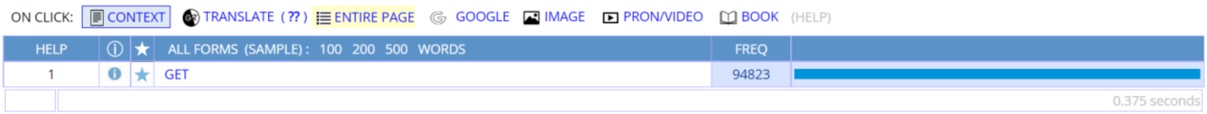
Sinclair (1991, as cited in Fauzi, 2020) stated that the bigger the corpus is, the more reliable it is to generalize the language use better. The British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) are a 'big-name' corpora having both spoken and written components. However, design of the corpus is much more important than its size (McCarty & Carter, 2007, as cited in Timmis, 2015). Smaller corpus is believed to be more useful for classroom purposes. Demographic factors (such as age, gender, social class) and genres and context of the language are related to the design of corpus.

3. Types of Corpus

There are varieties of corpora categorised according to their features and properties. In order to uncover or understand the language patterns, those who are interested in language study can explore different types of corpora with computational tools.

3.1. General / reference corpora

General corpora contain both written and spoken components. Friginal (2018) defines general corpora as "compilation representing language use of very large, diverse groups of people" (p. 16). It includes multiple registers and gives comparative and proportional views on how language is used. That is, it gives understandable information about a language. The word 'reference' reveals that general corpora are used to be base for educational materials such as grammars, dictionaries and other language reference books. BNC formerly known as 'the BYU corpora' is a 'big name' of general corpora created by Oxford University press containing 100 million words of text from variety of different genres (spoken, written, magazines, newspapers, academics etc.). It covers British English of late twentieth century that makes it synchronic corpora. Below you can see a screenshot from BNC website:



HELP	ⓘ	★	ALL FORMS (SAMPLE): 100 200 500 WORDS	FREQ
1	ⓘ	★	GET	94823

0.375 seconds

Figure 1. A screenshot from the BNC website (Source: https://www.english-corpora.org/bnc/)

The research results on the word ‘get’ reveal the frequency of use of the word as ‘94823’. If you click on context, you can see in which context the word ‘get’ is used such as news, technology, engineering, speech etc. ‘Translate’ directs us to google translate web page. By clicking on pron/video, we see the web page of youglish and listen to the pronunciation of the word searched in related context. Book part offers us the books containing that word in its name. In short, it gives us almost everything related to the word ‘get’ from a single web site.

COCA is another ‘big name’ in general corpora. It contains more than one billion words of data and the newest version was released in March 2020. It has also wide ranges of genres as for: spoken, fiction, magazine, newspapers, academics, blogs, TV programmes and movies. (Davies, 2008-)

3.2. Specialized corpora

This kind of corpora is called as pedagogic or teaching-oriented corpora (Friginal, 2018) and it is compiled for mostly teaching purposes or particular research aim. That kind of corpora contains specific texts/discourse about particular subjects (arts, politics, aviation etc.). We can say that specialised corpora is superior to general corpora when teaching objectives are in question because teachers can easily focus on their intended specialized area such as academic written English or professional English.

Michigan Corpus of Academic Spoken English (MICASE) and its British analogue British Academic Spoken English (BASE) are well known specialised corpora. The former contains transcripts of different kinds of lectures, discussion sessions, dissertation defences, campus/museum tours and interactive exchanges in academic environment at the University of Michigan while the latter contains recordings and transcriptions of lectures and seminars at the university of Warwick and Reading (Weisser, 2016). British Academic Written English (BAWE) is another academic writing specialised corpus developed with good quality students’ assignments from three UK universities (Çakmak, 2021).

3.3. Written and spoken corpora

Written corpora are only composed of texts produced in written form such as newspaper articles, academic reports, facebook or other social media posts, e-mails, text messages, tweets, amazon customer product reviews, etc. That kind of corpora gives the opportunity to the researcher to analyse long and complex sentences of language users. On the other hand, spoken corpora having transcripts of spoken language is less common than the former. The Santa Barbara corpus of Spoken English (SBCSAE) including face to face conversations, different kinds of speech event, telephone conversations, conversations during a lecture, etc. are available for free access on the website: https://www.linguistics.ucsb.edu/research/santa-barbara-corpus#Intro.

3.4. Annotated corpora

Providing corpora with interpretive additional linguistic information is the practice of annotation. It makes the corpora more useful and more comprehensive. Annotation can be practiced in many ways. The most common type is grammatically tagging or labelling. That is, labelling a word according to class which it belongs to. In other words, adding part-of-speech information that is called as POS tagging is a sort of annotation. Depending on different levels of linguistic analysis of a corpus, other methods of annotation are phonetic, semantic, pragmatic, discourse, stylistic and lexical annotations (Leech, 2004). The Brown Corpus, the Lancaster-Oslo-Bergen Corpus (LOB) and BNC are examples of annotated corpora.

3.5. Comparable and parallel corpora

Comparable corpora can be defined as multilingual corpora. The issue is that the components of two or more languages are compiled in the same way. LOB and Lanchester Corpus of Mandarin Chinese are examples of comparable corpora (Friginal, 2018). Mcenery and Xiao (2007) define parallel corpora as corpora containing texts in L1 and their translations in L2. Texts from similar domains and context in the L1 and L2 registers are compiled. The Canadian Hansard Corpus is a well-known parallel corpus consisting of English and Canadian French parallel texts collected from official records of Canadian Parliament (Linguistic data consortium, n.d.). In short, both parallel and comparable corpora are used primarily for translation purposes.

