

Building Corpora for Under-Resourced Languages in Indonesia

Totok Suhardijanto, Arawinda Dinakaramani

Department of Linguistics Universitas Indonesia, Computer Science Universitas Indonesia

Depok Indonesia, Depok Indonesia

{totok.suhardijanto, arawinda.dinakaramani}@ui.ac.id

Abstract

Indonesia has the second highest language diversity in the world, just under Papua New Guinea (Simons & Fennig 2019). There are 719 recorded regional languages spoken in Indonesia, 13 of which have become extinct (Lauder 2017). Presently, there are three categories of linguistic condition in Indonesia which consist of the national language, regional language, and foreign language. In accordance with the politics of language, the focus of language development in Indonesia lies in the national language, namely Indonesian or Bahasa Indonesia. Bahasa Indonesia is often cited as one of the great success stories of language policy and planning. However, the very success of Indonesian language threatens the other 699 languages in the archipelago (Cohn & Ravindranath 2014). As a result of the intense politics of national language, language resources focus on Indonesian language—even though its quantity and quality are still far behind other prominent world languages. Thus, the development of regional language resources in Indonesia has yet to become a government priority. Meanwhile, the number of regional languages that fall into the endangered category rises with each passing year.

Keywords: under-resourced languages in Indonesia, multilingual corpora, corpus management system

Résumé

Indonesia merupakan negara kedua yang memiliki keberagaman bahasa tertinggi di dunia setelah Papua Nugini (Simons & Fennig 2019). Tercatat ada 719 bahasa daerah yang dituturkan di Indonesia dan 13 di antaranya telah punah (Lauder 2017). Pada saat ini, di Indonesia, terdapat kondisi kebahasaan dengan tiga kategori bahasa yang hidup di dalam masyarakat, yakni bahasa nasional, bahasa daerah, dan bahasa asing. Sesuai dengan politik kebahasaan, fokus pengembangan bahasa di Indonesia terletak pada bahasa nasional, yaitu bahasa Indonesia. Bahasa Indonesia sering dirujuk sebagai salah satu dari kisah sukses kebijakan pembinaan dan perencanaan bahasa. Namun, kesuksesan bahasa Indonesia tersebut memberikan tekanan terhadap 699 bahasa lain di negara kepulauan tersebut (Cohn & Ravindranath). Akibatnya gencarnya politik bahasa nasional, pengembangan sumber daya bahasa (language resources) pun terfokus pada bahasa Indonesia—meskipun jumlah dan kualitasnya pun masih sangat jauh dari bahasa-bahasa utama di dunia. Dengan demikian, pengembangan sumber daya bahasa daerah di Indonesia belum menjadi prioritas pemerintah, padahal jumlah bahasa daerah yang masuk ke dalam kategori terancam punah terus bertambah tiap tahun.

1. Background

Although Indonesia is the second country with a variety of languages in the world, the documentation effort and development initiatives of language resources from the existing languages in the country are still far from the optimal condition. Even if there are any, according to Suhardijanto & Dinakaramani (2018), all are related to the following two conditions. First, most language resources and documentation were in fact initiated by foreign institutions or institutes. Second, if conducted by Indonesians, they are usually sporadic, individual, and limited. They are limited to be used to support their own research.

Out of 719 languages in Indonesia, 706 of them are still used, while 13 languages belong to the endangered category (Lauder 2016). Regional languages in Indonesia vary in the types of languages and the number of the users. From the types, Indonesian languages can be grouped into two big categories: Austronesian languages and Non-Austronesian languages. The Austronesian languages are spread in the western and eastern parts of Indonesia, while the non-Austronesian ones are spread only in the areas of Papua, Maluku, and Nusa Tenggara, all of which are located in the eastern part of Indonesia. From the number 259

of the users, 386 languages are spoken by more or less 5,000 users; 233 are owned by more or less 1,000 users; 169 are owned by more or less 500 users; and 52 languages belong to more or less 100 users (Gordon 2005). Meanwhile, according to Simons & Fennig (2018), there are only 20 languages spoken by more than one million people, including Javanese with the number of users around 84.3 million.

In the case of Indonesia, besides the high number of regional languages, there are some language problems that complicate the situation. In Indonesia, there are three language categories, namely national language, regional language, and foreign language. According to Riza (2008), the development of language resources for languages in Indonesia generally focus on the national language, which is Indonesian language only. This happens due to the lack of attention from the government towards the problems of regional languages in this country (Lauder 2016). As a result, the regional language development and documentation funding is very limited. Several efforts to establish language resources have been done sporadically and without coordination by the researchers having concern for the fate of the regional languages in Indonesia (see Suhardijanto 2017, Suhardijanto & Arawinda 2018).

