English Journal
Vol. 14, No. 2; September 2020, pp. 75-83

CORPUS LINGUISTICS, LANGUAGE CORPORA AND LANGUAGE TEACHING

Eska Perdana Prasetya[1], Anita Dewi Ekawati[2], Deni Sapta Nugraha[3], Ahmad Marzuq[4], Tiara Saputri Darlis[5]
English Education Program Faculty of Teacher Training and Education, Universitas Ibn Khaldun Bogor[1]
English Education Program Faculty of Teacher Training and Education, Universitas Muhammadiyah Prof. Dr. Hamka[2]
Sekolah Tinggi Penerbangan Indonesia[3]
Department of Arabic Language Education, Faculty of Languages and Arts, Universitas Negeri Jakarta[4]
Students of Applied Linguistic Study Program, Universitas Negeri Jakarta[5]
eska@uika-bogor.ac.id[1], anita.dewieka@uhamka.ac.id[2], deni.sapta@ppicurug.ac.id[3], ahmad.marzuq@unj.ac.id[4], tiarasaputridarlis@gmail.com[5]

ABSTRACT

This research is about Corpus Linguistics, Language Corpora, And Language Teaching. As we know about this science is relatively new and is associated with technology. There are several areas discussed in this study such as several important parts of the corpus, the information generated in the corpus, four main characteristics of the corpus, Types of Corpora, Corpora in Language Teaching, several types that could be related to corpus research, Applications of corpus linguistics to language teaching may be direct or indirect. The field of applied linguistics analyses large collections of written and spoken texts, which have been carefully designed to represent specific domains of language use, such as informal speech or academic writing.

Keywords: *Corpus Linguistics, Language Corpora, Language Teaching*

INTRODUCTION

The development in the science of Language Education is certainly related to one science, namely linguistics, modern linguistics is now getting livelier with the existence of a branch of knowledge, namely the corpus and language corpora. Both sciences are closely related to linguistics. Why studying linguistics is very important, because linguistics is related to language, and language is the main communication tool in every human being. In everyday life, humans use various forms of language to meet their daily needs. There are four broad forms of language, namely reading, writing, listening, and speaking. The most important human need is to be able to communicate with other people because humans are social creatures who cannot live alone.

According to Adolph (Hizbullah, 2016), The linguistic corpus itself is the development of modern linguistics enlivened by the emergence of a relatively "new" branch of science, namely corpus linguistics. So, it can be concluded that linguistic corpus is a science that studies a set of data that is natural, real according to its use, be it written data or oral data that is transcribed. Meanwhile, according to (Baker, 2010) 'Corpus linguistics is only a quantitative approach, just useful for identifying general patterns but not for any in-depth qualitative analysis.

From the above opinion, it can be concluded that Corpus linguistics is the study of language data on a large scale - computer-aided analysis of a very extensive collection of transcribed speech or written text and that a corpus is only a quantitative approach, only useful for identifying general patterns but not for analysis. qualitative profound.

The word corpus may sound foreign to our ears, plus corpora in learning. This is especially so for those who are not linguists or linguists. The corpus is a collection of authentic texts, both written and transcripts of large amounts of conversation that are stored electronically. The corpus has been used frequently in various areas of linguistic research.

THEORETICAL BACKGROUND

Corpus Linguistic

There are several opinions regarding the linguistic corpus, the first opinion from (T McEnery., 2012) explains that Corpus linguistics is the study of language data on a large scale - the computer-aided analysis of very extensive collections of transcribed utterances or written texts. As for the opinion of (S Dash, 2010) The uniqueness corpus linguistics lies in its way of using modern computer technology in collection of language data, methods used in processing language databases, techniques used in language data and information retrieval, and strategies used in application of these in all kinds of language-related research and development activities. And the opinion of (Keck, 2012) Corpus linguistics is an area of applied linguistics that uses computer technology to analysed large collections of spoken and written texts, or corpora, which have been carefully designed to represent specific domains of language use, such as informal conversation or academic writing.

From the three opinions above, it can be concluded that corpus linguistics is:

Corpus linguistics learns about Language with the help of:

a. Modern computer technology in language data collection

b. It is simply a collection of widely described speech

c. There are methods and techniques used in processing and developing language databases

d. The field of applied linguistics analyses large collections of written and spoken texts, which have been carefully designed to represent specific domains of language use, such as informal speech or academic writing.