3.6. Learner corpora

Learner corpora are specific type of non-native corpus. The data for learner corpora are provided by English language learners. International Corpus of Learner English (ICLE), coordinated by Sylviane Granger of the University of Louvain, is a famous corpus based on essays written by upper intermediate and advanced learners around the world (UCLouvain, n.d.).

Granger (2008) and Meunier (2021) claim that informing second language acquisition (SLA) research and providing useful inputs for applied projects for the creation/improvement of teaching materials and approaches are two main purposes of learner corpora. It is also a useful resource for developing assessment techniques for computer-assisted language learning (CALL) system (Kotani & Yoshimi, 2015). Among the four fields of pedagogical perspectives (material design, syllabus design, language testing and classroom methodology), material design had concrete achievement. For instance, the Longman Dictionary of Contemporary English (2003) and Cambridge Advanced Learner's Dictionary (2003) added common error parts by analysing learners' errors in learner corpora (Granger, 2008). Thus, we can assume that learner corpora are useful for error analysis so that we can see whether our learners use language accurately or not. The error analysis enables teachers or researchers to explore what the learners have learnt in grammar or in vocabulary and challenges they have while using the language (Kotani et al., 2016).

4. Corpus Linguistics and English Language Teaching

Bennet (2010) explains the idea behind the corpus linguistics with Sinclair's detection of the meaninglessness of the word alone. To Sinclair (2004), a word in itself does not carry meaning, but meaning is gained through several words in a sentence.

Corpus linguistics is a kind of methodology to study language expressed in its real life context. That is, it is the study of a language through corpus. It focuses on analysis of real samples of language use (Cheng, 2012, as cited in Szudarski, 2018; Roca Varela, 2012).

There are two different approaches in corpus linguistics: A corpus-based approach is seen as a research method in which corpus evidence is used to test or to exemplify the existing theories of languages (top-down analysis). Corpus-driven approach is seen as theory-generating in the field of linguistics. The investigated linguistic features come directly from analysis of the corpus, not from the categories pre-established by the researchers (bottom-up) (Friginal, 2018).

Corpus linguistics analysis differs from others by following a critical inductive approach to language. It reveals integrated relation between form and meaning. Therefore, a new pedagogical view called lexicogrammar is provided (Samburskiy, 2014). Since the representativeness of linguistic properties in real life is essential for the authenticity and descriptive nature of language, corpus linguistics tries to answer two main research questions mentioned in Bennet (2010, p. 2);

1. What particular patterns are associated with lexical or grammatical features?

2. How do these patterns differ within varieties and registers?

As we have already mentioned, corpora do not always give responses to all the questions related to language use and corpus linguistics does not provide all possible language use at one time. That is, no matter how big a corpus size is, all samples of a language use may not be present. Corpus linguistics does not provide us negative evidence. That is, non-existence of a particular language pattern in a corpus does not mean that it is not possible or incorrect. We should remember that it may not be very common in the register represented by the corpus.

Finally, corpus linguistics does not give the answer of the question ‘why’. It only tells us what it is. To answer the question, language users use their intuition. To Timmis (2015), English language teachers should be familiar with corpus research and practice because most of the time, they ask themselves questions about the number of words the students should learn, the necessity of grammar at a specific point, the usability of language in real life by the students, etc. Through corpora, they can observe and study the real-world language use, with relevant distributions of frequency and respond to their own questions by accessing actual occurrences of features, rather than relying on limited intuition (Friginal et al., 2020). Through several domains of applied linguistics such as grammar, lexis, pragmatics; SLA can be investigated. Frequency issues are in question. As for the examples of frequency, the following can be given as examples of some studies: the most frequent words and phrases in English, the difference(s) between written and spoken English, most frequently used tenses, use of modal verbs, different uses of words in formal and in informal situations, the number of words a learner must know to participate in everyday conversation, etc.

There are two ways of using corpora in language teaching and learning: *indirect* and *direct* use. Indirectly, many of the teachers and learners are benefiting from corpora without realizing it or having no idea what a corpus is. Teaching materials, reference books, dictionaries and course books are developed by experts with corpus-based approach. The list mentioned in Garcia (2014, p. 2) shows some examples of corpus-based ELT publications including dictionaries, grammars and textbooks:

Cambridge Dictionary of American English

Cambridge International Dictionary of English

Cambridge Grammar of English

Collins COBUILD English Dictionary for Advanced Learners

Collins COBUILD English Usage

Collins COBUILD Intermediate English Grammar

Longman Dictionary of Common Errors

Longman Dictionary of Contemporary English

Longman Grammar of Spoken and Written English

Macmillan English Dictionary

Macmillan Collocations Dictionary

Natural Grammar (Oxford)

Oxford Advanced Learner's Dictionary

Oxford Collocations Dictionary for Students of English

Practical English Usage (Oxford)

Touchstone series (Cambridge)

Vocabulary in Use series (Cambridge)

Through such kind of materials, learners have opportunity to contact indirectly with corpora. In other words, learners or teachers can access indirectly to the samples of real life language use without accessing directly to the corpora systems (e.g. BNC, COCA, etc).