Not only the government but also the legislative party seems not to prioritize the language problems in Indonesia yet. This is proven from no law draft on regional languages in Indonesia which has been integrated into the national legislation programs (*prolegnas – program legislasi nasional*) that becomes the duty of the House of Representatives, although the draft has been completed since 2016. According to Lauder (2016), the enactment of the law on the regional languages, in fact, is expected to be able to strengthen the position of the regional languages in Indonesia. It seems that the government and parliament of Indonesia still consider economic and political fields as the main priority of development in Indonesia.

This paper informs the efforts we have made to collect, compile, and build regional language resources in Indonesia with the funding obtained from various sources. In this study, the discussion focus is limited on the development of the corpus management system that becomes one of the phases in the regional language resources development in Indonesia. The corpus management system we are developing has several functions as follows:

- 1) save the corpus text data of regional languages in a digital form;
- 2) process and store corpus metadata so that it can be accessed by other software;
- 3) analyze corpus text data by corpus methods, such as keyword lists, concordances, n-grams, etc.

2. Corpus Manager

In the effort to build language resources, there are many activities that we have done. Those activities start from language documentation in the field to annotated corpus compilation that can be used to develop the next application of NLP (Natural Language Processing). We started from database development for making dictionaries and grammar books to the making of software or tools to manage language data.

As previously mentioned, this paper will discuss the development of the corpus management system for the existing languages in Indonesia. Since the database or corpora of regional languages are kept in the server of Universitas Indonesia, the corpus system developed is named Korpus Universitas Indonesia (Corpus of Universitas Indonesia).

Corpus management system (CMS) is generally a search engine system developed in a complex manner so that it can carry out the search towards the form of a language or a set of sentences. In a narrow understanding, CMS refers to the server or corpus query engine, while the client side is usually called as user interface (Kouklakis 2007). In this paper, CMS is understood as the combination between those two sides. Therefore, with that understanding, CMS hereinafter will be referred to as a corpus manager.

A corpus manager can be a stand-alone software installed on the user's computer or an online corpus tool that allows users to access corpus, or corpora, from any computer. A corpus manager is designed to have some features. The

basic feature is concordance. A user can use a corpus tool to search for a keyword and then the search results will be shown as the line of context for each occurrence of the keyword. Other features of a corpus tool include the ability to extract wordlist, lexical bundles or n-gram, keywords, particular structures, and also metadata information from the corpus.

Some corpus managers are designed for a particular corpus, while other corpus managers are designed to enable users to upload and analyze any corpus. Most corpus managers are used to access a prepared corpus, while some corpus managers are used to access a web as a corpus. Prepared corpus is a corpus that has been compiled with linguistic research in mind and specifically designed for linguists' purposes (Kilgariff & Kosem 2012). The web can be viewed as a corpus with vast quantities of texts for many languages that covers a wide range of text types and domains (Kilgariff, Baisa, Buta, Jakubichek 2014).

The users of corpus managers can be categorized into several types, such as lexicographers, linguistics researchers and students, and language teachers and learners. For this reason, a corpus manager should be designed to meet the needs of their target users. In the case of corpus manager development, one of the designs is the provision of as many as functionalities as possible to fulfill the users' needs. In relation to language resources, the users' needs of this corpus manager are not limited to the researchers in linguistics or other social sciences but also to the researchers and development in the field of natural language processing or artificial intelligence.

3. Language Corpora

Currently, the regional language data that become the focus in the language resources development are limited to the regional language data with the number of users above five million people. The languages entered into Korpus Universitas Indonesia cover Indonesian language, Javanese, Sundanese, Minangkabau language, Banjar language, and Bataknese.

The design of corpus data for our web-based application is decided by considering:

- 1) selection criteria: if applicable, we design a corpus that represents various texts or genres;
- 2) corpus size: the size is still growing;
- 3) data authenticity: from real data, no artificial data
- 4) storage media: each corpus data is digitalized, especially in the form of text file;
- 5) data manipulation: we build a web-based application to access and manage corpus data.

