

# Dictionary 4.0: Alternative Presentations for Indonesian Multilingual Dictionaries

**Arbi Haza Nasution, Totok Suhardijanto**

Informatics Engineering Department Universitas Islam Riau, Linguistics Department Universitas Indonesia  
Pekanbaru Riau Indonesia, Jakarta Indonesia  
arbi@eng.uir.ac.id, totok.suhardijanto@ui.ac.id

## Abstract

Building a multilingual dictionary for 719 languages in Indonesia is a challenging task. We have developed application to create the Leipzig-Jakarta list database for all indigenous languages in Indonesia. The database can be used to generate lexical similarity or lexical distance matrix between languages by comparing the word list. For starter, we covered 11 languages: Indonesian, Javanese, Sundanese, Madurese, Bima, Ternate, Tidore, Palembang Malay, Mandailing Batak, Malay, and Minangkabau. The application has two main features: exploring the existing translations and adding translations to a new language or editing existing translations through crowdsourcing. User acceptance test showed 3.48/4 score.

**Keywords:** multilingualism, multilingual dictionary, lexical network, lexical computation, computational linguistics

## Abstrak

Membangun kamus multibahasa untuk 719 bahasa di Indonesia adalah tugas yang berat. Kami telah mengembangkan aplikasi untuk membuat pangkalan data daftar Leipzig-Jakarta untuk semua bahasa daerah di Indonesia. Pangkalan data tersebut dapat digunakan untuk menghasilkan kesamaan leksikal atau matriks jarak leksikal antar bahasa dengan membandingkan daftar kata tersebut. Sebagai permulaan, aplikasi ini mencakup 11 bahasa: Indonesia, Jawa, Sunda, Madura, Bima, Ternate, Tidore, Melayu Palembang, Batak Mandailing, Melayu, dan Minangkabau. Aplikasi ini memiliki dua fitur utama: menjelajahi terjemahan yang ada dan menambahkan terjemahan ke bahasa baru atau mengedit terjemahan yang ada melalui mekanisme urun daya. Uji keberterimaan pengguna menunjukkan skor 3,48 / 4.

## 1. Introduction

According to (Eberhard et al., 2019), there are 719 languages in Indonesia, where 707 languages are still alive and 12 languages have become extinct. Extinct in this sense is that there are no longer any of the speakers. Among the surviving languages, 701 languages are local languages and 6 languages are not local languages. Furthermore, there are 18 languages that are used as administrative and / or educational languages, 73 languages are still growing, 188 languages are classified as strong, 347 languages are in difficulty, and 81 languages are in a danger of extinction. Furthermore, based on his observations, (Anderbeck, 2015) groups Indonesian languages into three groups. First, about two of the four languages in Indonesia today still have a vital life force and have a safe number of speakers (EGIDS (Expanded Graded Intergenerational Disruption Scale) 1-6a). In this group, intergenerational transmission of speakers still occurs and persists. Even though some of them are bilingual, they know when to use local and Indonesian languages. Second, one of the four languages in Indonesia is in fragile condition (EGIDS 6b Threatened) with speakers who continue to decline in number. Usually most young people still learn their mother tongue, but certain reasons make them change their orientation towards languages that are more economically advantageous. Third, the rest, one of the four languages in Indonesia seems to be dying (EGIDS 7-8b) or may have become completely extinct (EGIDS 9 and 10). Some, like the Marori language, may be lost in a generation. The other may be in two or three generations. With conditions like that, of course, we are like racing with time to document language.

Although some experts distinguish the terms of language documentation and language description (Austin and Sal-labank, 2011), in some ways, the two are interconnected. According to Austin, the documentation and description of languages differ in their purpose, points of interest, research methods, workflow, and outcomes. Descriptions or language descriptions basically aim at producing grammar, dictionaries, and collections of texts, the target users are generally linguists, and the material produced is sometimes written in a framework that is accessible to trained linguists. In contrast, language documentation is discourse-centered, the main objective being the direct representation of as many types of discourse as possible (Austin, 2007; Woodbury, 2003; Himmelmann, 1998). However, according to (Austin and Grenoble, 2007) the documentation project must rely on the application of theoretical and descriptive linguistic techniques so that the resulting output is sure to be utilized and understood by many communities. So, in other words, documentation and description are activities with objectives and outcomes that complement each other, and one of their important outcomes is the result of lexicographic work, the dictionary.