Corpora Language

For this section, at least the author gets some explanation from several sources. According to (Bennet., 2010) There is also an opinion from (Flowerdew, 2009). In addition, the linguistic. An important aspect related to creating corpora is the issue of copyright, especially if findings from a corpus will be distributed via a handout or published in any form. content of corpora is different from what is experienced by individuals in real life, many of them consist largely of written language. The final opinion from (Ide, 1998) By far the greatest need for the development of linguistic corporations is to ensure their usability and reusability in integrated platforms.

From some of the explanations above, it can be concluded that language corpora are one of the important aspects related to language corpora. It is a copyright issue, especially if the findings from the corpus will be disseminated through handouts or published in any form. In addition, the linguistic content of the corpora is different from what individuals experience in real life, many of which mostly consist of written language. By far, the greatest need for corporate linguistics development is to ensure its usability and reuse in an integrated platform.

Corpus Linguistics dan Language Corpora

In this section of the discussion, the author will first begin with some important parts of the corpus. According to (S Dash, 2010) there are several important parts of the corpus, namely:

A. Quantity, what is meant by Quantity here is that it must be in a large size containing a lot of data both in oral and written form.

B. Quality, all texts must be obtained from samples of actual speech and writing. The role of a linguist is very important here.

C. Representation, this should include samples of various texts.

D. Simplicity, it must contain plain text in a simple format.

E. Equality, the sample used in the corpus must be an even size.

F. Retrievability is Data, information, examples, and references should be easily retrieved from the corpus by end-users. It is concerned with language data preservation techniques in electronic format on computers.

G. Verifiability, the corpus must be open to any kind of empirical verification. We can use the data from the corpus for any kind of verification. This puts corpus linguistics one step ahead of the intuitive approach to language study.

H. Augmentation, this should be improved regularly. This will place the corpus to record the linguistic changes that occur in a language over time.

I. Documentation, the complete information of the components must be separated from the text itself. It is always better to keep documentation information separate from text and to include only a minimal header containing a reference to the documentation.

METHOD

In this study, the authors used a literature study, namely by searching for various written sources, such as e books and articles in a journal, as well as documents that are relevant to the problem being studied. So that the information obtained from this literature study is used as a reference to strengthen existing arguments.

FINDINGS AND DISCUSSION

The information generated in the corpus

According to (Flowerdew, 2009) A corpus can yield various types of information that can be of potential use in language pedagogy. It can provide information about the behaviour of words, multi-word phrases, grammatical patterns, semantic and pragmatic features, and textual property. According to him, there is also some additional information

in the corpus that can be produced in the corpus.

a. Word frequency

In the word frequency section, there are three things contained in this section, namely 1) Very useful in helping to prioritize what is taught. 2) The existence of relevant criteria such as reach, availability, coverage, learning ability, and prototype. 3) It is necessary for the learner to obtain various registers and genres, then the existence of comparative data, as provided by keyword analysis, which will provide some information.

b. Collocation

Collocations relate to how words usually appear (or don't appear) together. The repeating pattern highlighted by the concordance will show typical collocations, although the program can provide a list of collocates.

c. Colligation

A distinction can be made between collocation, which is a combination of individual words, and colligation, which refers to how lexical words are associated with certain words or grammatical categories. Hunston (2002, p. 13), once again gives an example of the word head which has the following colligations: of, over, on, back, and off. Once again colligations affect the meaning of the word, so Hunston gives examples such as department heads, hitting someone's head, dropping their heads back.

d. Semantic preference

The main focus of this section is how a word or phrase relates to a group of overlapping words which (1) has an element of meaning, (2) is related to a particular. genre or register, or (3) belonging to the lexical set in terms of synonym, meronym, antonym.

e. Semantic prosody

This kind of information is now beginning to be included in the dictionary, but learning activities in the form of analysing words in context and identifying their semantic prosody may be a more effective learning strategy, to the extent that learners are more likely to remember what they have discovered on their own.

