# *A view of corpus linguistics and language teaching*

**ANA CLARA SÁNCHEZ SOLARTE**
**GABRIEL VICENTE OBANDO GUERRERO**
*Universidad de Nariño*

Corpus linguistics is a topic that language teachers, especially EFL teachers should know due to its many possibilities for language instruction. The purpose of this article is precisely to look at the relationship between corpus linguistics and teaching. To do so we will define and look at the historical background of corpus linguistics and some research examples, then the applications and potential limitations of CL will be addressed.

In the past, research in grammar, morphology and syntax has been carried out taking the sentence as the basic element of language. These studies have contributed to bring some organization and coherence to the study of human communication. However, and as the investigation in language teaching and learning advances, the need for the inclusion of context as a vital element for understanding language has risen. From this premise a new trend has developed: corpus linguistics (CL). Even though this is not a new concept in the fields of TESOL, SLA and applied linguistics around the world, teachers might not be familiar with this topic. The purpose of this article is to present basic concepts related to corpus linguistics and to look into the applicability of corpus-based research in the foreign language classroom. We will start this review with a historical overview of the origins of CL.

## Definition and Historical Background of corpus linguistics

According to Conrad (2000), Corpus Linguistics is "the empirical study of language relying on computer-assisted techniques to analyze large, principled databases of naturally occurring language" (p. 548). A

corpus is a large body of language taken from oral (e.g., television or radio shows) or written (e.g., books or newspaper articles) production. The texts are transcribed and compiled in a computer. With the use of specialized software people interested in language are able to search for a specific word or phrase and know what words go before and after them. This information can be useful for deducing and analyzing pre-stated grammar rules. Sometimes as teachers we need more examples to illustrate certain grammar rules or we are not sure if the vocabulary used in certain instances is appropriate. Corpus linguistics is a powerful aid, which can enable EFL teachers to work with examples extracted form discourse produced by native speakers.

It is important to note that even though computers and computer software are regarded as essential tools in the task of collecting and organizing large amounts of information, the existence of a corpus is not subject to the invention of computers. Similarly, corpus linguistics is not a trend that appeared after computers.

Kennedy (1998) classifies the fields in which corpus studies did not include the aid of computers. According to this author the five fields of scholarship were biblical and literary studies, lexicography, dialect studies, language education studies, and grammatical studies. In the field of biblical studies it is important to mention the work by Alexander Cruden. This author took into account the major content words found in the Bible. The purpose of his work was to show that the parts of the Bible were factually consistent with each other.

In the field of lexicography, Kennedy (1998) gives the example of the creation of the Oxford English Dictionary (OED). By the time this dictionary was completed in 1928 it had taken the editors a lifetime to get the compilation of words. The first editors, Murray, Bradley, and two following editors had died before the first edition of the dictionary was completed.

In the area of dialect, Kennedy (1998) cites two important works based on corpus: The English Dialect Dictionary (Wright, 1898-1905) and The Existing Phonology of English Dialects (Ellis, 1889) (p.16). Language education research was also corpus-based. Kennedy (1998) cites the work of Thorndike in 1921, who compiled a total of 4.5 million words from sources like the Bible and classic English works, in order to guide the design of materials for teaching literacy to native

speakers of English in the United States.

In the field of Grammar, Kennedy refers to the work of Fries titled *American English Grammar* (1940) as one of the most complete compilations. He used a corpus of letters written to the United States government. What made this research interesting is that people of different social backgrounds wrote these letters. He analyzed different grammatical aspect (e.g., the past participle "done" used as preterite)

With computer advancements the work of compiling a corpus got to be an easier task. Corpus studies began to be available in order to be used in on going research.

Rundell (1996) states that one of the most recognized corpora is the Brown Corpus containing one million words. It was published in 1961 in the United States. The Lancaster-Oslo-Bergen Corpus (LOB) was also one of the first generation corpora. It became available in 1964. It was a British compilation; it imitated the Brown Corpora in size and in style. According to Kennedy (1998) other corpora modeled the Brown Corpus, adapting them mostly to different varieties of English. For example, the Kalhapur Corpus of Indian English was published in India in 1978 and the Wellington Corpus of Written New Zealand English was published in 1986.

