Parole: Journal of Linguistics and Education, 10 (2), 2020, | 176-181

Available online at: https://ejournal.undip.ac.id/index.php/parole

Corpus Linguistics and Corpus-Based Research and its Implication in Applied Linguistics: A Systematic Review

Ali Mohammed Saleh Al-Hamzi*, Ayoub Gougui, Yuni Sari Amalia, Totok Suhardijanto

Airlangga University, Surabaya, Indonesia

ABSTRACT

This article conveys a case-of-systematic survey of outstanding progress on corpora conducted by researchers affiliated with different common-section institutions all over the world. Such a range overview selected 20 outstanding types of research from multi research-pushing institutions all around the world. These projects employ corpus techniques and technology to treat an enormous domain of research queries that are relevant to linguistic studies, language teaching and learning, cultural studies, and discourse analysis. These varied implementations of corpus techniques and advances clearly explain the great stress and chances that corpora applied in linguistics can hand to those who have the intention to research, educate, and learn the language.

ARTICLE INFO

Paper type:
Systematic Review

Article history:
Received 22/06/2020
Revised 06/08/2020
Published 26/10/2020

Keywords:
- Corpora
- Corpus linguistics
- Applied linguistics
- Cultural studies
- Linguistic studies

1. Introduction

The implementation of a corpus in the field of linguistics has grown rapidly over the decades. Corpus in the simple term is defined as the compilation of texts that has been gathered for a specific reason (Cheng, 2011). Despite its definition above, the definition of corpus linguistics as a theory or a methodology has become a 'polemic' among the linguists. Some argued that corpus linguistics cannot be considered as a particular domain of research as it is just a methodological basis for language study (Tognini-Bonelli, 2001). Lindquist (2009) stated that corpus linguistics is not a branch of linguistics because "corpus" does not inform you what is studied. Instead, a methodology comprises a wide range of several related methods that can be used by researchers of various theoretical learnings. Besides, Weisser (2016) also argued that corpus linguistics is a methodology that provides the development of our insights about the way language works by 'consulting' real-life data. Therefore, it is obvious that we cannot learn how to conduct corpus research based on a theoretical basis solely. On the contrary, other scholars agreed that corpus linguistics is more than just a pure methodology (Tognini-Bonelli, 2001). Halliday (1993c:24) as cited in Tognini-Bonelli (2001), for instance, asserts that corpus linguistics combines the activities of data collection, theorizing, and argues which these kinds of activities lead to a qualitative change in the way we understand language.

In this regard, the increase of corpus mechanisms and systems has qualified those who have great intentions to do studies all over the world to carry out research in their various localities with lower obstruction. Through years, corpus mechanisms have moved out the domain of verifiable research in relevance to linguistic studies and language education to the use of corpora to serve certain functions in conducting research in specific areas of study. In fact, it has been progressively widespread for researchers to use corpus technology in their

E-mail Address: abudhia55@gmail.com *(Ali Mohammed Saleh Al-Hamzi),* ayoubtg@gmail.com *(Ayoub Gougui),* yuni.sari.amalia2017@gmail.com *(Yuni Sari Amalia),* tugasgakusei@gmail.com *(Totok Suhardijanto)*

research as well as compiling their own corpora for particular goals. This paper conveys a case-of-systematic survey of outstanding progress on corpora conducted by researchers to address a diverse range of topics in their latest academic efforts, concentrating particularly on how their fundamental studies have been managed in Applied Linguistics.