According to (Bibber, D. Ripper, 2014) the corpus consists of four main characteristics:

1. The Corpus Approach is empirical, analysing the actual patterns of language use in natural texts. In this section several things can be explained, namely:

a. Language is an authentic language.

b. The idea of principled corpora has already been mentioned.

c. Examples of corpora themselves consist of textbooks, fiction, nonfiction, magazines, academic papers, world literature, newspapers, telephone conversations at home or at work, cell phone conversations, business meetings, class lectures, radio broadcasts, and TV shows, among other acts of communication.

d. In short, every real-life situation in which linguistic communication occurs can form a corpus.

2. The Corpus Approach utilizes a large and principled collection of naturally occurring texts as the basis for analysis.

From this approach, it can be concluded that the characteristics of the Corpus Approach refer to the corpus itself. You can work with the written

corpus, the oral corpus, the academic oral corpus, etc.

3. The Corpus Approach makes extensive use of computers for analysis.

The characteristics of the Corpus Approach refer to the corpus itself. You can work with the written corpus, the oral corpus, the academic oral corpus, etc.

4. The Corpus Approach depends on both quantitative and qualitative analytical techniques.

These characteristics of the corpus approach highlight the importance of our intuition as linguist users.

Since the electronic corpus is new, we have yet to come to a common consensus on what is considered a corpus, and how it should be classified. The classification scheme I'm proposing here is as far as prudent at this point. This offers a plausible way of classifying corpora, with clearly defined categories where possible.

Types of Corpora

As noted earlier, a corpus is not simply a collection of texts that have been randomly selected without reference to some kind of framework that allows for balance and sampling; corpora is also not an opportunistic collection of texts. According to Hunston (Baker, 2010), the important differences made by many corpus linguists are of two kinds, namely between general corpora (sometimes referred to as reference) and special corpora.

a. General corpora

A general corpus can be seen as a prototype corpus that is usually very large, consisting of millions of words, with texts gathered from various sources, representing many linguistic contexts (written, spoken, electronic, public, private, fiction, non-fiction).

b. Special corpora

The corpora reference is often used in conjunction with specific corpora - the former provides information about the 'norms' of language which can then be compared with the latter to identify what is relatively frequent or infrequent in the more specialized variations of the language.

Corpora in Language Teaching

a. Corpus linguistics and language teaching

In this article, the authors will describe the relationship between corpus linguistics (CL) and language teaching (LT) and provide an overview of the most important pedagogical applications of corpora. As the Figure below wants to illustrate according to (Römer, 2009), the image below shows that this relationship is a dynamic relationship in which two fields greatly influence each other.

Language Teaching benefits from the resources, methods, and insights provided by Corpus Linguistics, it also provides an important impetus taken in corpus linguistic research.

b. Types of pedagogical corpus application

When we talk about the application of the corpora in language teaching, this includes the use of corpus tools, for example actual text collections and software packages for corpus access, and corpus methods, for example the analytical techniques used when we work with corpus data.

c. Corpus linguistics application for language teaching

According to Stubs in (Flowerdew, 2009) explains that Applications of corpus linguistics to language teaching may be direct or indirect. Direct application is a user at an advanced level of English at an academic level who uses a special language corpus to assist them in writing academic papers. While indirect applications are the application of corpus findings for the creation or improvement of dictionaries, reference grammar, and pedagogic materials, below is an explanation of direct and indirect applications:

1) Indirect applications

Use in developing reference material

In indirect applications, there are two sub-sections discussed, namely: Use in developing reference material and Pedagogic materials, one of the first applications in this field is Collins COBUILD English Language Dictionary (1987) edited by John Sinclair ; Other dictionaries have used the corpora to a greater or lesser degree, for example, Longman Dictionary of Contemporary English, Oxford Advanced Student Dictionary, Macmillan Dictionary of English for Advanced Students. As Leech (1997, p.14) points out, some of the advantages of corpus-based lexicography are that corpus data:

• can be searched quickly and completely,
• can provide frequency data,
• can be easily processed to generate updated word lists,
• can provide authentic examples for quotes,
• Can be easily used by the lexicographic team to update and verify other level descriptions such as dictionary definitions.