Among the six languages, the Javanese corpus possesses the most diverse text collection. Broadly speaking, it is divided into two categories, namely the spoken and written corpus, while the details can be seen in the following figure

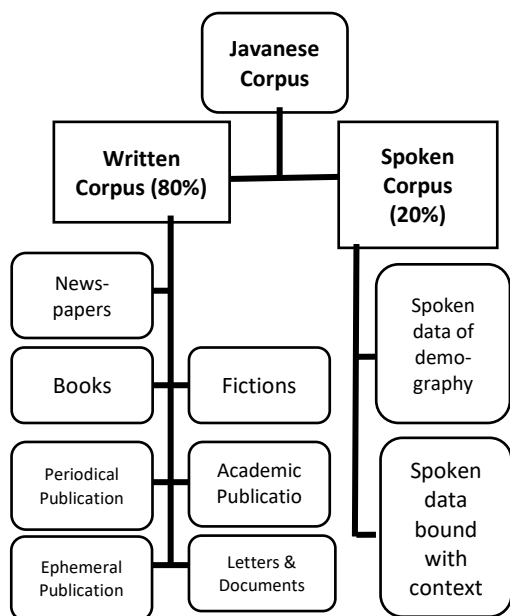


Figure 1: The Structure of Javanese Corpus

In the phase of data collection, written texts are generally obtained through data exploration in the field. Data in the form of hardcopy are digitalized via the scanning process and kept in the format of a text file (.txt) with the text encoding UTF-8. Some texts were obtained through text scrapping in the web, such as the Wikipedia text in Javanese.

Meanwhile, for spoken data, they are generally obtained via live recording. Therefore, the process of transcription, editing, and conversion into the format of text file are required as written text data should be. In terms of content, spoken data consist of spoken language data uttered based on demographic variables, such as age, sex, and others. Moreover, other data are spoken language data compiled based on the variety of context, such as lecture, preach, conversation, and others.

Still on the Javanese language corpus, the text data were collected from the period of time between 1940 and 2018. The biggest data portion is the fiction texts. It happened because the most publication encountered in the market in that period of time was fiction texts. In terms of Javanese language data, the least number of texts is academic texts due to the policy of using the national language, which is Indonesian language, as the language of instruction in education.

4. Design and Architecture

The corpus manager is built in PHP and designed based on the model-view-controller (MVC) architectural pattern. Since the users are mostly Indonesians, Indonesian is set as the default language for the application interface page.

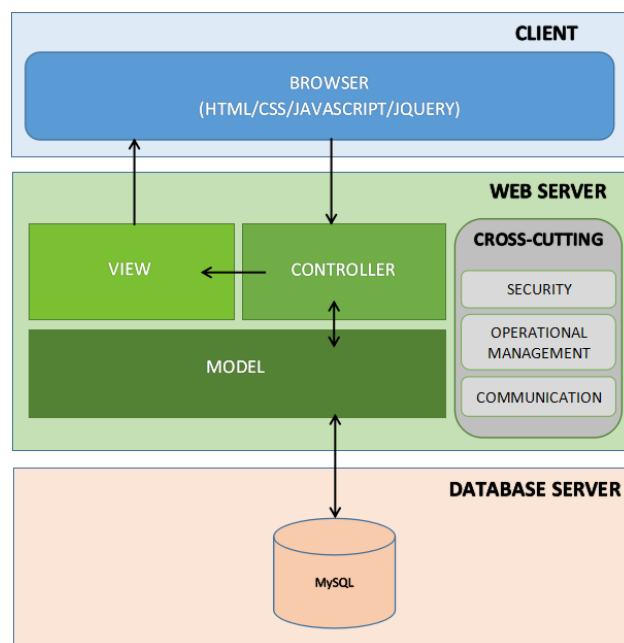


Figure 2: The System Architecture

Because the purpose is to support the regional language documentation, this corpus manager is also designed to be able to facilitate collaborative work in building and managing language resources. In addition, as the analysis tool, the corpus manager must be able to fulfill the users' needs which vary to search and explore language resources.

Since the functions of this corpus manager vary, the user classification consists of eight types, namely admin, chief editor, editor, data contributor, pending member, pending user, uncategorized, and annotator. The categories of data contributor and annotator are accommodated in this system because the work to build language resources really needs the role of those two user categories.

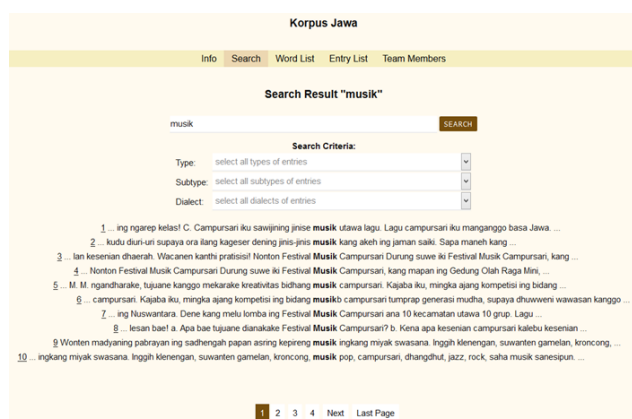
5. Features and Functions

The features in the corpus application have different accessibility permissions. These features are categorized into twofold: accessible to all users and accessible to specific users. Some features that are accessible to all types of users are select corpus, view word list, view concordance search result, and view text list.