2. Research Methods

The essence of this review is on the modern corpus and corpus-based research conducted by researchers affiliated with different common-section institutions all over the world. The study was bounded to input obtainable in common scopes, such as scholarly databases, journals, and books. Each selected newly published journal and outstanding study is checked out to select studies in relevance to how language corpus has been employed by those who conduct studies to treat a different sequence of themes in their recent scholarly efforts, concentrating particularly on how their essential studies have been fulfilled in Applied Linguistics. When a researcher taking part in corpus-based research is recognized, his/her names are employed as key terms to deal for research offspring in basic scholastic data. A total of 20 are divided based on focus. Table 2 and the list of references supply information of the 20 studies for those who have great intentions to read. The 20 studies are published outputs (books, journal papers), or ongoing work from 2000 onward. Based on their situations at the time of this study (May 2020), obtainable points for every presentation and project may differ. Outputs of published research are all often accessible, whilst outstanding projects usually have restricted information. Additional studies were conducted for those entries with finite inputs through diverse engines of searches and academic data utilizing keywords and names of authors. In these lines, the 20 ongoing and outstanding studies of corpus and their implication in applied linguistics are briefly described.

3. Results and Discussion

Extending across an enormous domain of themes, the 20 studies are categorized based on target readers or their major research attentions. Six majors are set; data contained in table 2 shows some studies distributed for every major. 12 Studies made to introduce teaching and learning of language are categorized in the “teaching and learning” group. Furthermore, 5 Studies are classified under “Linguistic research”, 1 study with its cultural major, and 2 Studies are put under “Corpus & Discourse Analysis. It is also noticed that some studies have some attention and thus can be classified under more than two sets. Regarding the facility of show and grasp of the research domain, we just treat the essential major of every study while classifying.

Table 1. Division of the 20 Studies by Focusing on Their Main Investigators

Teaching and learning	Linguistic research	Cultural studies	Corpus & Discourse Analysis
Barker (2010) Biber et al. (2012) Boulton & Tyne (2015) CHEN et al. (2014) Conrad et al. (2009) Fuster Márquez & Clavel Arroitia (2010) K Hyland et al. (2012) K Hyland & Wong (2013) Ken Hyland (2013) Lee & Webster (2012) Ma et al. (2012) Walsh (2010)	Gries (2009) Kaszubski (2003) Moisl (2015) Cheng (2011) Lukin et al. (2017)	Wiegand & Mahlberg (2019)	Baker (2006) Partington & Marchi (2015)

Table 2: Division of Studies Depending on Their Majors

Teaching and learning	Linguistic research	Cultural studies	Corpus & Discourse Analysis
12	5	1	2

3.1. Teaching and Learning

A sum of 12 studies has employed corpora to respond to a research question binding to language teaching and learning, containing one edited book (Ken Hyland, 2013). Projects under this umbrella show the most considerable category. These projects generally resolved the basic corpora or corpus, contrasted, and added notes upon it with a view to reinforcing learners' realization of particular language terms. For example, the study themed "A corpus of textual revisions in second language writing" by (Lee & Webster, 2012) used a corpus of second language learners' writing from the researcher's own academy. These structured texts involve tutor's returns, authentic drafts, and attached rewritings of the learners' written thoughts. Incorrect actions made in second language learners' written communicating thoughts are counted for their appearance then categorized at levels of text classification. This study highlights the steps of writing and introduces the layout of writing assistive devices.

Besides, to study the main shifts that corpora have fetched to language education, first, however, Hyland (2013) in his study titled "Corpora and innovation in English language education" presents a requisite arranger on what corpora are and how they are employed. Corpora have been at the front of two of the most important variations in language education in latest years. On one aspect they have supplied teachers and materials preparers with more firm depictions of how language is designed and employed, disclosing the prevalent appearance of phrasal units as the foundation of idiomatic language use. On the other hand, they have simplified modern methodologies of teaching, sharing in a broad change from teaching as telling knowledge to teach as authentic learning. At present, corpora supply a means for applicants to occupy a more robust and reflexive part in their learning by discovering genuine instances of language. In the latest decade, then, we have seen corpora hard-done-by in domains as various as syllabus plan (Walsh, 2010), methodologies of classrooms (Ken Hyland, 2013), student grammar (Conrad et al., 2009) and assessments (Barker, 2010) in what has been something of an uprising in situations to language teaching.