Since it may be sometimes difficult and time consuming, teachers do not need to select or edit the data according to their target learners. Material design and development units at schools or universities take the responsibility of selecting suitable course materials (e.g. textbooks, worksheets, other visuals, etc.) and deciding on how to teach the related syllabus to the learners. In other words, they determine the methodology, syllabus and textbooks so that teachers use them in their classroom. At that point Jones and Waller (2015) suggest corpus-informed syllabi, textbooks and methodologies in language teaching. Since it is a stubborn fact that corpora cannot always provide all answers in ELT textbooks, they do not argue in favour of textbooks purely led by corpus. Learning materials containing the information a corpus can provide us is rather desirable than not having it at all (Jones & Waller, 2015).

Although there are many corpus-based ELT materials published by the native users of language as shown on the list above, non-native published textbooks underrepresenting authentic language still needs to be revised or changed. For example, in Turkey, Peksoy (2013) in his MA study, analysed ELT course books used in high schools in Turkey in order to find out appropriateness of the language of those books to the language used in countries where English is the native language. For the purpose of corpus-based analysis, the researcher scanned the content of all high school English students' books and workbooks on computer and uploaded them to Sketch Engine, which is an online corpus-analytic programme. Specific grammar topics were searched; similarities and differences were compared. It was finally found out that the English language course books at the level of high school underrepresent authentic language in terms of certain grammatical items and frequency of their collocation. According to the results of this particular study, especially the course books written by Ministry of National Education (MoNE) book committee should be revised or changed.

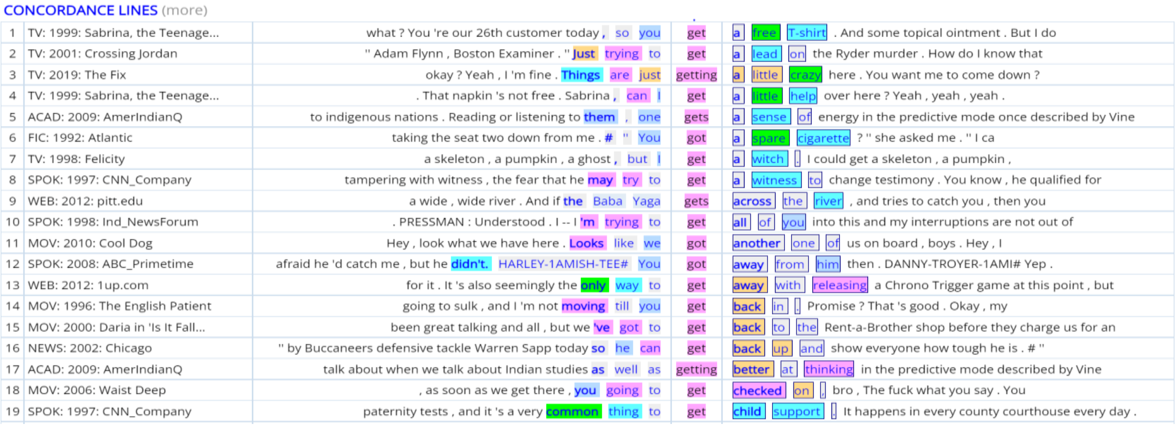
In another study, Çakmak (2021) similarly analysed the authenticity of dialogues in B1-B2 levels ELT course books of the MoNE by comparing their content with the Spoken British National Corpus of 2014, in terms of functional categories with pragmatic functions, discourse markers, face and politeness, conversational routines, etc. She also investigated the opinion of native speakers about the authenticity of course books. The dialogues were analysed in detail with a native speaker in collaboration. To make the comparison, the researcher created the course book corpus and uploaded it to the Sketch Engine. Frequency count of each item and concordance of each word were analysed. The result of the study revealed that the dialogues in the course books do not represent the functional categories with pragmatic functions at a desired level. The interview with the native speaker showed that there were a lot of inauthentic and incorrect language uses in the course books. Therefore, there are inconsistencies between real life language use and EFL course books. Those two studies having almost the same aim conducted eight years apart, show that textbook (course book) design is still a significant problem in our country today.

In addition to the discussion about the representativeness of authenticity in teaching materials above, Garcia (2014) stated that language is not such a limited and easy entity that can be simply put in a book or reflected on a teaching material. Those published formats of corpora are very restricted. Nevertheless, corpora support learners with more language uses and show distinct variations of language. Therefore, they provide scaffolding in education. That way, there is another option for language education which is *direct use of corpora*. When we take the teaching syllabus, published books or other corpus-based teaching materials into consideration, we notice that indirect use of corpora are mostly related to what to teach, but corpora provide more than what to teach. Corpora enable teachers how to teach.

Fligelstone (1993, as cited in Şimsek, 2020) and Leech (1997, as cited in McEnery &Xiou, 2010) put forward three different direct uses of corpora in teaching language: teaching about, teaching to exploit, and exploiting to teach. The first one is related to teaching corpora/corpus linguistic as an academic subject at universities like the other sub-branches of linguistics and training other researchers/teachers with the aim of integrating corpus linguistics in language teaching. ‘Exploiting to teach’ is about exploiting corpus resources to teach in order that the teachers can decide on what to teach and select the tools they are going to use in their classroom. Regarding the last type of direct use of corpora, McEnery et al. (2006) explained ‘teaching to exploit’ as teaching students how to benefit from corpora for their own purposes. That is called data driven learning (DDL) or ‘discovery learning’ (McEnery & Xiao, 2010).

