To Data Driven Research and Beyond: An Overview of Corpus Linguistics Research in Indonesia

Hamamah

{hamamah@ub.ac.id}

Faculty of Cultural Studies, Universitas Brawijaya

Abstract. Nowadays, corpus materials can be sourced from diverse linguistic data from any language in the world and play a vital role as data source in many disciplines, even in non-linguistic fields. This study wants to navigate to what extent corpus-based research in Indonesia has developed from 2011-2022 in the Google Scholar database. The findings suggest that 405 corpus-related articles in Indonesian context still revolve around corpus construction and analysis; only few studies utilize corpus for language learning, such as DDL and foreign language learning. Not to mention that documentation of Indonesian local languages has not been done collectively, and rather individually, making it challenging for other researchers to navigate local language corpora. By giving insights of the development of corpus research in Indonesia, it is hoped that many people are inspired and encouraged to carry out research and empirical applications of corpus and corpora outside the field of linguistic discussion.

Keywords: corpus research, data-driven research, language documentation, local languages, Indonesian corpus

1 Introduction

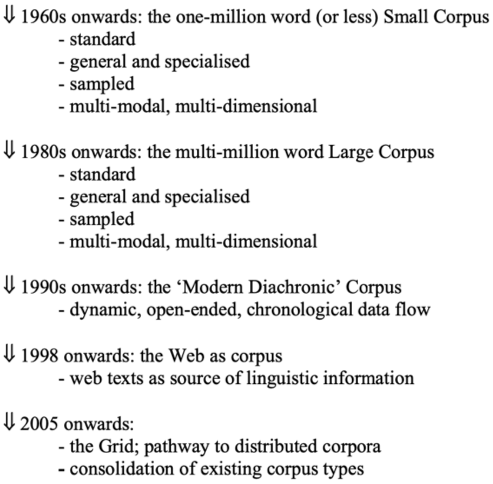
The realm of language documentation is currently inseparable with corpus building, with corpus linguistics as a subdiscipline in the late 1950s is often associated as a new empirical approach, in which computer being the fundamental tool in its collection and construction [1], [2]. Today, corpus as a collection of texts, either spoken or written, gravitates towards several nuances, inter alia, "…machine readable form, sampling representativeness, finite size and the idea that a corpus constitutes a standard reference for the language variety it represents" [1, p. v]. Hitherto, corpus linguistics is not necessarily inquiring a particular element of a language, i.e., phonology, phonetics, or morphology, but "… a set of procedures, or methods, for studying language" [2, p. 2], in which the data is gathered from corpus, "… and theorizing are no longer separate activities, but the textual instance is valued as a window on to the linguistic system" Halliday, in [3, p. 18].

It is popularly acknowledged that modern computerized corpus construction started in the spirit of documenting real-life textual instances of American English, with the publication of

ICOLLEC 2022, November 11-12, Malang, Indonesia
Copyright © 2023 EAI
DOI 10.4108/eai.11-11-2022.2329401

Computational Analysis of Present-Day American English in 1967 by Henry Kučera and W. Nelson Francis based on the Brown Corpus (1961-1967) pioneered by the two linguists. Following the Brown Corpus, W. Nelson Francis with four other founders, Geoffrey Leech, Jostein Hauge, Arthur O. Sandved, Stig Johansson, and Jan Svartvik, constructed the International Computer Archive of Modern English (ICAME), which is in itself a corpora, where scholars from several institutions around Europe and America contribute to its development [3, p. 23]

From there onwards, English became a primary language to be machine documented in the spirit of collecting native speakers' repertoire for one greater purpose: to comprehend how English grammar evolves. Svartvik [3] gives example of how the data in the Survey of English Usage Corpus was collected from many textual and spoken utterances, but all revolves around educated speakers or recorded from academic events. Similar practices are then adopted by other "small corpora", such as the Lancaster-Oslo-Bergen Corpus, and London-Lund Corpus, with the characteristics being amounting to only one-million word or less, and documenting standard variety of general and specialized English collected from multi-modal and multi-dimensional medium [4]. Following similar characteristics, with additional criteria such as diverse linguistic parameters—including 'principled selection' to ensure balance in representativeness as well as applying 'random' vs. 'non-random' sampling techniques—and demographic parameters (i.e., age, gender, social class and ethnicity, and research needs) for comprehensive metadata, multi-million-word large corpora, or super-corpora were then developed in the early 1980s, i.e., the Birmingham Corpus, the Bank of English, and the British National Corpus [4].