Many projects also review relevant factors related to the use of corpora in higher education. In a certain study titled "Corpus Linguistics and its Applications in Higher Education" by Fuster Márquez & Clavel Arroitia (2010), they are set out to depict implied essentials of corpus linguistics and its progress in relevance to theoretical linguistics and its implementations in modern teaching pursuits. In these projects, they try to set up how diverse types of corpora have shared the progress of direct and indirect programs in language teaching. They identify Data-Driven Learning because of its relation to applied linguistics literature and check in detail merits and demerits. Finally, they resume problems in connection with the application of corpus linguistics in the classroom since knowledge of the restrictions of corpus linguistics is necessary for its coming prosperity.

Another study focused on "Corpus-based study of language and teacher education" by Boulton & Tyne (2015). This study indicates that Corpora have still to make fundamental progress in the field of teaching. Corpora absolutely include the resolutions to all issues relating to teaching, and enormous employment with learners is probable to have the opposite effect. However, the directory recommends that they can supply a further collection of devices and mechanisms for a diversity of aims for fully several learners in some situations, can develop their knowledge of the language and metacognitive dexterities, constructing and stimulating existing information, which leads to life-long learning.

Furthermore, the application of corpora in the field of language teaching and learning conveys a case-of-systematic survey of outstanding progress on corpora conducted by researchers to address a diverse range of topics in their latest academic efforts, concentrating particularly on how their fundamental studies have been managed in Applied Linguistics. In this relevant topic, a number of articles work toward studying and presenting several types of research and projects in language corpora. It is, therefore, categorized into two primary sections in relation to their focus. On one hand, it normally supplies a preface to the domain of corpus linguistics and its high relevance to language teaching and learning. On the other, it conveys a concise summary of some concerning articles of the domain. Stepping from the articles presented at the 4th International Conference that was conducted to convey many topics in the domain of Corpus Linguistics (CILC2012, Jaén, Spain). These studies have several characteristics in general. They all use overall implementation of corpora and at the same time deal with topics relating to teaching and learning, the implied proposition being that a real and alternately useful link can be founded between teaching and research. With regard to this source, one and all frames a clear taking of the way that various corpora can be taken advantage of several ends.

Finally, two studies concentrated on improving systems of learning to reinforce learning depending on regarding corpus. Here students of Mainland /Hong Kong learn some lessons to improve their pronunciation (Cheng, 2011) that is depending on the learner's pronunciation corpora. The purposes of the suit are to characterize frequent pronunciation difficulties of leaners, along with providing remedies to learners and instructors in order to develop the accurate pronunciation of those who learn English. Another system relevant to corpus-based learning was existed to improve lexical-grammatical terms learning for majors of the English language (Ma et al., 2012). This framework focused on a number of certain corpora that have academic ends.

3.2. Linguistic Research

In relevant to linguistic research, there are 5 studies involving two improved books on corpus linguistics and corpus-based research in language studies (Cheng, 2011) and the book titled “Cluster Analysis for Corpus Linguistics” By Moisl (2015). In this book, Corpus linguistic-specific implementation using grouping is depicted as a Web application aimed specially to ease explorations in quantitative dialectology – or dialectometry – by giving researchers in dialectology to manage computer-supported explorations and computations even if they have comparatively little computational experience. Two projects also discussed different phenomena in linguistics or special characteristics of the language. One study by Kaszubski (2003) makes the use of corpora in applied linguistics. It indicates that Corpus linguistics is heterogeneous. Many users consider it as a technological rise to classic linguistic methodologies—a good approach nevertheless one not taking into consideration the certainty that corpora lighten both what is known in language and what is still unknown. Some other corpus linguists therefore devote themselves to line new clue in corpora to challenge existing theories. It is potential to identify yet a third approach, which indicates the progress of devices for text processing. Corpus linguists of this standing resort to link less significance to the quality of a corpus or the fineness of findings, and more to the competence and power of computer software. Another study done by Gries (2009) indicates that corpus linguistics is one of the rapid-rising systems of methods in contemporaneous linguistics. Furthermore, It investigates some of the centric presumptions (‘formal distributional differences reflect functional differences’), concepts (corpora, representatively and balancedness, markup and annotation), and methods of corpus linguistics (frequency lists, concordances, collocations), and discusses a few ways in which the discipline still needs to mature.