Discovery learning approach increases the autonomy of the learners. With the corpus based approach to language pedagogy, *three Is* of the exploratory teaching approach prevail the *three Ps* of traditional teaching approach. To remember, three Ps are: presentation, practice and production while three Is are: illustration, interaction and induction. Looking at real data signifies the illustration phase which is the first step of discovery learning. During the second phase -interaction-, the learner observes, discusses and shares opinion about the data. Finally, the learners generate rules with inductive reasoning. Learners discover those patterns and rules through analysing concordances’ lines.

A concordance is a listing of each occurrence of a word or pattern in corpus presented with the other words surrounding it. Concordances are generated by computer and learners can access the multiple examples of authentic data in the forms of concordance lines. They enable learners to analyse the different language forms in context of use (Rashikawati, 2019). Key Word in Context (KWIC) is the most common concordance format in corpus linguistic. The meaning of searched item is not given explicitly on the screen. It is usually deduced through the examples provided with concordance lines. The figure 2 below shows the DDL principle of presenting the data and asking learners to notice linguistic pattern (Timmis, 2015):



CONCORDANCE LINES (more)

1	TV: 1999: Sabrina, the Teenage...	what ? You 're our 26th customer today , so you	get	a free T-shirt . And some topical ointment . But I do
2	TV: 2001: Crossing Jordan	" Adam Flynn , Boston Examiner . " Just trying to	get	a lead on the Ryder murder . How do I know that
3	TV: 2019: The Fix	okay ? Yeah , I 'm fine . Things are just	getting	a little crazy here . You want me to come down ?
4	TV: 1999: Sabrina, the Teenage...	. That napkin 's not free . Sabrina , can I	get	a little help over here ? Yeah , yeah , yeah .
5	ACAD: 2009: AmerIndianQ	to indigenous nations . Reading or listening to them , one	gets	a sense of energy in the predictive mode once described by Vine
6	FIC: 1992: Atlantic	taking the seat two down from me . # " You	got	a spare cigarette ? " she asked me . " I ca
7	TV: 1998: Felicity	a skeleton , a pumpkin , a ghost , but I	get	a witch . I could get a skeleton , a pumpkin ,
8	SPOK: 1997: CNN_Company	tampering with witness , the fear that he may try to	get	a witness to change testimony . You know , he qualified for
9	WEB: 2012: pitt.edu	a wide , wide river . And if the Baba Yaga	gets	across the river , and tries to catch you , then you
10	SPOK: 1998: Ind_NewsForum	. PRESSMAN : Understood . I -- I 'm trying to	get	all of you into this and my interruptions are not out of
11	MOV: 2010: Cool Dog	Hey , look what we have here . Looks like we	got	another one of us on board , boys . Hey , I
12	SPOK: 2008: ABC_Primetime	afraid he 'd catch me , but he didn't. HARLEY-1AMISH-TEE# You	got	away from him then . DANNY-TROYER-1AMI# Yep .
13	WEB: 2012: 1up.com	for it . It 's also seemingly the only way to	get	away with releasing a Chrono Trigger game at this point , but
14	MOV: 1996: The English Patient	going to sulk , and I 'm not moving till you	get	back in . Promise ? That 's good . Okay , my
15	MOV: 2000: Daria in 'Is It Fall...	been great talking and all , but we 've got to	get	back to the Rent-a-Brother shop before they charge us for an
16	NEWS: 2002: Chicago	" by Buccaneers defensive tackle Warren Sapp today so he can	get	back up and show everyone how tough he is . # "
17	ACAD: 2009: AmerIndianQ	talk about when we talk about Indian studies as well as	getting	better at thinking in the predictive mode described by Vine
18	MOV: 2006: Waist Deep	, as soon as we get there , you going to	get	checked on , bro , The fuck what you say . You
19	SPOK: 1997: CNN_Company	paternity tests , and it 's a very common thing to	get	child support . It happens in every county courthouse every day .

Figure 2. A screenshot of concordance output of ‘get’ from the BNC website (Source: https://www.english-corpora.org/bnc/)

Concordances provide rich information such as meanings, use, collocation, word families, etc. Researchers, teachers, and even the learners can filter, sort out and count in order to get their desired result (Sketch Engine, n.d.). For instance, Quilichevna (2020) in his/her study on corpus based approached in vocabulary teaching underlines the role of teacher in using corpora to teach languages and give some sample corpus based activities applied to concordance lines from any text. The researcher stated that with the concordance activity, students cannot only notice negative meaning but also positive meaning and word combinations of focused word. They can also try to find meaning of a given word and its different usages. Corpus based tasks are claimed to help students to be exposed to how English is used for communication by native speakers.

In short, students are able to discover the languages rules and patterns through authentic materials. In order to push the learners to discover and to make generalization of rules, those concordance lines are to be supported with questions and prompted by teachers or material makers. Those questions enable learners to think critically and to participate actively in the learning process. For instance, instead of looking an unfamiliar word up in a dictionary, they will try to find out the meaning from the context.

5. Applications of Corpora in Language Teaching

Friginal (2018) in his book entitled “*Corpus linguistic for English teachers: New tools, online resources, and classroom activities*” explains the application of corpus linguistic in language teaching as integrating corpus tools and the corpus analysis as part of classroom activities or homework assignment. Students benefit from corpora online databases when they deal with research projects or find answers to questions related to language patterns. Both teachers and students explore language patterns of use from concordances and corpus-based materials. Since language is a systematic and identifiable phenomenon through empirical, frequency and pattern-based approaches, corpus linguistic methodology provides those relevant data about English vocabulary and grammar as they are used in their natural and authentic context.