⇓ 1960s onwards: the one-million word (or less) Small Corpus
- standard
- general and specialised
- sampled
- multi-modal, multi-dimensional

⇓ 1980s onwards: the multi-million word Large Corpus
- standard
- general and specialised
- sampled
- multi-modal, multi-dimensional

⇓ 1990s onwards: the 'Modern Diachronic' Corpus
- dynamic, open-ended, chronological data flow

⇓ 1998 onwards: the Web as corpus
- web texts as source of linguistic information

⇓ 2005 onwards:
- the Grid; pathway to distributed corpora
- consolidation of existing corpus types

Fig. 1. English language corpus evolution [4].

In the 1990s, corpus was evolved to reflect the changing nature of language, and is much more dynamic, open ended, and includes chronological data flow. Such a corpus is called "Modern Diachronic", and two notable examples include the RDUELS unit at Birmingham using *Times* newspapers and at Liverpool using *Independent* newspapers, both dating back to 1988. Renouf [4, p. 36] mentions that the construction of modern diachronic corpus was mainly "scientific", or "theoretical", with the spirit of inquiry to monitor the chronological change in variety, grammar and lexis of a language over time. With the discovery of the Internet, in 1998 and the years that follow, web texts are finally seen as valuable sources to build and develop corpora of any purposes, as the Web becomes ubiquitous with linguistic materials to analyze. By considering the expansive language variety and the sporadic language phenomenon existing within the virtual interactions, not to mention its collection feasibility as well as the continuously updating language use, many researchers look upon it as great opportunity to broaden their linguistic analysis beyond face-to-face interactions [4]. While of course there are practical and theoretical issues to take into account, especially the fact that web texts are nonetheless real-time data which can "…form an arbitrary and instantial corpus that changes like the sand with each new search", the demand of its construction grows, as there appears a possibility of obtaining utterances in text which were observed too difficult to collect in conventional text corpora [4, pp. 42–43].

Despite English texts (and transcriptions) being the first established work of modern and computerized corpus linguistics (see more McEnery and Hardie, 2013; Svartvik, 2007), textual and spoken instances of any language can be documented into a corpus and analyzed using corpus linguistics. They even play vital role as a data source in many disciplines, as has been projected, inter alia, in a compilation of papers entitled *Corpus linguistics around the world* edited by Wilson et al., [5]. In the book, corpus-based research is no longer focusing on linguistic fields, but also penetrating the fields of cross-cultural rhetoric, social psychology, and economy. Other languages have also interestingly being documented in corpora, such as Basque, Chinese, Danish, Dutch, German, Maltese, Russian, Slovene, and Spanish [6].

That being said, considering how English (and its variety) is no longer the primary language to document, corpus is no longer categorized by its quantity, but by its collected utterances, as well as purpose of analysis and collection instead. Bennett [7] mentions that there are eight major types of corpora used extensively in today's corpus-based research, i.e., generalized, specialized, learner, pedagogic, historical, parallel, comparable, and monitor. Each corpus certainly has its own functions and aims in any scientific research that utilizes corpus as its data source. For example, in the context of data-driven language learning, generalized, specialized, learner, and pedagogic corpora may be possibly used to assist foreign language learners for real-life examples of how specific words and phrases are used in different contexts. Learner corpora, for instance, can as well be a window to document the sporadic distinctions and unique language use by a group of language learners around the world [7].