As technology continued to advance, what are considered second-generation corpora began to appear. Rundell (1996) cites some examples of more complete corpora. In 1980, John Sinclair assembled a corpus of 7.3 million words at Birmingham University. This corpus was then expanded to 20 million words. It is relevant to mention that this corpus is important since the editors of the first COBUILD dictionary used it in their listing of words and definitions. This is one example of the way corpus linguistics may affect practical aspects of language learning.

By the end of the decade the Longman-Lancaster Corpus began running with 30 million words. In 1994 the British National Corpus (BNC) contained 100 million words collected from written and spoken texts. The number of words continued to grow from then on, and by 1996 Longman had collected 100 million words complementing the BNC, and the COBUILD Bank of English had 320 million words. This corpus was used to publish dictionaries and series of textbooks for

language teaching.

These examples illustrate the evolution of corpus linguistics over the years, a field of research that is gaining importance all over the world. However, the expansion and recognition of corpus studies is still developing.

## Unpopularity of Corpus Linguistics

During the 1960s and the 1970's corpus linguistics research did not cease, but researchers at that time were trying to work despite Chomsky's theories. There is a great distinction to be made in this matter. Corpus linguistics was viewed differently before and after Chomsky. As stated by McEnery and Wilson (2001) corpus linguistics has been accepted and rejected throughout time, but it is worth mentioning that the early linguistic studies relied almost totally on corpus linguistics. It was the only way language could be analyzed due to the need of a concrete collection of data. Many early studies were based on new ideas; there were no previous data banks that allowed a researcher quick access. People interested in language had to start from scratch and most likely a corpus was the first thing in mind to organize the language.

Studies in the fields of phonetics and language acquisition were based on recorded speech, which are in fact a form of corpus linguistics. The usefulness of corpus linguistics began to be questioned after the ideas proposed by Chomsky. He changed the direction away from empiricism (i.e., a method based on observation and recollection of data) and towards rationalism (i.e., the development of a theory of mind). From this idea, Chomsky argued that corpus linguistics couldn't be a useful tool in linguistics because the research should be done looking at language competence rather than performance (p.6). From this, corpus linguistics "encourages us to model the wrong thing – we try to model performance rather than competence" (p.6).

Another major criticism was the finite characteristic of language. According to Chomsky, as mentioned by McEnery and Wilson, language can never be described by enumerating sentences (p.12). In other words, corpus linguistics can be a waste of time. In addition, Chomsky also argued that "why look through a corpus of a zillion words for facts which may be readily available via introspection?" (p.11).

These notions can be refuted today with simple arguments. The idea that performance is not useful and that introspection is the only base of a study, has been proven wrong by language acquisition studies. In these cases data needed to be recorded precisely, effectively and extensively. The only way changes could be analyzed was to witness the actual evolution that began to occur in the speech of children. Only with a corpus could this information become physical transcriptions of language production rather than abstract ideas.

For instance, Kennedy (1998) cites the existence of the Child Language Data Exchange system (CHILDES), a corpus that consists of some 20 million words and was brought together using recorded speech of over 500 children. According to Kennedy this corpus was used in research conducted by Brown and his students at Harvard in the 1960's and beyond, and in studies conducted by Bloom, Clark, Fawcett, Fletche, Berko-Gleason, Snow, Wells, Slobin and Weir. Corpora played an important role in these research studies since patterns found in early child language could be counted and identified from the available data.