5.1. Corpora and vocabulary teaching

Vocabulary knowledge is important for foreign language learning. However, memorising vocabulary is not enough for acquisition. Vocabulary knowledge in terms of semantic accuracy, which we can call meaning in context and its accurate use, is significant for achieving the goal of learning a foreign language, which is successful communication in writing and speaking (Mukhamadiarova et al., 2020). Being exposed to a word in its real context facilitates its memorization and internalization. Corpora provide that meaning in context.

There are two main effective approaches to teaching vocabulary: explicit vocabulary teaching and incidental learning. The former occurs when the teacher takes the learners' attention in a definite way to learn words. On the other hand, incidental learning is the result of learner's exposures to written or spoken text without explicitly directing attention (Çalışkan & Kuru Gönen, 2018). While dealing with a large size of English vocabulary, it is important to combine those two approaches. In order to integrate those approaches for an efficient instruction of vocabulary, it is essential to use variety of computer assisted techniques including corpora (Çalışkan & Kuru Gönen, 2018).

Regarding some previous studies conducted using corpus based activities, DDL and concordances, corpora were found useful and effective in vocabulary instruction. Roca Varela (2012), for example, aimed to show pedagogical usefulness of corpora for vocabulary teaching by presenting four different corpus-based activities and also different ways of working with corpora resources in the classroom. With those activities, researchers had the objective of making Spanish students encounter the words in its real life context. First, the students studied the meaning and use of the word 'carpet' which is often confused by Spanish students with the word 'carpeta' in their native language that lead them to use it for the same meaning and in the same way. After having highlighted examples illustrating most clearly the meaning and the use of 'carpet' on the random concordance lines of BNC, the sample sentences made the word clear for the students. As a second activity, the researcher used Collins Wordbanks Online to make the students aware of the syntactic, semantic and frequency differences of confusing words (signify/mean). In their third activity, corpora were used to look into polysemic nature of some words (*suburb* in the study). The students realized that 'suburb' does not denote an unpleasant or dangerous place. For the final activity, students were asked to compare learner's use of some words with native users by consulting Santiago University Learner of English Corpora and the British National Corpus. At the end of the comparison, students realised that they misused the words. Judging by the activities, the researcher emphasized that the advantages of corpora use is enormous in classroom and that those databases show the learners how the mechanics of language use works and help them to solve lexical problems and to be more accurate in English.

However, they claim that vocabulary teaching should not be based on only corpora; otherwise, it might be boring. Good combinations of different techniques are advisable.

Similarly, Ergül (2014), in her master’s thesis, investigated the effectiveness of corpus based activities instead of using textbook activities and using dictionaries in vocabulary teaching and compared students’ attitudes towards corpus-based and textbook activities. 34 students of intermediate level of English were divided into two groups (control and experimental groups). For the control group, ‘North Start 3 Reading and Writing’ text books and vocabulary activities were covered while with the experimental group those exercises were replaced with corpus-based vocabulary activities by choosing appropriate concordance lines from COCA. A pre-test was applied in order to see the effectiveness. After a six-week experiment, a post test, an attitude questionnaire and interviews were administered. The post-test results revealed that experimental group had higher scores than the control group. That is, corpus-based material was found more successful in teaching than the text book and dictionary activities. Attitudes questionnaire and interviews showed that students had positive attitude towards corpus-based activities in vocabulary teaching.

In line with Ergül (2014), two other recent researches on learner’s and teacher’s perceptions of corpus application in classroom revealed that the attitudes toward using corpora in classroom while teaching vocabulary are positive. To illustrate, Sinha (2021) examined 32 EFL learners’ perception of corpus application as a vocabulary learning tool, challenges they encounter while using corpus data and their thought about teacher support while using a corpus for learning new words. Data were collected through a perception questionnaire after teaching certain academic words using a corpus. She found out that most of the students have both positive and negative attitudes towards classroom application of corpus. Generally they think that corpus is an effective tool for learning a new word; however, the absence of teacher guidance impact the success of corpus based vocabulary learning because of the nature of corpus data which often renders learning difficult for learners inexperienced in using corpora.

Çalışkan and Kuru Gönen (2018) aimed to explore teachers’ opinion on the use of concordance lines for vocabulary teaching and to investigate the perception of language teachers on vocabulary teaching based on corpus-based pedagogy after having received training. Participants were three EFL teachers at a Turkish state university and the data were collected through semi-structured interviews, an open-ended questionnaire and reflective logs. The teachers had four week training on corpus pedagogy and on how to design and adopt concordance lines for their vocabulary instruction. The results revealed that EFL teachers had positive attitudes towards corpus application in classroom and found it efficient for language pedagogy. Unfortunately, the study demonstrated that teachers had no idea about using corpus materials before the training. Therefore, training created an awareness of incorporating corpus-pedagogy into classroom for vocabulary teaching.

A different comparative study about vocabulary teaching was conducted by Sezgin and Öztürk (2020). They compared the language used in TV series (Sherlock and Doctor who) to the spoken part of BNC in order to find out whether those TV series reflect the real life spoken language in terms of vocabulary. British TV Series Corpus (BTSC) was collected for that comparison. The study revealed that the TV series corpus covered 98.54% of the most frequent lemmas in the spoken part of BNC. Hence, the researchers claimed that TV series are effective materials used as both in class and extra-curricular activities for teaching vocabulary and listening/speaking skills.