Some features that are accessible to a specific type of users are as follows. First, the add editor feature can only be accessed by a chief editor. The submit entry feature can only be accessed by corpus team members. Then, the view, edit, validate, and reject entry features can only be accessed by the editor team. The function of viewing and generating multiple word expression or n-gram can only be accessed by registered users. The Add Annotator feature can only be accessed by a chief editor and an editor member. The function to view, edit, validate, and reject annotation results can only be accessed by chief editor and editor member. Finally, the features of processing and managing text-

annotation can only be accessed by the annotator team, chief editor, and editor member.

Figure 3: The Screenshot of Javanese Concordance



Feature

From all functionalities expected to exist in this corpus manager, there are only the functions of concordance, generating word-list, and managing corpus covering uploading, editing, and text sending validation from the contributor.

6. Conclusion

This corpus manager will continue being developed in terms of its design, functionalities, and the number of language data managed. There are several functionalities that are not available yet in this system, such as generating n-gram, collocation, searching based on structure, and others. Some of the corpus managers have been available as a stand-alone software, but they have not been integrated into the system. The number of languages integrated into the system will be attempted to keep on increasing.

7. Acknowledgements

This research was supported by The Directorate of Higher Education, The Ministry of Education Republic of Indonesia with Research Grant No. 1/E1/KP.PTNBH/2019 and No: 516/UN2.R3.1/HKP.05.00/2018.

8. Bibliographical References

- Cohn, A. and Ravindranath, M. (2014). Local Languages in Indonesia: Language Maintenance or Language Shift? *Linguistik Indonesia* 32.2: 131-148.
- Dinakaramani, A. and Suhardijanto, T. (2019). Building a web-based application for language resources in Indonesia. *Journal of Physics: Conference Series* 1192 (2): 12-22.
- Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL Internasional.
- Kilgarriff, A. and Kosem, I. (2012). *Electronic Lexicography*. Ed. S. Granger and M. Paquot. Oxford: Oxford University Press, pp 83-106.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014).

The Sketch Engine: ten years on. *Lexicography*, 1: 7-36, 2014.

Lauder, M.R.M.T. 2016. Preventing the Extinction of the Regional Languages through Policy Formation. The paper presented at the 9th National Seminar of Mother Tongue IX, "Prevention Strategy for Indigenous Language Extinction", Denpasar, Bali. 26-27 February 2016.

Riza, H. (2008). Resources Report on Languages of Indonesia. The 6th Workshop on Asian Language Resources.

Simons, G.F. and Fennig, C.D. (eds.). (2019). *Ethnologue: Languages of the World*. 21 st ed. Dallas: SIL International.

Suhardijanto, T. (2016). Developing language resources for under-resourced languages in Indonesia. The paper presented in the International Conference on Knowledge Creation and Intelligence Computing 2016, State Polytechnics Institute of Manado, Manado, Indonesia, 15—17 November 2016.

Suhardijanto, T. and Dinakaramani, A. (2018). Developing Language Resources for Indigenous Languages in Indonesia: Annotated Javanese Corpus Building. *Proceeding of Asia Pacific Corpus Linguistics Conference* 2018.

Suhardijanto, T. and Dinakaramani, A. (2019). Korpus Beranotasi: Ke Arah Pengembangan Korpus Bahasa-Bahasa di Indonesia. *Prosiding Kongres Bahasa Indonesia*. *Prosiding Kongres Bahasa Indonesia*, pp. 339-355.

Kouklakis, G., Mikros, G., Markopoulos, G., and Koutsis, I. (2007). *Corpus Manager A Tool for Multilingual Corpus Analysis*. *Proceedings from Corpus Linguistics Conference*. University of Athens: 1–12.

9. Language Resource References

- Suhardijanto, T., Puspitorini, D., and Dinakaramani, A. (2019). Javanese Language Corpus. URL: < https://korpus.ui.ac.id/c/korpus_jawa>.
- Suhardijanto, T., Puspitorini, D., and Dinakaramani, A. (2019). Minangkabau Language Corpus. URL: < https://korpus.ui.ac.id/c/korpus_minang>.
- Suhardijanto, T., Puspitorini, D., and Dinakaramani, A. (2019). Sundanese Language Corpus. URL: < https://korpus.ui.ac.id/c/korpus_sunda>.