

The Effectiveness of Construction Corpus in EFL Classroom in Fostering English Proficiency

Dr. Elsadig Ali Elsadig Elnadeef

Dr. Ayman Hamad Elneil Hamdan

King Khalid University,
College of Art and Sciences –Dhahran Aljanoub

Abstract

The study presents the significance of implementing corpus linguistics in English classroom. It focuses on defining corpus linguistics, aims of corpus linguistics, what we need to construct a corpus, types of corpora, using corpora in English classroom, target features of corpus and processes of corpus construction. The study accentuates that corpus linguistics focuses on linguistic performance, rather than competence and it focuses on linguistic description, rather than linguistic universals and corpora can be brought to the classroom in different ways, soft or hard version. Moreover, the study explains that corpus linguistics aims at discovering pattern of authentic language use through analysis of actual usage and language pattern in terms of lexical, structure, lexis, grammatical, discourse, phonological and morphological aspect.

Key words: Corpus linguistics, corpora, concordancing, colligation, token , lemma

1.0 Introduction

English teacher proficiency involves mastering particular declarative knowledge and procedural knowledge detailed as: pedagogical content knowledge, linguistic knowledge, curriculum and syllabus knowledge, psycho-social knowledge y, and technological knowledge. Linguistic Knowledge requires tracing the progressive diversion trend of linguistics. Linguistics has gradually moved from using words and clauses as the unit of analysis to considering texts as a whole and finally to seeing texts as instances of construction of knowledge and ideology. Cognitive and functional approaches assume that language should not be studied in isolation, but in the context of actual communicative interactions. This implies that corpus linguistics is an important tool for working within the cognitive-functional framework.

2.0 Definition of Corpus Linguistics and Corpora

Corpus linguistics is the study of linguistic phenomena through large collection of machine readable texts and it is concerned with the analysis of data obtained from a corpus. Moreover, it makes lists of words in a particular text with their context and this process is known as concordancing. It is also concerned with colligation which refers to combining texts into a

corpus and collecting patterns based on syntactic group. It is concerned with the relation between micro-linguistic events and macro-social structures by drawing attention to patterns at the micro-linguistic level. (Tognini, 2001). Corpus linguistics focuses on linguistic performance, rather than competence and it focuses on linguistic description, rather than linguistic universals. Moreover, it focuses on quantitative, as well as qualitative models of language and it is based on empiricist, rather than rationalist view of scientific inquiry. (Leech 1992) .A corpus is a large principled systematic collection of naturally occurring examples of language or texts or pieces of language stored electronically and selected and ordered according to explicit linguistic criteria in order to use as a sample of language and they have been computerized (Sinclair 1996).

3.0 Aims of corpus linguistics

Corpus linguistics aims to discover pattern of authentic language use through analysis of actual usage and language pattern in terms of lexical, structure, lexis, grammatical, discourse, phonological and morphological aspect. It only concerns in the usage patterns of the empirical data and what reveals to us about language behavior. (Simpson and Mendis ,2003)

3.1 Searching corpora provides answers to questions like these:

What are the most frequent words and phrases in English?

What are the differences between spoken and written English?

What tenses do people use most frequently?

What prepositions follow particular verbs?

How do people use words like can, may and might?

Which words are used in more formal or informal ones?

How often do people use idiomatic expressions?

How many words must a learner know to participate in everyday conversation?

How many different words do native speakers generally use in conversation?

4.0 What do we need to construct a corpus?

- A personal computer (Windows , MAC- Linux..)
- Use concordance program to analyze plain text files (extension text)
- Know language and basic ideas in corpus linguistics e.g. Word list –frequencies –type – token- concordance (Sinclair, J.1996)

-Techniques used include generating frequency word list and concordance lines, collocate, cluster and key lists.

Word list: usually arranged from highest to lowest frequencies of types

A type: unique form of a word

A word: defined as a running letters separated by space and punctuation

“To be or not to be, that is the question”

To- be- or- not- to – be- that- is- the- question

It has 8 types (to/be) occur twice

It has 10 token (count every word)

5.0 Types of Corpora

1. General or reference corpora aim to represent language in its broadest sense and it provides a sample data from which we can make generalization about spoken and written discourse as a whole and frequencies of occurrence, and co-occurrence of particular aspects of language in discourse

2. Specialized corpora explain a corpus of texts of a particular type, such as newspaper, editorials, geography textbooks academic articles in a particular subject, lectures, casual conversations, essays written by students etc. it aims to be representative of a given type of text. It is used to investigate a particular type of text.

3. The Michigan corpus of academic spoken English (MICASE) An example of specialized corpus that is designed with a particular research project in mind.

MICASE is normally an open access corpus and it is available without charge to people who wish to use it (www.isra.umich.edu/eli/micase/index.htm) .

4. The British academic spoken English (BASE) corpus (www.rdg.ac.uk/Acadepts/II/basecorpus/). It has been developed at the University of Warwick .

5. The British academic written English corpus BAWE or the British academic written English is a specialized corpora that is based on the written discourse alone, developed at the university of Warwick and the university of Reading and Oxford university in the UK (www.Warwick.ac.uk/fac/soc/celte/bawe/).

6. The TOEFL spoken and written academic language corpus which includes both spoken and written discourse.

7-Learner corpora is a kind of specialized corpora contains written texts and spoken transcripts – International Corpus of Learner English-ICLE= Standard Speaking Test Corpus SST comprise of oral interview tests of Japanese learners

8-Pedagogic corpora contain language used in classroom settings. It can include academic textbooks, transcripts that learners encounter in an educational setting and it can be used to ensure learners are learning useful language, to examine teacher-student dynamics or self-reflective tool for teacher development.