2) Direct Applications

In this section it consists of Corpora and syllabus design and Data-driven learning If one accepts the corpus view of language, that is, it consists of mostly repeating patterns (what Sinclair, 1991 calls the "idiom principle"), then important implications apply to syllabus design. Instead of being organized in grammatical form, the syllabus can be designed around the most important repeating patterns (see Sinclair & Renouf, 1988; Willis, 1990; Willis & Willis, 1989). This type of syllabus is referred to as a lexical syllabus, although it is somewhat misleading, as it is designed around a lexical pattern, not singular. words.

In this section the writer gets various materials that can be presented regarding research related to this corpus. There are several types that could be related to corpus research. Among others are

a. Computational Linguistics (Computer Linguistics)

According to (Musthofa, 2010) states that Computational Linguistics has four meanings as follows: 1. A scientific study of language from a computer point of view. 2. An interdisciplinary study of language involving natural language and computers. 3. A study that focuses on natural language processing with its various phenomena automatically by computers. 4. The study of engineering computer application programs for natural language processing with various complexities.

From the above explanation, it can be concluded that computational linguistics can be related to language from a computer point of view, interdisciplinary

studies of natural and computer anatomy, natural language processing and computer engineering.

b. Cultural Studies

According to Edward in (Rai, 2015) Culture is that complex whole which includes knowledge, belief, art, morals, law, custom, and other capabilities and habits acquired by man as a member of society. According to the explanation above, it can be concluded that cultural studies is related to knowledge, beliefs, arts, morals, law, customs, and other abilities and habits acquired by humans as members of society.

c. Discourse Analysis and Pragmatics

According to (Deborah Schiffrin & Hamilton, 2001) Discourse analysis is frequently equated with conversational analysis, and pragmatics with speech act theory. And regarding pragmatics further they add pragmatics is sometimes said to encompass discourse analysis - or the reverse. It has been suggested that discourse analysis is more text-cantered, more static, more interested in product (in the welcomeness of texts), while pragmatics is more user-cantered, more dynamic, more interested in the process of text production.

In the above statement it can be concluded that Discourse Analysis is often equated with speech analysis, and pragmatics with speech act theory. pragmatics is sometimes said to include discourse analysis - or vice versa.

d. Language Acquisition

There is an explanation of Language Acquisition according to (Aljoundi, 2008) language acquisition is a matter of growth and maturation of relatively fixed capacities, under appropriate external conditions has been criticized by Sampson (2005) and Gethin (1999) on the role of adult speech which cannot be ruled out to help children in working out the regularities of language for themselves.

From the explanation above we can conclude that Language Acquisition is a matter of growth and maturation of capacity that is relatively fixed, under appropriate external conditions it has been criticized by Sampson (2005) and Gethin (1999) regarding the role of adult speech that cannot be ruled out to help children in practicing language regularity for themselves.

e. Lexicography

The author gets the theory of lexicography according to (Bergenholtz & Gouws, 2012) Lexicography is divided into two related disciplines:

• Practical lexicography is the art or craft of compiling, writing and editing dictionaries.

The theoretical lexicography is the scholarly discipline of analyzing and describing the semantic, syntagmatic and paradigmatic relationships within the lexicon (vocabulary) of a language, developing theories of dictionary components and structures linking the data in dictionaries, the needs for information by users in specific types of situation, and how users may best access the data incorporated in printed and electronic dictionaries.

In the explanation above, the authors conclude that lexicography can be seen from two sides, namely practical lexicography and theoretical lexicography. When viewed from the practical factor of lexicography of arts or

crafts composing, writing, and editing dictionaries. However, when viewed from theoretical lexicography, it analyses and describes the semantic, syntagmatic, and paradigmatic relationships in language lexicons (vocabulary).

f. Social Psychology

Social Psychology according to (DANSABO, 2015) social psychology is that aspect of psychology that explores the relationship between the individual's behaviour and the specific social situation in which the individual is operating. Social Psychology is an aspect of psychology that explores the relationship between individual behaviour and certain social situations in which the individual operates.

g. Stylistics

Stylistics according to (Tariq, 2018) Stylistics is the study of language and the language of literature in all its manifestation and is, one of the moderate approaches to literature. It can be said that Stylistics is the study of language and literary language in all its manifestations and is one of the moderate approaches to literature.

h. Speech

Speech according to (Abdullah, 1988) 'speech communication' deals with aspects of communication theory as well as the practice of speaking in social contexts with special reference to ESL (English as a Second Language). The author can conclude that speech relates to aspects of communication theory as well as the practice of speaking in a social context with special reference to ESL (English as a Second Language).