As aforementioned, non-linguistic fields can still benefit from the development of corpus building, however, the main idea of corpus building itself is to document language. The subject of language documentation itself has already been an interesting topic in Indonesian academia, as Indonesia is home to hundreds of local languages and scholars have tried their best to document local languages to revitalize it. However, many specialized corpora are still individually collected by linguists and researchers from all around Indonesia; they have not been collectively pooled in one-stop website that people can access easily for their research purposes. Currently, there are several corpora documenting Indonesian language and its variety from various institutions and researchers:

1. **Korpus Indonesia** (2018): this corpus is compiled by the Indonesian Ministry of Education and Culture
2. **Postag Indonesia** (around 2016)

3. **Leipzig Corpora Collection – Indonesian:** this corpus is compiled by Leipzig University, source material from Indonesian web texts since 2013, an example of Web corpus
4. **SEALang Library Indonesian Text Corpus** (around 2010-2011): this monolingual corpus compiles Indonesian texts retrieved from a variety of Internet sources
5. **IDN-ENG IDENTIC** (2012): an example of parallel corpus, usually for translations
6. **Kamus Alay – Colloquial Indonesian Lexicon** (2018): source material from Instagram
7. **Indonesian Speech Recognition** (2012): an example of spoken corpus
8. **Tokyo Institute of Technology Multilingual Speech Corpus (TITML) – Indonesian (TITML-IDN)** (2011): this corpus is developed for acoustic models training for automatic speech recognition system in Indonesia
9. **OpenSubtitles Indonesian** (2018): this corpus has limited access through SketchEngine subscription which documents Indonesian movie subtitles
10. **Indonesian Web Corpus (INDNESIANWaC)** (2010): this corpus can be freely accessed through SketchEngine
11. **C-SMILE (Corpus of State University of Malang Indonesian Learners' English)** (around 2019): an example of learner corpus
12. **CINTA (Corpus of Indonesian Texts in Academia)** (around 2019): an example of specialized pedagogic corpus.

With this understanding, this study wants to navigate how far corpus-based research in Indonesia has been from 2011-2022. This study utilizes Google Scholar database to perceive the dissemination of corpus-based research in Indonesia and its association with linguistic and non-linguistic fields or disciplines. It is decided that the timeline is deemed suitable for obtaining such information as corpus in Indonesia has only started flourishing around 2010 with the appearances of SEALang Library Indonesian Text Corpus and Indonesian Web Corpus (INDONESIANWaC). It is expected that through the findings of this study, Indonesian researchers are more encouraged to utilize corpora in their data-driven scientific venture, especially in multidisciplinary research, and specifically, to build corpus to document local languages in Indonesia, especially ones in the verge of extinction and loss.

2 Method

This study utilizes computational mapping analysis method on a Google Scholar database within the period of 2011-2022 to navigate around research articles written in the context of corpus research in Indonesia. The data was first retrieved using Publish or Perish, a software program to obtain and analyze raw academic citations [8]. Several keywords were input in their designated columns to retrieve the metadata of the articles. In the "title words" column, 'corpus' was written; in the "keywords" column, 'corpus AND Indonesian AND Indonesia' was written.

This needed to be done to make sure that the articles retrieved are about corpus research in Indonesia on Indonesian language (and its varieties).

When the data was finally mined, Publish or Perish provided a table containing information on the metadata of each article that matched or met the designed threshold. The metadata includes the citation number and total of citation per year, rank, name(s) of the author(s), title, year of publication, publication site/journal, publisher's name, and type of article. It was then saved in the RIS/RefManager format in the form of bibliographic data to be processed in the software tool VOSviewer to provide better bibliometric visualization. Noun phrases from the title and abstract of each article would then filtered down to exhibit co-occurrence network of terms, which would then appear as nodes in the visualizations [9]. 2442 terms appeared as the result, and by applying the rule of minimum co-occurrence of terms to at least 4 times, 108 terms met the threshold and used in the visualization.

3 Findings and Discussion

This study finds out several discoveries related to corpus research in Indonesia. According to the Google Scholar database, there are 405 academic articles available which are either corpus-based or use corpus as their data source or corpus linguistics as their primary research method/procedure in the span of 2011-2022. One of the earliest works is written by Larasati et al., [10] where they described their invention of a morphology tool for Indonesian language called MorphInd. The paper is cited by 97 other articles (8.82) per year since it was published.