It is clear that language can be analyzed by looking at the information provided in naturally occurring speech that a corpus can easily provide. As McEnery and Wilson (2001) point out "naturally occurring data has the principal benefit of being observable and verifiable by all who care to examine it" (p. 14). We do not have the capacity of reading minds; therefore a corpus is our only chance of studying language. On the other hand, it is not the intention of corpus linguistics to study all the possible producible language. We can analyze the language that is available to us and examine immediate data without a corpus analysis. However aspects such as the frequency of production of a linguistic item can only be studied by looking at a corpus of language. As McEnery and Wilson put it: "if the corpus linguist can often seem the slave of the available data, so the non-corpus linguist can be seen to be at the whim of his or her imagination"(p.15).

Biber, Conrad, and Reppen (1998) discuss the types of research that can be conducted using corpus linguistics. The first aspect to consider, according to these authors, is that the goal of corpus-based investigations is not just to report qualitative findings, but also to look

at complex association patterns in the language. Two main research questions can be obtained in relation to these patterns, one in relation to the use of a linguistic feature (i.e., a lexical item or a grammatical construction) and the other focused on the characteristics of texts or varieties. It is important to note that there can also be non-linguistic associations in this type of research, mainly referring to distribution of features across registers, dialects, and time periods. For instance, in the area of pragmatics and spoken discourse Kennedy (1998) mentions the London-Lund Corpus as a compilation mostly used to conduct this type of research. Altenberg in 1990 used this corpus to study the frequency of discourse items that had "the pragmatic function for planning and structuring interactive discourse, for softening or intensifying what is being said and for provoking feedback through backchanneling" (p. 175). Altenberg studied a total of 50,000-word sample and found a total of 4,516 discourse items dealing with responses, hesitations, softeners, initiators, hedges, expletives (e.g., God, heavens), thanks, apologies, attention signals, response elicitors, politeness markers, orders and others.

It can be seen that corpus linguistics can be used in a wide variety of studies with potential applications in fields ranging form sociolinguistics to vocabulary teaching.

## Corpus Linguistics and the ESL/EFL Classroom

Throughout the years there have been different interests in relation to the process of second language learning. For a while the teacher was the center of attention and the only source of knowledge. This attention was later shifted towards the students. Characteristics like motivation and personality were considered important. Nowadays, a stronger emphasis is made on a balance between what is being taught and the type of instruction the students receive. An important concept that needs to be considered in regard to instruction and learning is noticing. According to Schimidt (1990) a student needs to be aware of certain grammatical structures in the input. The importance of this concept lies in the fact that in order to "acquire" a grammatical structure, a vocabulary item or the syntactic organization of a structure, it is necessary to see how it differs from the L1 or from the previous

knowledge the learners had in the L2. Noticing means realizing how a certain element in the L2 behaves and only by realizing this, it is possible to use that element appropriately. Thus, teachers should be able to provide L2 learners with adequate input to enhance the opportunities for noticing in the classroom. Here is where a corpus and corpus linguistics play an important role. In a study by Sandra Fotos (1992), she concluded that noticing strategies helped the students to be aware of the structures. The problem for many teachers is that they do not have access to naturalistic types of input, therefore they might be providing the students with artificial language and no real noticing can occur. Corpus linguistics can be a helpful source for a more naturalistic type of input. Kennedy (1991) states that teachers need to provide repeated exposure to salient structures of the language, and that corpus linguistics should be the source that provides this information by looking at the frequency and use of these structures. The more the students are exposed to a salient structure, the better the possibility of their internalizing the structure and using it.

It is true that non-authentic materials can also provide enough noticing opportunities, but it might be difficult for teachers to know if the materials contain sufficient real life structures. Many times the language contained in textbooks is artificial. If the teacher does not have access to real language, he or she will simply be teaching an artificial type of language. On the other hand, it might also be the case that the teacher's proficiency is not sufficient for providing realistic examples outside the textbook, as might be the case in many countries, where English is taught as a foreign language. In some other instances, the oral proficiency of the instructors does not provide students with enough output for noticing to take place. So students cannot infer rules from the language used in the classroom, hence the importance of having access to CL. A corpus could make available the necessary language portions that would enable students to analyze the structures and vocabulary used in authentic situations. Furthermore, the students could also take part in the learning, by looking at the results of corpus-based studies in regard to specific grammatical structures that might need additional explanation.