In the context of endangered languages, dictionaries have a very crucial role, namely storing what is left of endangered languages and cultures by recording valuable information that might be lost (Cristinoi and Nemo, 2013). The bilingual dictionaries are also useful for natural language processing researchers, especially for those related with enrichment of language resources like bilingual dictionary (Nasution et al., 2016; Nasution et al., 2017b; Nasution et al., 2017a; Nasution et al., 2018) or machine translation

(Nasution et al., 2017c; Nasution, 2018). Furthermore, in many cases, the existence of a dictionary can help revive a language and change the attitudes of speakers of that language which ultimately encourage them to use it as often as possible. Even so, (Cristinoi and Nemo, 2013) mentioned that there are some problems related to lexicography in the realm of language documentation. First, the compilers of endangered language dictionaries are generally people or linguists who care. Certainly, the result is different from the general dictionary compiled by a professional team. Secondly, dictionaries made for endangered languages are certainly far from direct economic profit. Third, the endangered language dictionaries have limited distribution, that is only to linguists or the public who have an interest in the language concerned. Fourth, in the work of lexicography in endangered languages there are several problems that must be resolved, for example what characters are used, which variations are considered standard, and so on. Fifth, data collection of endangered languages is more difficult because it only relies on the ethnographic work of researchers or notes from concerned community members. Sixth, the dictionary of threatened languages is usually used for research purposes, documenting specific languages and cultures, protecting language and cultural heritage that will be lost without written traditions on the language or culture, helping indigenous people communicate in dominant foreign languages, helping non-native speakers to understand the native speakers and their cultural background, and provide orthography or standard written form for the entire vocabulary.

Because of the problems mentioned above, the data collection of endangered language dictionaries is generally done with a limited number of vocabularies, generally focus on general vocabulary or even basic vocabulary lists. The list is a lexical artifact which is a vocabulary whose references are universally available in many languages in the same region. In the condition of Indonesia which is multilingual, of course the problem becomes more complex. Over time, how do lexicographic studies contribute to language documentation efforts, especially in terms of recording important and varied information about language and culture in Indonesia? Making multilingual dictionaries is not an easy task, especially from the point of computational lexicography (Walker, 1995). Thus, in this paper, we try to build a model that can accommodate the diversity of languages in Indonesia. This can be further elaborated with the question: how to compile dictionaries for languages in Indonesia? What is the correct format of multilingual dictionaries that can help document languages in Indonesia? These two questions will be answered in this paper.

## 2. Methodology

In the 1950s, linguist Morris Swadesh published a list of 200 words called the Swadesh list, which were thought to be 200 lexical concepts found in all languages that were most unlikely to be borrowed from other languages (Swadesh, 1955). Swadesh then reduced the list to 100 items based on intuition where a drastic removal from a 200-word list was the best solution, with the consideration that quality is at least as important as quantity. Al-

though the new list has weaknesses, but the list is relatively light to process because of the small amount. Automated Similarity Judgment Program (ASJP) (Brown et al., 2008) is an open source software with the main objective to develop a Swadesh list database for all languages in the world where lexical similarity or lexical distance matrix between languages can be obtained by comparing the word list. However, the list of 100 Swadesh words was cut down to 40 words that are considered the most stable of forms of change, maintained over time and not replaced by other lexical items from the language itself or elements borrowed from other languages (Holman et al., 2008). The lexical distance between regional languages in Indonesia has been visualized using the ASJP database (Nasution and Murakami, 2019; Nasution et al., 2019). However, there are doubts about the validity of the lexical distance between some regional languages such as between Sundanese and Javanese which should be closer to the lexical distance but only 21.8% of lexical similarities are produced. Therefore, alternative word lists are needed that can produce more accurate lexical distances.

In addition to the Swadesh list, linguists also use the Leipzig-Jakarta list (100 words) (Tadmor et al., 2010) to test the level of chronological separation of languages by comparing words that are resistant to loans. The Leipzig-Jakarta list is available in 2009 (Sakel and Everett, 2012). The mobile application developed in this paper aims to develop the Leipzig-Jakarta list database for all regional languages in Indonesia where lexical similarity or lexical distance matrix between languages can also be further obtained by comparing the word list. The application built will be tested for user satisfaction with quantitative analysis using a questionnaire. The proposed framework is depicted in Figure 1. The data will be used to generate visualization of Indonesian Indigenous Languages Lexical Similarity with Knowledge Graph.