As aforementioned, when processed in the VOSviewer, 108 terms were used in the mapping, with 'corpus' occurring as many as 219 times. The terms 'Indonesian' and 'Indonesia' follow in the third and fourth place with 104 and 86 occurrences respectively. This was foreseeable as in the initial stage of data collection, 'corpus', 'indonesian', and 'indonesia' were included as a set of keywords to mine the articles from the Google Scholar database. It is also possible that the term 'study'—following in the second place—is exhibited to occur 158 times as 'study' often appears as a frequent synonym to 'research', or in the sentence within the abstract "this study…", or in the article title "a corpus-based study". These four terms also happen to have the lowest relevance scores, making them the most general keywords occurred in the 405 articles. While van Eck and Waltman [11] state that general terms i.e., 'conclusion', 'new method' and 'interesting result' do not contribute much, these four terms are the primary nodes to understand to what direction corpus research has been in Indonesia, and therefore, despite their low relevance score representing their generality, are still significantly valuable.

On the other hand, several terms related to corpus research in Indonesia that sporadically occur—often lower than 10 occurrences—and have the highest relevance score include 'natural language processing', 'local language', 'madurese', 'corpus assisted critical discourse analysis', 'islam', and 'ddl'. Referring to the explanation given by van Eck and Waltman [11, p. 31] that "terms with a high relevance score tend to represent specific topics covered by the text data… and the focus shifts to more specific and more informative terms", these highly relevant terms then can be assumed to represent the rare research interests related to corpus in Indonesia. On top of that, the fact that 'madurese' and 'local language' occur only 3 times and 4 times respectively only corroborates the lack of corpus research on Indonesian local languages, or construction of local language corpora in Indonesia, compared to corpus documenting Indonesian language as listed in the previous chapter.

Moving on, the overlay visualization processed in the VOSviewer demonstrates that while corpus research in Indonesia had started in 2011, it was not until late 2017 that research utilizing corpus became much more ubiquitous. The term ‘sign language’ in the far right (see Figure 2), for example, with its weak link strength, only occurs 3 times in corpus-related research published around 2014. It is possible that there was an attempt to develop a sign language corpus in Indonesia, or a documentation of Indonesian Sign Language around 2014. Another example worth noting is the appearance of ‘parallel corpus’, which is closely connected with ‘corpus’, ‘indonesia’, ‘indonesian text’, ‘malay’, and ‘chinese’, occuring 12 times in corpus-related research published around 2016. This is an interesting find considering in 2012, there is indeed a Indonesian-English parallel corpus for translators called IDN-ENG IDENTIC, making it highly possible that there were attempts to make another parallel corpus with 3 different languages—Indonesian, Malay, and Chinese (although Indonesian is technically the standardized variety of Malay).

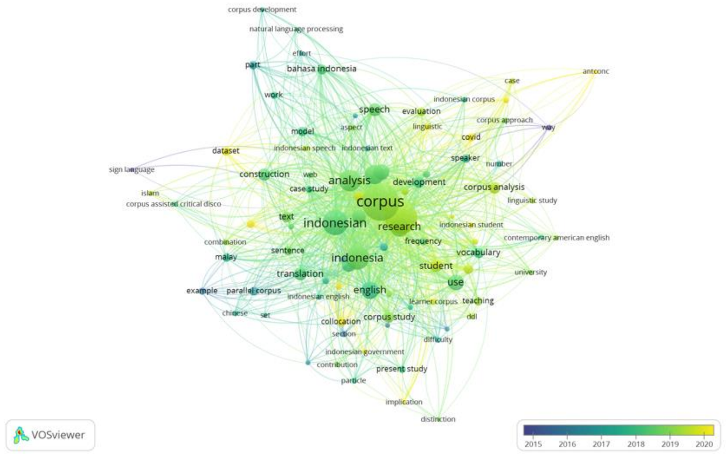


Fig. 2. Overlay Visualization from VOSviewer.

Furthermore, building and development of Indonesian corpora or corpora documenting languages in Indonesia are still an undergoing movement. The terms ‘corpus development’, ‘development’, ‘construction’, and ‘indonesian corpus’ have started to appear in 2017 and 2018 publications, and they still appear along with ‘learner corpus’ in 2019 publications. The fact that Korpus Indonesia, Kamus Alay, C-SMILE, and CINTA were started to become available around the same time making it possible that the publications are perhaps reports on the construction and development of either corpus. There is also a great possibility that these corpora—or other individually-made corpus—are used as data source, as evidenced by the terms ‘corpus analysis’, ‘indonesian student’, and ‘learner corpus’ overlapping in 2019 publications.