9- A monitor corpus is a dataset which grows in size over time and contains a variety of materials. The relative proportions of different types of materials may vary over time such as The Bank of English (BoE) and the Corpus of Contemporary American English (COCA).

10-Balanced corpora is known as sample corpora, try to represent a particular type of language over a specific span of time.

11- Parallel Corpus which is a collection of texts that has been translated into one or more other languages.

(Teubert, 2005)

6.0 Using Corpora in English Classroom

Corpus linguistics is used as primary data for developing dictionaries and grammars, informing textbooks and other teaching materials, preparing materials for classes (e.g. as a source of examples), data-driven learning in the classroom and studying learner language, developing new theories. (Thomas and Short,1996).

Using of corpora in the English classroom is usually referred to as Data Driven Learning (DDL). It is considered as a branch of CALL. It attempts at raising learners' awareness of the target language by teaching the students to discover behavior of language items (Breyer, 2008).

Corpora can be brought to the classroom in different ways –soft or hard version-Leech 1997): In the soft version, it the teacher who has to access to corpora, and he hast to master the skill of using it and software. On the other hand, leaners need to be equipped with computer skills to left and the right context.

6.1 Processes of Corpus Construction:

- Sorting concordance: teaching the possibility of sorting concordance according
- Finding a lemma- student possibility of finding the whole lemma of a word

- Extending the context- illustrate the possibility of exploiting small specialized corpora for traditional linear reading
- Comparing synonym – observe two different lexemes in the same concordance and it was suggested this could be applied to analysis differences between synonym.
- Using regular expressions emphasize the advantages offered by POS tag corpus
- Collocation-external based on the collocation potential of particular words
- Colligation –frequency of the occurrence of word with particular part of speech (it is collocation of lexical and grammatical element)-guiding question-make a concordance of word ---note down its frequencies
- Word class –connection between grammar and lexis

(Fuentes and Rokowski, 2002)

7.0 Creating corpora

When creating corpora, there must be three factors: the corpus must be principled, it must use authentic texts, and it must have the ability to be stored electronically.

In creating the corpus, you need to identify and understand grammatical patterns in English from authentic texts which are defined as those that are used for a genuine communicative purpose. Learner corpora can be created from the compilation of student work taken from particular class for one particular student, or from a series of students and classes. Students can analyzed their language use and pinpoint areas that need further instruction or document progress that has been made.

8.0 The Corpus Approach

The corpus approach is comprised of 4 major characteristics:

- 1- It is empirical, analyzing the actual patterns of language use in natural texts.
- 2- It utilizes a large and principled collection of natural texts as the basis for analysis
- 3- It makes extensive use of computers for analysis
- 4- It depends on both quantitative and qualitative analytical techniques

9.0 Target Features

Corpora focuses on certain areas of language teaching such as phraseology (study of phrases and determined the meaning of a word is found through several words in sequence through

phrase. Phraseology includes the study of collocation, lexical bundles(it is a frame which has set words around a variable word or words- we use a frame to express future –going to- is expected to –is supposed to – is about to- is due to ----lexical bundles are recurring sequence of three or more words ---idioms have a meaning not derivable from their parts unlike lexical bundles) and language occurring in preferred sequences) , lexico-grammar(there is no difference between lexis and grammar-they are intertwined that they cannot be productively studied separately e.g. know-matter-suppose) , registers(register is defined as situation of use-we use different language with different audiences// pronouns is used slightly more than nouns but nouns are used significantly more than pronouns in fiction, news and academic writing .

10.0 Problems with language on the web

- Biased distribution of text-types and genres
- Repeated and reused text
- Unknown provenance (who wrote this, when, why?)
- Native and non-native producers of language
- Mixture of varieties
- Unclear separation of linguistic elements of the webpage
- Accessing the hidden web
- Accessing language embedded in audio and video streams
- Repeated and reused text
- Lack of persistence of source data and location identifiers
- Unknown sampling and ranking strategies of search engines.

(Simpson and Mendis ,2003)

11.0 Problems with language in the corpus

- Limited size
- Expensive, time-consuming and slow to make
- Usually limited to out of copyright texts
- Not up to date
- Design decisions were made by someone else

- Not comparable to other corpora
- Access restrictions
- Difficult to deploy in the classroom
- Often, only restricted online access with limited functionality, processing, and not connected to other resources

(Orpin, 2005).

References

- Breyer, Y. (2008). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, 22(2), 153-172.
- Fuentes, A. C. & Rokowski, P.E. (2002). Use of corpus in the business English classroom: From lesser to more specific. Paper presented at the IVACS Conference, June 2002: Ireland, UK.
- Leech, G. (1992). Corpora and theories of linguistic performance. In J. Startvik (Ed.), *Directions in corpus*.
- Orpin, D. (2005). Corpus linguistics and critical discourse analysis. Examining the ideology of sleaze. *International Journal of Corpus Linguistics*, 10(1), 37-61.
- Simpson, R. & Mendis D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 27 (3), 419-441
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. McH. (1996) Preliminary recommendations on corpus typology' EAGLES Document TCWG-CTYP/P
- Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), 1-13.
- Thomas, J. and Short, M. (eds.). (1996). *Using corpora for language research*. London and New York: Longman.
- Tognini Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.