## 3. Results

For the initial research, 100 Leipzig-Jakarta word lists were translated into 11 languages: Indonesian, Javanese, Sundanese, Madurese, Bima, Ternate, Tidore, Palembang Malay, Mandailing Batak, Malay, and Minangkabau. The application has two main features: exploring the translations of the 100 Leipzig-Jakarta word list and adding translations to new languages or changing translations that are already available. The exploration interface for translating 100 Leipzig word lists into 11 languages with the details of the translated words including the definition, synonyms and example use of the word in a sentence are shown in Figure 2.

To add a translation to a new language or change an already available translation, the user should register to the system first using the registration form. After entering the user's email address for verification, the user can click on the language selection dropdown, then the user can choose the destination language according to the language selection feature, the last step, the user can type the translation according to the language of choice, and click the "SUNTING / TAMBAH KATA" (which translated to EDIT / ADD WORDS) button, then the translation added / edited will be

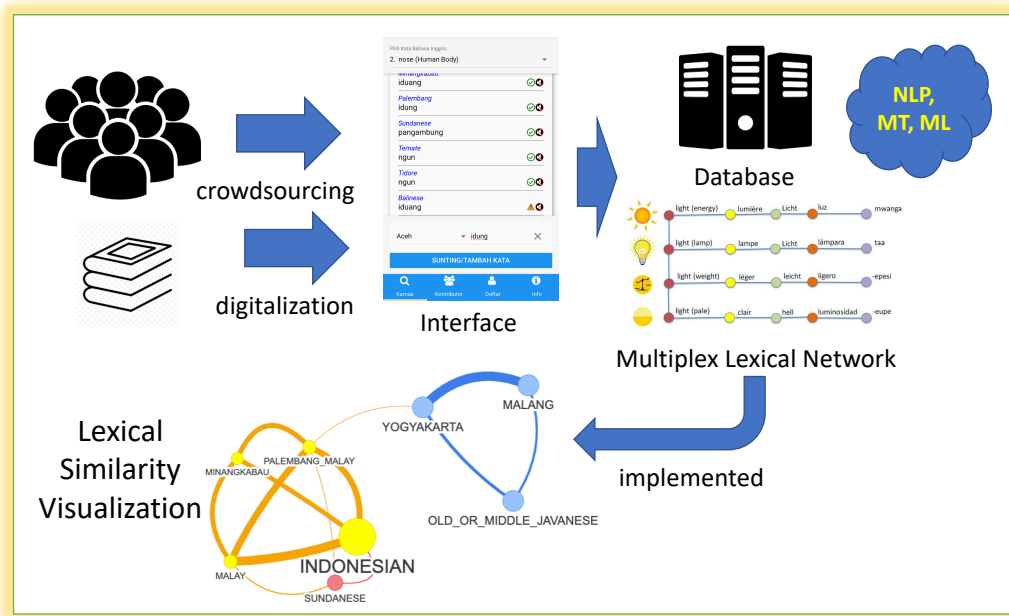


Figure 1: Proposed framework.

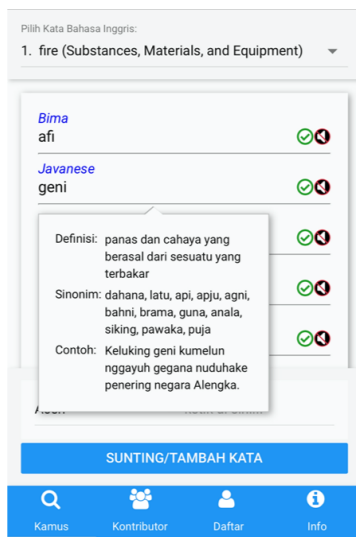


Figure 2: The definition, synonyms and example in sentence.



Figure 3: Leader board of contributor.

#### 4. Conclusion

verified by the linguist. Finally, the user will get a poin for each translation added or edited, and another poin when the new translation or edition has been verified. The leader-board is shown in Figure 3.