Yet, from the overlay visualization, corpus research in Indonesia from 2011 to 2022 is still scattered around, and it is difficult to determine the trend for each year, as corpus in Indonesia is still not considered a popular data source or corpus linguistics a sought-after procedure/method. Specific corpus terms and computational analysis tool, for example, inter alia ‘antconc’, ‘collocation’, and ‘dataset’ only appear in studies published around 2020 onwards, and rarely mentioned or overlapped with former publications. Regardless, it is quite refreshing to see the term ‘covid’ in corpus-related research published around 2020—perhaps with the limited conventional methods of data collection in the era of COVID-19 pandemic, corpus linguistics and corpora have become the feasible method and source to use.

4 Conclusion and Future Plans

This study presents mappings of how corpus-based research or corpus linguistics have been utilized in either linguistic or non-linguistic fields in Indonesian context to encourage Indonesian researchers to employ empirical applications of corpus. In the context of foreign language teaching and learning, for instance, research on learner corpus, pedagogic corpus, and the use of corpus for Indonesian EFL learners are still sporadic. Therefore, it is greatly recommended that corpora are used in data-driven learning (DDL), as it has been proven internationally to become a beneficial learning method for language learners, particularly second or foreign language learners as they are encouraged to independently look for real-life language use through corpora. It is worth mentioning that corpus-based DDL and BIPA research in Indonesia has been conducted, however, the publication frequency is still very low.

Documenting local languages in corpus has also not been collectively coordinated. Looking back to the list of available corpora in Indonesia, it is undeniable that researchers often use or build their own specialized corpus from either spoken or textual materials from particular local language for their own research purposes. They rarely put together a specific database for their corpus/corpora for other people to access, especially one that can be accessed offline [12]. Additionally, the language that most of these corpora document is none other than Indonesian, although a few does project its varieties; Kamus Alay on documenting colloquial Indonesian lexicon, OpenSubtitles Indonesia on conversational Indonesian, and C-SMILE on Indonesian learners’ English, for instances. The term ‘madurese’, for example, appears only 3 times with weak link strength in 2016 publications that its node almost disappears behind other terms in the overlay visualization, suggesting that there are only a few published studies on Madurese corpus. Through these findings, this study wants to inspire and encourage future Indonesian researchers to carry out more research in documenting local languages in corpora to not only maintain, but also revitalize and prevent Indonesian local languages from language death or loss, as well as empirical applications of corpus outside the field of linguistic discussion as it has been repeatedly proven to be potential in many data-driven research around the world.

References

[1] Lüdeling, A. and Kytö, M.: Corpus Linguistics: An International Handbook. Mouton de Gruyter, Berlin, Philadephia (2009)

[2] McEnery, T. and Hardie, A.: The History of Corpus Linguistics. Oxford University Press (2013)

[3] Svartvik, J.: Corpus linguistics 25+years on. Corpus Linguistics 25 Years on, Amsterdam (2007)

[4] Renouf, A.: Corpus development 25 years on: from super-corpus to cyber- 27 corpus. Corpus

Linguistics 25 Years on, Amsterdam (2007)

[5] Wilson, A., Archer, D. and Rayson, P., Eds.: Corpus linguistics around the world. Rodopi, Amsterdam (2006)

[6] Wilson, A., Archer, D. and Rayson, P.: Preface: Corpus linguistics around the world. Corpus linguistics around the world, Amsterdam (2006)

[7] Bennett, G.: An Introduction to Corpus Linguistics. Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers, Michigan (2010)

[8] Harzing, A.: Publish or Perish: Explains the use of Publish or Perish and its metrics. Harzing.com (2016)

[9] van Eck, N. and Waltman, L.: Visualizing Bibliometric Networks. Measuring Scholarly Impact: Methods and Practice, Switzerland. pp. 285–320 (2014).

[10] Larasati, S., Kuboň, V., and Zeman, D.: Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. pp. 119–129 (2011)

[11] van Eck, N. and Waltman, L.: VOSviewer Manual: Manual for VOSviewer version 1.6.18. (2022)

[12] Adriansyah, A.: Penyusunan Korpus Berita Terbuka Berbahasa Indonesia. Jurnal Teknologi Terpadu, Vol. 2, No. 2 (2015)