5.2. Corpora and grammar teaching

Contrary to grammar books written using traditional approaches, corpora provide descriptive statements rather than prescriptive ones. That is, corpora do not say that something is certainly wrong or right but it describes how the language is used. Like in vocabulary, context is an essential feature in grammatical choice. Jones and Walter (2015), in their book '*Corpus linguistic for grammar: a guide for research*', explained what a corpus can tell about the aspect of grammar as follows:

- Which area, in particular contexts, are more frequent than other forms? (e.g. comparison of the frequency of past simple and past perfect production in spoken narratives)
- The difference between written and spoken forms in particular context.
- Specialised uses for English for Specific Purposes. (e.g. Business English or English for tourism)
- How language patterns colligate and collocate. Discovery of likely combinations of words.
- How a particular form can be used to have negative, positive or neutral connotation in a semantic prosody.

As understood from the statements above, many areas of grammar are still under-described in pedagogically simplified grammar books. In a similar methodology with vocabulary instruction, DDL is very useful and practical in grammar instruction. Nevertheless, it is not so common in most of the ELT classrooms. Unfortunately, traditional methods keep their place. DDL is a bridge between corpora and the classroom. Just like in vocabulary learning, students start to investigate grammatical and lexico-grammatical patterns in language. DDL in grammar teaching necessitates both product and process approaches. Instead of activities focusing on the teaching of prescriptive rules, activities raising learner's consciousness should be practiced in the classroom.

To Wang (2018), there are three main characteristics of corpus-based grammar teaching method. First, easily accessible and large amount of natural occurring texts from different fields increase students' context awareness. Then, students notice the specific use of grammatical expressions themselves with the help of the teacher. Finally, the combination of top-down and bottom-up approaches promote discovery learning of students. As for the example of an academic study, Oghigian and Chujo (2010) developed a series of corpus-based activities for beginner level of English in teaching vocabulary and grammar. They preferred not to ask the students to read and understand concordance lines because it might be difficult at beginner level. Instead, they asked the students to focus on KWIC which means the key word in the centre of the data. In the first task, they asked the students to observe the various forms of the word ‘develop’, which was thought to be useful for understanding derivations of the assigned word, identifying part of speech or verifying the correct form. Another activity was given with a list of words ‘information, data, homework, passenger’ to ask them search to find out which words are countable and which words are uncountable. The third activity made the students notice the existence of an adjective between an article and the word ‘organization’. In short, all the proposed activities were good examples of discovery learning. The researchers stated that there were significant gains in related areas and students showed positive attitudes towards corpus-based activities in questionnaires.

Another study conducted in China by Wang (2018) aimed to explore the effectiveness of corpus-based grammar teaching method compared to traditional grammar teaching method. For the study, 40 undergraduate Chinese students were divided randomly into two groups -control and experimental - following a pre-test that showed no significant difference between participant groups. As for the contrastive experiment, three conjunctions expressing the subordinate clause for causality - because, since, for - were searched by the teacher for concordance lines from COCA and BNC. Then, a mini-text made up of 40 concordance lines was also searched. By showing the mini-text to the experimental group, the teacher made the students observe the lines to find out collocates of the concerned three conjunctions. Then, the students were asked to generalize their usage pattern and to summarize the similarities and differences. In contrast, the control group was taught the words traditionally by using the course book and some mechanical exercises. The results revealed that the experimental group has higher grammar proficiency. In addition, interviews and in-class observations showed that students developed positive attitudes towards corpus-based grammar teaching method. Students’ sensitivity of language use, focusing on semantic meaning, pragmatic sensitivity and context awareness increased.

In line with the previous researches summarized above and the studies mentioned for vocabulary teaching, Pookharoensil (2012) also revealed similar results on the effectiveness of corpus-based method and the attitudes of learners toward corpus-based activities. In the study, the researcher presented grammar topics such as conditionals and who vs. whom to 17 Thai graduate students in corpus-based method. After the instruction, the students completed an opinion and attitude questionnaire and they were also interviewed individually. At the end of the study, most of the participants showed positive attitudes towards concordance-based instruction and they enjoyed concordance lines. Most of them thought that they could increase their knowledge with corpus-based teaching of grammar.

Before moving on to suggestions, we should advert that corpora also contribute to language production. Before writing or speaking classes, learners can gain the vocabulary and grammar which helps them to talk or write about a particular topic and during writing, learners can consult corpora to find answers to the questions emerging when they write. This may not be the issue for speaking. Following the speaking and writing activities, both learners and teachers can benefit from corpora in terms of error correction. Corpora encourage students for self-correction. The teacher can discover the existence of student's language use which is seen prescriptively wrong in course books but descriptively right in real life use in corpora. That is, corpora help teacher while they are correcting students. They find answers for the questions emerging during the correction of students work and create new exercises taking corpora as reference so that they can be practiced in the classroom to revise common learner errors (Garcia, 2012).

6. Conclusion and Suggestions

In this paper, we reviewed the basic information about corpus linguistics which is now seen as one of the sub-branches of linguistics and its applications in English language education. First, the definition of the word 'corpus' (plural corpora) was made; then, types of corpora and the most famous corpus systems related to each type were introduced. The different types of use of corpora in ELT were presented shortly. A few examples of research about the practices of corpus use in ELT were also mentioned.