The application that was built was tested by 36 random users with quantitative analysis using a questionnaire with 7 questions as shown in Table 1. Based on the results of the user satisfaction questionnaire with dictionary 4.0, the average value for the whole questionnaire item was 3.48. This shows that the design and appearance of the Dictionary 4.0 Application is quite interesting, easy to use and accepted by users.

Until now, in this study, a multilingual dictionary prototype model with the functionality to collect data of various languages was quickly compiled. Therefore, the focus in this paper is on the issue of setting up a language data collection system through a crowd sourcing mechanism. Meanwhile, in terms of usage, acceptance testing has been carried out to see how well the application design according to the user. Based on these tests, we obtained quite interesting results, which is 3.48 from a scale of 4. The next stage of this research is to upgrade the dictionary 4.0 application that is capable of managing multilingual dictionary services with dedicated functions for general users and registered users. In addition, the language similarity comparison function

Item	Mean	Median	Standard Deviation
Appealing design and appearance	3.47	3	1.078
The design and appearance of the application is easy to understand	3.41	3	1.043
The navigation menu is easy to understand	3.37	3.5	1.157
The colors used in the application are suitable and not excessive	3.72	4	0.958
The application is easy to use	3.47	4	1.078
Easy to explore each word translation	3.47	4	1.047
It is easy to propose revision to existing translations or add new translations	3.47	3.5	1.047

Table 1: Results of user satisfaction questionnaire of dictionary 4.0

will be included using the lexical distance approach as in the ASJP database program.

## 5. Acknowledgements

This research was partially supported by Universitas Islam Riau.

## 6. Bibliographical References

- Anderbeck, K. (2015). Portraits of language vitality in the languages of indonesia. *Language documentation and cultural practices in the Austronesian world: Papers from*, pages 19–47.
- Austin, P. K. and Grenoble, L. (2007). Current trends in language documentation. *Language documentation and description*, 4:12–25.
- Austin, P. K. and Sallabank, J. (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.
- Austin, P. K. (2007). Training for language documentation: Experiences at the school of oriental and african studies. *Documenting and revitalizing Austronesian languages*, pages 25–41.
- Brown, C. H., Holman, E. W., Wichmann, S., and Velupillai, V. (2008). Automated classification of the world's languages: a description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(4):285–308.
- Cristinoi, A. and Nemo, F. (2013). Challenges in endangered language lexicography.
- Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). *Ethnologue: Languages of the world*.
- Himmelman, N. P. (1998). Documentary and descriptive linguistics.
- Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.
- Nasution, A. H. and Murakami, Y. (2019). Visualizing language lexical similarity clusters: A case study of indonesian ethnic languages. *Journal of Data Science and Its Applications*, 2(2):45–59.
- Nasution, A. H., Murakami, Y., and Ishida, T. (2016). Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3291–3298, Paris, France, May.
- Nasution, A. H., Murakami, Y., and Ishida, T. (2017a). A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2):9:1–9:29, November.
- Nasution, A. H., Murakami, Y., and Ishida, T. (2017b). Plan optimization for creating bilingual dictionaries of low-resource languages. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 35–41, Sept.
- Nasution, A. H., Syafitri, N., Setiawan, P. R., and Suryani, D. (2017c). Pivot-based hybrid machine translation to support multilingual communication. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 147–148, Sept.
- Nasution, A. H., Murakami, Y., and Ishida, T. (2018). Designing a collaborative process to create bilingual dictionaries of indonesian ethnic languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3397–3404, Paris, France, may. European Language Resources Association (ELRA).
- Nasution, A. H., Murakami, Y., and Ishida, T. (2019). Generating similarity cluster of indonesian languages with semi-supervised clustering. *International Journal of Electrical and Computer Engineering (IJECE)*, 9(1):1–8.
- Nasution, A. H. (2018). Pivot-based hybrid machine translation to support multilingual communication for closely related languages. *World Transactions on Engineering and Technology Education*, 16(2):12–17.
- Sakel, J. and Everett, D. L. (2012). *Linguistic fieldwork: A student guide*. Cambridge University Press.
- Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.
- Tadmor, U., Haspelmath, M., and Taylor, B. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, 27(2):226–246.
- Walker, Z. C. (1995). *Automating the lexicon: research and practice in a multilingual environment*. Oxford University Press.
- Woodbury, A. C. (2003). Defining documentary linguistics. *Language documentation and description*, 1(1):35–51.