Corpus linguistics is concerned with the occurrence of language

features in natural speech in different settings; for this reason, the authenticity of the input needs to be closely related to register. Conrad (2000) points out that corpus-based research has shown that there are several differences across Standard English. These differences relate to the choice of words depending on the purpose and the situation of use, for example, the vocabulary used by a teenager talking to his friends and the speech the same teenager uses to talk to his/her mother. Even though the vocabulary is basically the same the choice of a word in a specific context may vary. This clearly shows that the type of input needs to change in a classroom depending on the purpose behind communication.

Teachers are aware of the fact that learning a language implies more than knowing vocabulary or grammar. Training students in recognizing these subtleties can also be tackled by using corpora. Carter and McCarthy (1995) mention differences such as age, gender, dialect, and social classes as important parts or influences on native speaker's speech. Therefore, a deeper analysis of a corpus in needed in order to identify these differences. Another feature of teaching that can be reconsidered thanks to the use of corpus linguistics is the teaching of vocabulary. Decarrico (2001) points out that when teaching vocabulary there are two important aspects that need to be taken into consideration: the number of words and the type of instruction. Meara (1995), as cited by Decarrico, affirms that a vocabulary of approximately two thousand words is helpful to achieve lexical competence. Corpus linguistics can certainly aid in this process in order to figure out what are the most commonly used words in a language. A student may not have access to all the words he/she needs. In fact, it could be the case that the classroom instruction is oriented towards vocabulary that might never be used outside the classroom.

Decarrico (2001) claims that vocabulary cannot be learned only in isolation, and suggests that knowing what words co-occur with others is also necessary. She suggests that this concept of collocation can help the learner memorize words and define their semantic areas. Sinclair (1991) also relates the principle of idiom to collocation, which states that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments"(p. 110). This idea

can be helpful for the learner since the learning of certain words can trigger the learning of successive vocabulary items, that is, students should not be pushed to learning isolated lists of vocabulary, but rather try to build word associations in their minds.

Another important aspect in which corpus linguistics can have a future contribution is on the design of educational materials. The following aspect has to do with the incidence of corpus linguistics in the creation, sequencing and design of such type of materials. McEnery and Wilson (2001) argue that there is a need for a critical analysis of textbooks. They cite some studies of this type of analysis which include topics that might have been misleading or incomplete in foreign language textbooks: Quantification and frequency, (Kennedy 1987); doubt and certainty, (Holmes, 1988); future time expressions in German, (Mindt, 1992); and vocabulary in Swedish (Ljung, 1990). The methodology used in these studies was similar. All the researchers compared the vocabulary included in the textbooks to vocabulary obtained from a corpus. As a conclusion they agreed that the textbooks many times provide unreal language, limiting the students' exposure to language that they might encounter in a day-to-day life situation.

These findings are important and teachers are aware of the artificiality of the contents found in textbooks. However, this does not mean that coursebooks will change in a short term. Publishing houses may not be willing to take risks and consider corpus linguistics as a base for the design of textbooks since this might mean huge economic investments. Materials created by teachers on the other hand, can be a good starting point to apply what research shows about lexicon and grammar since they are contextualized, they can be modified and teachers have enough freedom to apply them whenever they need.

The existent gap between the language used in academic life and that used in everyday life was analyzed by Coxhead (2000) who conducted a corpus-based research study in order to evaluate The General Service List (GSL). The GSL was developed by West in 1953 and contained 5 million words related to ESL /EFL learners. For the current study a total of 3.5 million words was used. The corpus was obtained from several types of academic texts, including textbooks, journal articles and academic sections from previous corpora. One research question for

this study was aimed towards the evaluation of the GSL and is stated as follows: "Which lexical items occur more frequently and uniformly across a wide range of academic materials but are not among the first 2,000 words of English given in the GSL?" (p.218).