The previous researches on corpora applications in ELT revealed that integrating corpora in language pedagogy is effective for teaching vocabulary and grammar, which are essential language areas to be well mastered in order to be able to write or speak on a particular topic. The result of perception studies also indicated that there are mostly positive attitudes towards the use of corpus-based activities or corpus-based methodology in the classroom. However, Friginal (2018) finds longitudinal studies critical in order to frame the relations between language learning and corpus-based methodology. Similarly, Rasikawati (2019) claims that more research are needed to maintain the support for the efficacy of the approach in various contexts. As for suggestions, he advises the teachers to conduct action research in order to bridge the gap between research and instruction so that teachers may be able to practice the corpus-based DDL approach more and evaluate its efficacy. On the other hand, Çalışkan and Kuru Gönen (2018) and Xodabande and Nazari (2022) state that corpus linguistic course is necessary for in-service EFL teachers because most language teachers are not enthusiastic to incorporate corpus work while teaching or doing exercises. The reason for that is mostly due to not having enough information about corpus and they do not know how to apply it properly in their classes. As a solution to the situation, Çalışkan and Kuru Gönen (2018) propose to revise the teaching programmes' curriculum by integrating a corpus-based pedagogy in both methodology and materials design courses. How to design a corpus-based teaching material should be taught and relevant material design assignments can be given to student teachers. For in-service language teachers, professional development units should organize trainings on how to incorporate corpus into classes in order to create corpus awareness.

REFERENCES

Bennet, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press.

Oxford Learner's Dictionary.(n.d). Corpus. In *Oxford Learner's Dictionaries*. Retrieved July 20, 2022, from https://www.oxfordlearnersdictionaries.com/definition/english/corpus?q=corpus

Çakmak, Z. (2021). *A corpus based study on authenticity of dialogues in the B1-B2 levels ELT coursebooks used in Turkey* [Master's thesis, Ondokuz Mayıs University]. YÖK Ulusal Tez Merkezi.

Çalışkan, G., & Kuru Gönen, S. İ. (2018). Training teachers on corpus-based language pedagogy: Perceptions on vocabulary instruction. *Journal of Language and Linguistic Studies, 14*(4), 190-210.

Davies, M. (2004) *British National Corpus* (from Oxford University Press). https://www.english-corpora.org/bnc/

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/.

UCLouvain. (n.d.). *The International Corpus of Learner English (ICLE)*. Retrieved July,20, 2022, from https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

UC Santa Barbara. (n.d.). *Santa Barbara Corpus of Spoken American English*. Retrieved July 20, 2022, from https://www.linguistics.ucsb.edu/research/santa-barbara-corpus#Intro

Ergül, Y. (2014). The effectiveness of using corpus-based materials in vocabulary teaching [Master's thesis, Pamukkale University]. YÖK Ulusal Tez Merkezi.

Fauzi, A.R. (2020, November 20-21). *Designing an English vocabulary workbook based on corpus-based approach: What actual learning task to incorporate target vocabularies into speaking*. [Paper presentation]. International Conference on English Language Teaching (ICON-ELT), Malang, Indonesia.

Frankenberg-Garcia, A. (2012). Integrating corpora with everyday language teaching. In Thomas, J. & Boulton, A. (Eds.). *Input, process and product: Developments in teaching and language corpora*, (pp. 36-53). Masaryk University Press.

Frankenberg- Garcia, A. (2014). How language learners can benefit from corpora, or not. *Recherches en didactique des langues et des cultures (11)-1*. https://doi.org/10.4000/rdlc.1702

Friginal, E. (2018). *Corpus linguistic for English teachers: New tools, online resources, and classroom activities*. Routledge.

Friginal, E., Dye, P. & Nolen, M. (2020). Corpus-based approaches in language teaching: Outcomes, Observations, and Teacher Perspectives. *Boğaziçi Journal of Education, (37)*1, 43-64.

Gabrielatos, C. (2005).Corpora and language teaching: Just a fling or wedding bells? *The Electronic Journal for English as a Second Language (TESL-EJ), (8)*4.

Granger S. (2008). Learner corpora. In Lüdeling, A. & Kytö, M. (Eds.). *Corpus Linguistics. An International Handbook*, (pp. 259-275). Walter de 5 Gruyter.

Jones, C.,& Waller, D. (2015). *Corpus linguistics for grammar*. Routledge.

Kotani, K. & Yoshimi, T (2015, November 1). *Design of a learner corpus for listening and speaking Performance.* [Paper presentation]. 29th Pacific Asia Conference on Language, Information and Computation, Shangai, China.

Kotani, K., Yoshimi, T., Nanjo, H., & Isahara, H. (2016).A corpus of writing, pronunciation, reading and listening by learners of English as a foreign language. *English Language and Teaching, (9)9.* http://dx.doi.org/10.5539/elt.v9n9p139.

Leech, G. (2004). *Adding linguistic annotation*. In Wyne, M.(Ed.), Developing linguistic corpora: a guide to good practice. AHDS, University of Oxford. Retrieved July 21, 2022, from https://users.ox.ac.uk/~martinw/dlc/chapter2.htm

Linguistic Data Consortium. (n.d.). *Hansard French/English*. University of Pennsylvania. Retrieved July, 20, 2022, from https://catalog.ldc.upenn.edu/LDC95T20

Mukhamadiarova, A. F., Caserta, L.F., Kulkova, M. A., & Reuters, K. (2020). Use of Corpus' Technologies for The Development of Lexical Skills. *Utopia y Praxis Latinoamericana, 25(7), 185-193 https://www.redalyc.org/articulo.oa?id=27964362017*

Meunier, F. (2021). Introduction to Learner Corpus Research. In Tracy-Ventura, N. & Paquot, M. (Eds.). *The Routledge Handbook of Second Language Acquisition and Corpora*, (pp. 23-33). Routledge.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What are they up to? In Anderman, G. & Rogers, M. (Eds.). *Incorporating corpora: The linguist and the translator*, (pp. 18-31). Multilingual Matters.

