Utilization of corpus for sustainability of linguistic research

U Novella

Japanese Department Bina Nusantara University, Kampus Kijang, Jalan Kemanggisan Ilir III No.45, RT.12/RW.6, Palmerah, Jakarta Barat, 11480, Indonesia

Email: utari.novella@binus.edu

Abstract. This research aimed to introduce a corpus to Indonesian to improve the sustainability of research in the field of language. In English and Japanese study, language research had been highly developed because the existence of a linguistic corpus supports it. For example, in Japanese, the corpus was managed by an organization called NINJAL (National Institute for Japanese Language and Linguistics), where the corpus could be accessed free or paid. This research method was a literature review. It is found that for the sustainability of research in Indonesia, the existence of a corpus is very much needed. It is better if the corpus in Indonesia is created and managed by an organization and can be accessed free or paid for researchers who want to research Indonesian in more detail.

Keywords: corpus, Japanese corpus, Indonesia corpus, linguistic research

1. Introduction

Nowadays, there are many advanced technologies created by humans to facilitate human life. The higher the technology, humans will be able to live a high-quality life filled with comfort. Society 5.0 is a smart society where technologies such as big data, the internet, Artificial intelligence (AI), and robots have become part of people's lives [1]. Technologies can be seen in various aspects of life, as in the fields of science, industry, agriculture, information, etc.

Society 5.0 is a concept created by the Japanese, which is motivated by the many social challenges that need to be overcome by the Japanese community in the future. One of which is the declining birth rate (shoshika) and the increasing number of older people (koreika). It is this challenge that makes Japanese society active in creating innovations to facilitate their community in the future and to provide solutions to the problems they will face. The concept of society 5.0, many benefits are felt by the Japanese community, including in the field of public health. The public can already consult with physicians remotely using internet technology without having to go to the hospital. Transactions without using cash are also an advantage of fintech. Even the development of infrastructure using robots is clear evidence of society 5.0.

In the field of language research, technology can also be applied to access a language "bank" database electronically with a computer. The Language Bank has another term called the corpus. The corpus comes from the word corpora, namely language materials collected in a source [2]. Language material contained in the corpus can be used as data for research. Japanese is one of the languages with a complete and updated corpus, which makes research in Japanese very productive and sustainable. In the Japanese corpus, there are various genres of language use, such as spoken or written language,
language based on gender, and language in terms of history. The author raised the Japanese Language corpus because, according to the author's field of Japanese Linguistics, when researching Japanese, it is beneficial with the Japanese corpus. By using corpus searching data becomes more accessible, and in terms of time is also very efficient. Meanwhile, when making comparisons in Indonesian, because there is no Indonesian corpus, it takes longer because they have to collect data one by one from various sources. Based on these reasons, the author wants to raise the theme of how the existence of this corpus can make the sustainability of research in the field of language. It attracts the interest of language researchers to research Indonesian, or examine the comparison of Indonesian with other languages to be more productive.

2. What is corpus linguistic
In linguistics, it is inseparable from the elements that make up the language. To examine the details contained in a language, it takes a collection of texts that consists of words, sentences, spoken, or written words that represent the language. Verbal sources can include recordings, interviews, discussions, and written sources in books, novels, newspapers, magazines, web, blogs, journals, etc. This collection of materials is called the corpus, and then the corpora will be stored in an electronic database called the corpus. The text has been collected in the corpus needs to be affirmed permission to use. With the consent of the text owner, data in the corpus is in the right condition. Corpora will not be useful if it is not stored using computer technology. Currently, computer corpora may save many millions of running words, whose features can be analyzed using tagging (the addition of identifying and classifying tags to words and other formations), and the use of concordance programs [3].

Corpus linguistics is a study of language based on a vast collection of real-life languages stored in the corpus. Corpus linguistics is not a separate branch of linguistics, corpus linguistics as a methodology rather than a traditional branch of linguistics like semantics, grammar, phonetics, or sociolinguistics [4]. Although the methods used in corpus linguistics were first adopted in the early 1960s, the term corpus linguistics did not appear until the 1980s [5]. Corpus linguistics can be applied in almost any area of language studies. An object of a study is authentic, naturally occurring language use. Using computer technology can quickly show that quantitative data shows what happens frequently and rarely occurs in the language. In general, data that can be analyzed by this method are divided into two broad groups. They are the data obtained from the researcher and informant's intuition, and the data from the language used daily naturally that can be from the light conversation [6]. It can also analyze a collection of texts collected and managed systematically. The purpose of systematic is to prepare the structure and content of the text-based on principles outside the language, such as the context of time, place, and others.

In language research, the scope of the use of corpus linguistics is not only about the form of words but also includes their meaning. Scope of corpus linguistics is as said by [7] that corpus linguistics is concerned not just with describing patterns of form, but also with how kind and meaning are inseparable. Corpus-based studies can see a variety of things in that language. Some fields that rely on the corpus are lexicography, grammar, sociolinguistics, translation, teaching, stylistics (dialectology), dialectology, and historical linguistics. The corpus will make language research in various fields will develop and sustainable.