Meanwhile, there are several types of speech, according to (Telg, 2011) Speeches can be divided into the following categories:

1. the informative speech

The goal of an informative speech is to provide information completely and clearly so that the audience understands the message.

2. The persuasive speech

Persuasive speeches are given to reinforce people's beliefs about a topic, to change their beliefs about a topic, or to move them to act.

3. Speeches for special occasions

Speeches for special occasions can be informative, persuasive, or both, depending on the occasion. Two of the more common types of speeches for special occasions are the speech of introduction and the speech of welcome.

CONCLUSION

In the discussion of Corpus Linguistics and Corpora, we can discuss several sub-themes, including the definition of the corpus and corpora itself, the point is a linguistic method that uses data from language materials collected in a source. From this understanding, it can emerge and explain the scope of the Corpus Linguistics and Language Corpora, the types of corporations, the information generated in the corpus, the corpora in language teaching and research related to the corpus. the field of applied linguistics that analyses large collections of written and spoken texts, which have been carefully designed to represent specific domains of language use, such as informal speech or academic writing. Computational linguistics can be related to language from a computer point of view, interdisciplinary studies of natural

and computer anatomy, natural language processing and computer engineering.

REFERENCES

Abdullah, F. S. (1988). Speech Communication. *British Telecom Technology Journal*, *6*(2), 7–21.

Baker, P. (2010). Sociolinguistics and corpus linguistics. *Sociolinguistics and Corpus Linguistics*, 1–189.

Bennet., G. R. (2010). Using CORPORA in the Language Learning Classroom: Corpus Linguistics for Teachers. *Language Value*, *2*, 140–143.

Bergenholtz, H., & Gouws, R. H. (2012). What is lexicography? *Lexikos*, *22*(July), 31–42. https://doi.org/10.5788/22-1-996

Bibber, D. Ripper, D. (2014). English corpus Linguistics. In *English Corpus Linguistics*. https://doi.org/10.4324/9781315845890

DANSABO, M. T. (2015). *INTRODUCTION TO SOCIAL PSYCHOLOGY* (pp. 1–59).

Deborah Schiffrin, D. T., & Hamilton, and H. E. (2001). *The Handbook of Discourse Analysis Edited*. https://doi.org/10.1007/978-3-030-56711-8_3

Flowerdew, J. (2009). Corpora in Language Teaching. *The Handbook of Language Teaching*, *June*, 327–350. https://doi.org/10.1002/9781444315783.ch19

Hizbullah, N. (2016). *Pembelajaran Bahasa Arab Di Indonesia* *. 385–393.

Ide, N. (1998). Encoding Linguistic Corpora. *Proceedings of the Sixth Workshop on Very Large Corpora*, 9–17.

Keck, C. (2012). Corpus Linguistics in Language Teaching. *The Encyclopedia of Applied Linguistics*, 2002–2005. https://doi.org/10.1002/9781405198431.wbeal0256

Musthofa, M. (2010). COMPUTATIONAL LINGUISTICS (Model Baru Kajian Linguistik dalam Perspektif Komputer). *Adabiyyāt: Jurnal Bahasa Dan Sastra*, *9*(2), 247. https://doi.org/10.14421/ajbs.2010.09203

Römer, U. (2009). Corpora and language teaching. *International Journal of Corpus Linguistics*, *14*(4), 549–556. https://doi.org/10.1075/ijcl.14.4.05buy

S Dash, N. (2010). Corpus Linguistics : A General Introduction Niladri Sekhar Dash. *Corpus Linguistics: A General Introduction.*, *c*, 25. http://www.academia.edu/919592/Corpus_Linguistics_A_General_Introduction

T McEnery., A. H. (2012). *Corpus Linguistics:Method,Theory and Practice*.

Tariq, M. (2018). *Style , stylistics and stylistic analysis : A re-evaluation of the modern-day rhetorics of literary discourse*. *March 2018*, 46–50.

Telg, R. (2011). *Speech Writing and Types of Speeches 1*. *August*, 2–4.