Besides, two studies have concentrated on specific language arts. In addition, Lukin et al., (2017) worked toward showing a new corpus, PersonaBank, composed of 108 specific stories from weblogs that have been commented with their story intention schemas, a profound picture of the fabula of a story. The themes of the stories and the base of the story intention scheme picture are depicted, along with the process of commenting on the stories to make the story intention schemes and the difficulties of conforming the tool to this new private narrative scope. It is also investigated how the corpus can be employed in implementations that tell the story differently using various styles of telling, co-telling, or like a content planner.

3.3. Cultural Studies

One project has a cultural focus. In this book titled Corpus Linguistics, Context, and Culture by Wiegand & Mahlberg (2019) Corpus Linguistics, Context and Culture explain the possibility of corpus linguistic methods for discussing language manners across a domain of contexts. Arranged in three parts, the chapters extend from accurate case studies on lexico-grammatical patterns to essential talks of meaning as section of the 'discourse, contexts, and cultures' topic. The last part of 'learner contexts' particularly indicates the requirement for mixedmethod approaches and the regard of pedagogical implying for actual-world contexts. Beyond its assistance to existing discussions in the domain, this edited volume emphasizes fresh trends in cross-disciplinary work.

3.4. Corpus & Discourse Analysis

Discourse Analysis can also profit from corpus linguistics research. Two studies have made the use of corpus linguistic research to reinforce the capacity and efficiency of discourse analysis. One study titled "Using corpora in discourse analysis" by Partington & Marchi (2015) indicates that the most explicit merit of merging corpus purses into discourse analysis is the possibility it displays of dissecting large numbers of symbols of any specific discourse type, which helps the analyst to explore representative discourse bodies, representative ways of stating things, and representative messages, besides the local bodies, meanings, and messages obtainable to classic primary reading. It also supplies a way of determining possibly motivating linguistic features – for example, positions of remarkable evaluation – in a large content of texts, which the analyst can then home in upon. In addition, it eases arbitrage among discourse types, shedding light on the relative frequency and the potential various roles of the linguistic features they present, for example, variations in collocational patterning or "profile" of the "same" lexical items or group of items. The other research (Baker, 2006) employing Corpora in Discourse Analysis discusses programs for performing discourse analysis using mechanisms that are placed in corpus linguistics. Earlier research on critical discourse analysis has concentrated on studying a single text or tiny bodies of texts, but researchers engaging in critical discourse analysis are embarking on acknowledging the possibility of employing corpora either to complete their feedbacks or as a good methodology in itself. A corpus-based approach enhances to supply a quantitative guide of the presence of discourses by helping researchers to state unvarying linguistic types of language employment and to reveal invisible meanings in lexical terms e.g. by checking how specific words are put together in a recurrence better than chance. Furthermore, corpus linguistics permits researchers to reveal linguistic evidence for prevailing/majority and resistant/minority discourses as a large corpus is likely to present a chain of ideological places - something which an analysis of a single text may be less likely to uncover.

4. Conclusion and Suggestion

This systematic review presents a glance at the latest studies that are done by researchers from various geographical locations all over the world. The variety of topics, the copra volume employed and languages studied to provide us a powerful tone that corpus and corpus-based research is a lifelike domain of research. Though we are not in a set to consider potential research trends, common types were regarded among the 20 studies.