3. Corpus in Japanese language research
In this section, we will look at the role of the corpus in Japanese language research following the author’s field. Language research in Japan is very much developed compared to Indonesia. Not only foreigners who research Japanese, but Japanese people are also interested in studying their languages. This phenomenon is evidenced by the large number of reference books on Japanese linguistics written by Japanese ranging from general to very specific. For example, Japanese has a variety of types of particles, one of which is particle marking particles wa. This particle only consists of 2 letters, namely  と and て, but many books that specifically only discuss this particle. This example shows that although only a particle has many aspects of being studied in terms of linguistics. If it is searched one by one in books,
magazines, newspapers, or that are heard orally through everyday conversation; there would be many situations that use ‘wa’ particles. However, if searched one by one, it will take a long time to classify it. The existence of the corpus will be beneficial for detecting and classifying these particles. For example, using wa particles verbally or in writing, the difference in the use of wa particles in terms of gender, even the use of wa particles, can also be seen at specific periods (in terms of linguistic history).

Japan has a large corpus of languages managed by the National Institute for Japanese Language and Linguistics (NINJAL), which was first formed in 1948. NINJAL has an online version of the corpus that is free to access without the need for registration is known as Shonagon; some need registration and paid are known as Chunagon. There is also an offline version of the corpus for a fee. Display the English version, as shown below.

Figure 1. NINJAL English Version
(https://pj.ninjal.ac.jp/corpus_center/bccwj/en/)

The classification of data available in this corpus varies greatly, including modern Japanese corpus, Corpus Annotation, Showa Speech Corpus, and others. Corpus shown in the image below:
Figure 2. The Various of Corpus (https://pj.ninjal.ac.jp/corpus_center/en/)

For example, research on the use of *wa* particles in Japanese sentences can be done by the corpus linguistic method by searching and classifying *wa* particles using a corpus like an image below.

Figure 3. The Search Column (https://shonagon.ninjal.ac.jp/search_form)
In the search column, it is filled with *wa* (は) particles and selects the year according to the *wa* particle usage in a particular year needed for research.

![Search Results](https://shonagon.ninjal.ac.jp/search_result)

**Figure 4.** The Searching Results of ‘*wa*’ Particle (https://shonagon.ninjal.ac.jp/search_result)

After searching, the number of *wa* particles that appear, the sentence that accompanies the particle, the media, title, author, year, and publisher will be written. Corpus is beneficial for Japanese language researchers to find research data. So is the case with foreign researchers interested in Japanese, with this corpus very helpful in research. Corpus gives a positive value to the sustainability of Japanese language research. Language research will not stop, because language is growing, language data is easily obtained through corpus and is efficient in terms of time.

4. **Corpus in Indonesian**

In Indonesia, the term corpus is well known to language learners or language researchers. However, the absence of an official Indonesian corpus makes it difficult for language researchers to obtain Indonesian data [8]. Another thing caused by the lack of language data sources in Indonesia is the low publication of Indonesian research papers at the international level. If Indonesian had a corpus like Japanese, there would be many aspects that could be researched from the variety of spoken and written Indonesian, Indonesian based on historical linguistics, and so on. Even researchers will not be overwhelmed when they have to collect data one by one to conduct a language study. The corpus of Indonesian will also enrich the contrastive research between languages in the world and Indonesian. It is because Indonesian as a mother tongue will be easier to explore and then compared to other words to see the linguistic phenomena similarities or differences in language. To be more extensive, with corpus technology can create a database for local languages in Indonesia and make language research even more developed.

Japanese Literature students at Bina Nusantara University are very minimal using corpus linguistics, especially Japanese corpus. It is because they do not understand what is meant by corpus, or some know what corpus is. Still, they do not know how to access and use it. So when they write a thesis with the field of linguistics, they usually will look for words raised as the focus of research on comics, novels, magazines, and so on manually by searching for these words one by one in the reading material.
they make data. This is very inefficient in terms of time, and the data obtained is also general. For example, it does not distinguish between spoken and written languages.

Even though Indonesian does not have an official corpus such as Japanese or other languages, it still needs to be given basic knowledge of the corpus and how to use it. The specific purpose of Indonesian students learning a foreign language can become familiar using the corpus in that language. In terms of the sustainability of research in the Indonesian language, it would be perfect if the Indonesian Language has a corpus to make it easier for researchers to conduct more varied research.

5. Conclusion
Corpus linguistics is needed for the sustainability of language research. Without the corpus will make the language research is not developed. Corpus can be used as a method for language research such as sociolinguistics, translation, and so forth. Therefore, Indonesians who study foreign languages need to be equipped with knowledge about the corpus so that when they research a foreign language, they can use it to support their research. For the sustainability of research in Indonesia, the existence of a corpus is very much needed. It is better if the corpus in Indonesia is created and managed by an organization and can be accessed free or paid for researchers who want to research Indonesian in more detail.

References
[1] Okamoto M 2017 Standardization Activities on Society 5.0 in Japan Retrieved from https://bsn.go.id/uploads/download/0_standardization_activities_on_society_5.0_in_japan_okamot
[2] Kridalaksana H 2008 Kamus Linguistik (Jakarta: Gramedia Pustaka Utama)
[3] McArthur T (ed.) 1992 The Oxford Companion to the English (Oxford & New York: Oxford University Press)
[4] Baker P 2010 Corpus Methods in Linguistics (New York: Continuum International Publishing Group)
[5] Nordquist R 2020 Definition and examples of Corpus Linguistics Retrieved from thoughtco.com/what-is-
[6] Nesselhauf N 2011 Corpus Linguistics: A Practical Introduction Retrieved from http://www.as.uniheidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduct
[7] Cheng W 2012 Exploring Corpus Linguistics: Language in Action (London: Routledge)
[8] Waskita D 2017 Corpus Linguistics: Method, Theory, and Practice Jurnal Sosioteknologi 16 145-147