McEnery, T., & Xiao, R. (2010). What corpora can offer in language teaching and learning. In E. Hinkel(Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 364-380). Routledge.

Osgigian, K. & Chujo, K.(2010).An effective way to use corpus exercises to learn grammar basics in English. *Language Education in Asia, 1*(1), 200-214. http://dx.doi.org/10.5746/LEiA/10/V1/A17/Oghigian_Chujo,

Peksoy, E. (2013). *The corpus-based analysis of authenticity of ELT coursebooks used in high schools in Turkey*[Master's thesis, Dicle University]. http://acikerisim.dicle.edu.tr/xmlui/handle/11468/864

Phoocharoensil, S. (2012). Language Corpora for EFL Teachers: An Exploration of English Grammar through Concordance Lines. *Procedia- Social and Behavioral Sciences, 64, 507-514. https://doi.org/10.1016/j.sbspro.2012.11.060*

Qilichevna, T. M. (2020). Corpus based approach in vocabulary teaching. *European Journal of Research and Reflection in Educational Sciences, (8)*2, 172-176.

Rasikawati, I. (2019). Corpus-based data driven learning to augment L2 students' vocabulary repertoire. *International Dialogue on Education, (6)2, 83-98*.

Roca Varela, M. L. (2012). Corpus Linguistics and Language Teaching: Learning English vocabulary through corpus work. *ES. Revista de Filología Inglesa,* 33, 285-300. https://dialnet.unirioja.es/servlet/articulo?codigo=4546831

Samburskiy, D. (2014) Corpus-Informed Pedagogical Grammar of English: Pros and Cons. *Procedia- Social and Behavioral Sciences, 154,* 263-267 https://doi.org/10.1016/j.sbspro.2014.10.148

Sezgin, H., & Ozturk, M. S. (2020). A corpus analysis on the language on TV series. *Journal of Language and Linguistic Studies, 16*(1), 238- 252. https://doi.org/10.17263/jlls.712787

Sketch Engine. (n.d.). Concordance – a tool to search a corpus. In *Sketch Engine*. Retrieved July,20, 2022, from https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/#toggle-id-2

Sinclair, J. (2004). *Corpus and text- Basic principles.* In Wyne, M.(Ed.), Developing linguistic corpora: a guide to good practice. AHDS, University of Oxford. Retrieved July 21, 2022, from https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Sinha, T.S. (2021). EFL learners' perception of and attitude to corpus as a vocabulary learning tool. *The Reading Matrix: An International Online Journal, (21)*2, 106-119. https://readingmatrix.com/files/25-3ds10cs0.pdf

Şimşek, T. (2020). *Corpora in foreign language teacher education: Introducing a corpus literacy course to ELT pre-service teachers* [Doctoral Dissertation, Çukurova University]. YÖK Ulusal Tez Merkezi.

Szudarski, P. (2018). *Corpus linguistics for vocabulary*. Routledge.

Timmis, I. (2015). *Corpus linguistics for ELT research and practice*. Routledge.

UCLouvain. (n.d.). *The International Corpus of Learner English (ICLE)*. Retrieved July,20, 2022, from https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Wang, X. (2018). Effect of a corpus-based grammar teaching method in the Chinese EFL environment. *Advances in Social Science, Education, Economics and Management Research (ICEEMR)*, *182*, 294-297. http://dx.doi.org/10.2991/iceemr-18.2018.67

Weisser, M. (2016). *Specialised Corpora*. Retrieved July, 20, 2022, from http://martinweisser.org/corpora_site/spec_corpora.html

Xodabande, I., & Nazari, M.(2022). Impacts of corpus linguistics course on in-service

To Cite this Chapter:

Onay Taş, S. & Baykal, N. (2022). Corpus studies and ELT. In A. Önal & K. Büyükkarcı (Eds.), *Essentials of foreign language teacher education*, (pp. 43-61). ISRES Publishing.

ABOUT THE AUTHORS

Inst. Semra ONAY TAŞ

ORCID ID: 0000-0003-3488-0184

semra.onay@dpu.edu.tr

Kütahya Dumlupınar University

Semra ONAY TAŞ is a French Language instructor at Kütahya Dumlupınar University since 2012. She had her bachelor's degree of French Language Teaching from Dokuz Eylül University, Buca Education Faculty in 2011. After having also graduated from Eskişehir Anadolu University, English Language Teaching department in 2021, she started her master's degree education in ELT department at Süleyman Demirel University. Since she gives mostly elective language courses at different departments of the university, her research interests mainly focus on attitude, motivation and enjoyment in foreign language learning. She is also enthusiastic about different language teaching tools and approaches.

Prof. Dr. Nazlı BAYKAL

ORCID ID: 0000-0002-6248-7614

nazlibaykal@sdu.edu.tr

Suleyman Demirel University

Nazlı Baykal is a Professor in the Department of English Language Teaching, Suleyman Demirel University, Isparta, Turkey, where she teaches Linguistics, Language Acquisition, Applied Linguistics courses. Her research interests include the relationship between language and ideology with reference to newspaper and political discourse, issues of language and identity and multimodal teaching materials in ELT