The findings obtained from this study are that the Academic Word List included 570 word families (e.g., analyze, concept, data, and research were the most frequent word families); 10% of the total tokens in the Academic Corpus, and more than 94% of the words in the list occur in 20 or more of the 28 subject areas of the Academic Corpus. According to the author these findings are important for the creation of future teaching materials, focusing on useful vocabulary items. Coxhead suggests that the research should be oriented towards the comparison of Academic Corpus with larger corpora, obtain information over specific words in relation to their meaning across different subject areas, investigate if the use of technical vocabulary lists is useful or simple reading would be enough, and finally compare the written academic English with the spoken academic English. But teaching the differences between academic and common English is not the only area in which corpus linguistics can help teachers; we sometimes forget that the language we use for speaking is very different from the language we use to write.

McCarthy and Carter (1995) focused on justifying the importance of teaching spoken grammar (SG). They stressed that students need to be given choices between written and spoken grammars and it is suggested that inductive learning has advantages over traditional approaches. The goal of this research is to emphasize the fact that students should be given different choices to be able to interact appropriately at written and spoken levels. In order to illustrate the grammatical forms of speaking and their interpersonal connotations, two samples from the Nottingham corpus are used. A detailed description of the corpus is also provided. After this some grammatical features are analyzed (e.g., reporting verbs, tags) and it is underlined that written grammars do not account for some uses found in spoken data. It is also noted that speaking does feature some expressions that are not likely to be found in writing (e.g., the use of the verb *tend*) and that is why they are not explained. For the authors the main pedagogical considerations are that language can be learned from text instead of created sentences, and

materials should be reformed to include spoken grammar. Two conclusions are drawn: spoken data needs further research, and current methodologies need to be reconsidered to move from presentation, practice and production to illustration, interaction and induction so that learners can improve their communicative skills.

Another important work was the one conducted by Hunston, Francis and Manning (1997). They found that a connection between vocabulary and grammar is evident by focusing on the use of patterns. According to the authors, "words that share the same patterns tend to share aspects of meaning" (p. 208). Most of the verbs with the pattern "V + by + –ing", can have the following meanings: either *start* or *finish* (e.g., with the verbs *begin, close, end, finish*) or *respond to* or *compensate for something* (e.g., with the verbs *atone, counter, react, reply*).

In order to obtain these patterns a corpus of 250 million words was used taken from COBUILD. The background given by the authors suggest that textbooks generally include sections that deal with vocabulary and grammar separately. This may not be really helpful for the students' learning of these structures due to the irregularities of the rules.

The authors suggest that patterns are crucial for four aspects of language learning: understanding, accuracy, fluency and flexibility. If a learner can identify a pattern in a sentence he or she might be able to guess the meaning of an unknown word and develop understanding. Also, if a student is aware of some patters in the language he or she may be able to produce more accurate sentences and also promote fluency. In addition, flexibility can be achieved since a student might be able to choose from the patterns he or she knows well and understands.

These studies can provide a base so that teachers can direct their students to learn grammar more effectively especially oriented more towards meaning.

## Implications

The first conclusion that can be drawn from this article is that corpus linguistics can be said to be a research strategy that is growing rapidly and most likely will continue to grow in the future. The main

reason can be said to be the need to get a real grasp of the language. By using corpus-based research in the classroom teachers are able to have a better control of what they teach. Due to its empirical characteristic of corpus linguistics, it is allowing instruction to make objective statements based on language, rather than presenting affirmations based on perceptions. In this way it is possible that teachers might be able to present better explanations and provide examples that in the long run might help students acquire the language more effectively.

Corpus linguistics is not just counting words or simply providing examples of language with no analysis. Today corpus linguistics has evolved to be used in many linguistic aspects and most importantly it has the potential of changing instruction in the classroom.

A change for classroom instruction that comes from corpus linguistics is the knowledge of frequency, not only at the lexical level but also of structures of the language. Teachers spend too much time concentrating on structures that the student will have little or no access to in a real life communicative situation. With this type of analysis the teacher will be able to make choices and construct a more realistic syllabus.