In manifestations of languages studied, English language corpora are employed for the plurality of the studies (12 out of 20), while several studies use corpora of language (one out of 20) limited to the cultural studies. Meanwhile, corpora of Linguistic research (five) are used in some of the studies. The remaining two indicate the employment of corpora in discourse analysis. The language employed in the study includes both the spoken and written forms and involves a broad series of types like naturally occurring dialogue, academic writing, literary works, and media texts.

To sum up, a considerable numeral of noticeable corpora has been used and or improved by those who have great intentions to conduct researchers for their specific study. There are many instituted corpora used by the researchers. On top of that, several corpora were put together by the researchers themselves as part of their studies, which proves that language corpora have been improved and/or applied, by researchers to treat a various range of topics in their recent well-read efforts especially when their primary research projects have been conducted in Applied Linguistics.

References

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Barker, F. (2010). How can corpora be used in language testing? In *The Routledge handbook of corpus linguistics* (pp. 661–674). London: Routledge, https://doi.org/10.4324/9780203856949.

Biber, D. E., Reppen, R., & Friginal, E. (2012). Research in corpus linguistics. In *The Oxford Handbook of Applied Linguistics, (2 Ed.)*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195384253.013.0038

Boulton, A., & Tyne, H. (2015). Corpus-based study of language and teacher education. *The Routledge Handbook of Educational Linguistics*, 301–312.

Chen, H. C., Chan, K. Y., Wong, P. M. J., Chee, E., Wang, L., & Wang, Q. (2014). A corpus-based online pronunciation learning system: The Pedagogical applications of a spoken corpus for improving Hong Kong/Mainland university students' English pronunciation. *The Second Asia Pacific Corpus Linguistics Conference (APCLC 2014)*.

Cheng, W. (2011). *Exploring Corpus Linguistics: Language in Action*. London: Routledge, https://doi.org/10.4324/9780203802632

Conrad, S., Biber, D., Daly, K., & Packer, S. (2009). *Real grammar: A corpus-based approach to English*. New York. Pearson Education ESL.

Fuster Márquez, M., & Clavel Arroitia, B. (2010). Corpus linguistics and its aplications in higher education. *Revista Alicantina de Estudios Ingleses*, (23), pp.51-67.

Gries, S. T. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Hyland, K, Chau, M. H., & Handford, M. (2012). *Corpus Applications in Applied Linguistics*. London: Continuum. http://dx.doi.org/10.5040/9781472541611.

Hyland, K, & Wong, L. L. C. (2013). *Innovation and change in English language education*. London: Routledge, https://doi.org/10.4324/9780203362716.

Hyland, Ken. (2013). Corpora, innovation and English language education. In *Innovation and change in English language education* (illustrate, pp. 218–232). London. Routledge.

Kaszubski, P. (2003). Corpora in Applied Linguistics. *ELT Journal*, *57*(4), 416.

Lee, J., & Webster, J. (2012). A corpus of textual revisions in second language writing. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 248–252.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh. Edinburgh University Press.

Lukin, S. M., Bowden, K., Barackman, C., & Walker, M. A. (2017). Personabank: A corpus of personal narratives and their story intention graphs. *ArXiv Preprint ArXiv:1708.09082*.

Ma, Q., Wang, L., He, A., & Decoursey, M. (2012). A corpus-based online learning system: Improving undergraduates' use of lexico-grammatical items. *Research Information Core Hub*.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Berlin. De Gruyter Mouton.

Partington, A., & Marchi, A. (2015). Using corpora in discourse analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (Cambridge Handbooks in Language and Linguistics, pp. 216-234). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139764377.013

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* . Amsterdam: J. Benjamins. https://doi.org/10.1075/scl.6

Walsh, S. (2010). What features of spoken and written corpora can be exploited in creating language teaching materials and syllabuses. In *The Routledge handbook of corpus linguistics*.

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. John Wiley & Sons. doi:10.1002/9781119180180

Wiegand, V., & Mahlberg, M. (2019). *Corpus Linguistics, Context and Culture*. Berlin. De Gruyter Mouton.