The technological advancements today allow a more organized type of research. The size of a corpus is not an evident problem anymore today. Computers can help the researcher organize, count and categorize data that is essential.

Research has offered many possibilities in different areas of linguistics, but are these findings really changing the practice within the classroom? It is not enough to say that certain teacher materials are written using corpus linguistics or that a textbook is written taking into account corpus-based approaches, what is really needed is a creation of new syllabi that includes this information.

Other areas of linguistics like sociolinguistics may also benefit from corpus linguistics studies. For example, there are differences in the language produced by males and females. Registering these differences will help a great deal in research regarding preferred style, speech communities and learner differences.

Pedagogical decisions should be made considering what the results of research in corpus linguistics tell us. It is important that learners are exposed as much as possible to the structures and expressions of a

language, but special focus should be done on the structures that frequently appear throughout corpora. Focusing on this will not only provide learners with reliable information about the use of a language, but it will also save time and

The results of research in corpora need to be systematized and accessible to teachers. It is valuable to have information about vocabulary or collocations, but this data should produce visible changes in the professional practice of teachers.

In EFL settings corpus linguistics may be a great help in order to provide learners with sufficient, contextualized, authentic materials. However, due to the long process of collecting, classifying and analyzing language samples, it is very difficult for schools to have access to this important resource. Thus, one of the future concerns for EFL teachers, administrators and researchers in countries in which English is taught as a foreign language should be either the creation of a corpus or taking the necessary steps to have access to databases.

## THE AUTHORS

Ana Clara Sánchez Solarte is an assistant professor at the University of Nariño. She holds an M.A in TESOL form the University of Northern Iowa. She currently teaches writing classes and EFL methodology.

Gabriel Vicente Obando Guerrero is an assistant professor at the University of Nariño. He holds an MA in TESOL from the University of Northern Iowa. He is charge of conversational classes and teacher training classes.

## REFERENCES

Bahns, J. (1993). Lexical collocations: A contrastive view. ELT Journal, 47, 56-63.

Bauer, L. & Renouf, A. (2001). A corpus-based study of compounding in English. Journal of English Linguistics,29, 101-123.

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. New York:

Cambridge University Press. Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics.

Applied Linguistics, 15, 169-189.

Botley, S., & McEnery, T. (2001). Proximal and distal demonstratives: A corpus-based study. Journal of English

Linguistics, 29, 214-233. Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. Applied Linguistics,16, 141-158.

Chafe, W. (1991). The importance of corpus linguistics to understanding the nature of language. In J. Svartvik (Ed.), Directions in corpus linguistics. Berlin: Mouton de Gruyter.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? TESOL Quarterly, 34, 548-559.

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34, 213-238.

Decarrico, J. (2001) Vocabulary learning and teaching. In M. Celce-Murcia (Ed.), Teaching English as a second or foreign language (pp.285-298). Boston, MA: Heinle & Heinle.

Fotos, S. (1992). Consciousness raising and noticing through focus on form: Grammar task performance versus formal instruction. Applied Linguistics, 385-407.

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connectors. ELT Journal, 51, 208-216.

Kennedy, G. (1991). Preferred ways of putting things with implications for language teaching. In J. Svartvik (Ed.), Directions in corpus linguistics. Berlin: Mouton de Gruyter.

Kennedy, G. (1998). Introduction to corpus linguistics. New York : Longman.

McCarthy, M., & Carter, R. (1995). Spoken grammar: what is it and how can we teach it? ELT Journal, 40, 207-217.

McEnery, T., & Wilson, A. (2001). Corpus linguistics: An introduction. Edinburgh: Edinburgh University Press.

Mukeherjee, J. (2001). Principles of pattern selection: A corpus-based study. Journal of Linguistics,29, 295-313.

Rundell, M. (1996). The corpus of the future, and the future of the corpus [On-line]. Available: http://www.ruf.rice.edu/~barlow/futcrp.html

Schmidt, R. (1990) The role of consciousness in second language learning. Applied Linguistics, 11, 129-158.

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