International Conference for Language Technologies For All: Enabling Linguistic Diversity And Multilingualism Worldwide
Paris, 4-6 December 2019

BUILDING CORPORA FOR UNDER-RESOURCED LANGUAGES IN INDONESIA

Totok Suhardijanto (Universitas Indonesia)
Arawinda Dinakaramani (Universitas Indonesia)

There are 719 local languages spoken in Indonesia, 13 of which has become extinct (Ethnologue 2019). Presently, there are three categories of linguistic condition in Indonesia which consists of the national language, regional language, and foreign language. The national language, namely Indonesian or Bahasa Indonesia, is often cited as one of the great success stories of language policy and planning. But the very success of Bahasa Indonesia threatens the other 699 languages in the island nation (Cohn 2014). As a result of the intense politics of national language, language resources are focused on Bahasa Indonesia. 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 rise with each passing year.

INDONESIAN CORPUS PROJECT

This project is an effort to compile and develop local language resources in Indonesia with funding support collected through research funding from various sources. At this time, the corpus management application system that has been developed has the following functions:

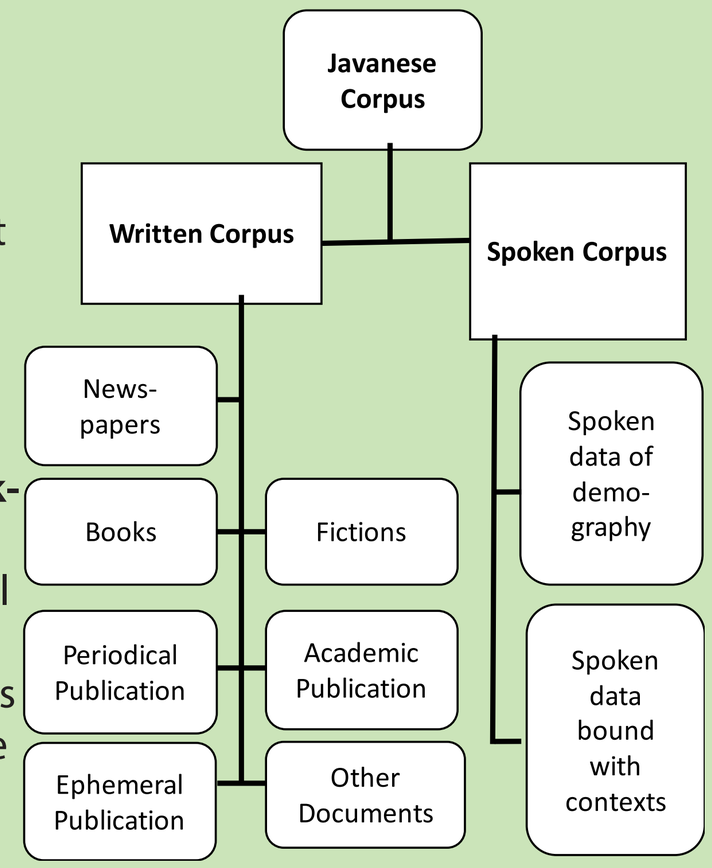
1. save the corpus text data of regional languages in digital form
2. process and store corpus metadata so that it can be accessed by other software
3. analyze corpus text data by corpus methods such as keyword lists, concordances, n-grams, etc.

The users of corpus tools can be categorized into several types, such as lexicographers, linguistics researchers and students, and language teachers and learners.

LANGUAGE DATABASE

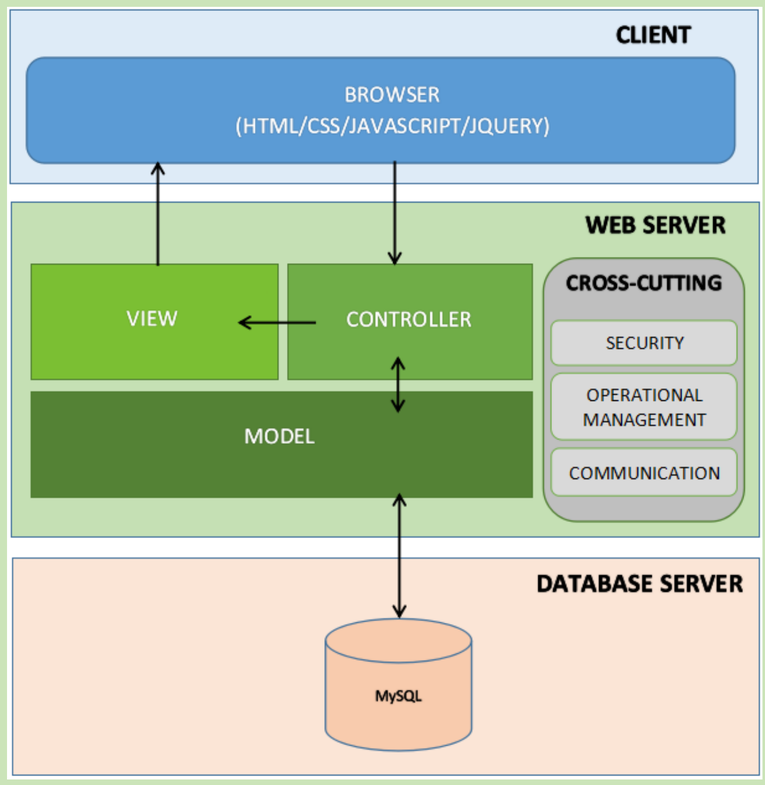
The design of corpus data for our web-based application is decided by considering:

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

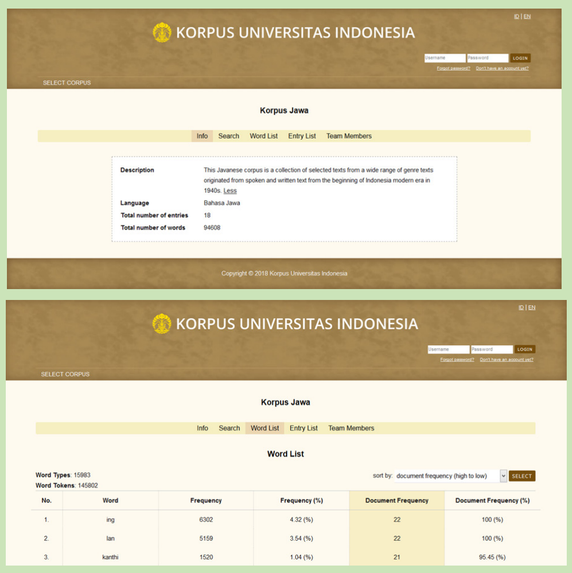


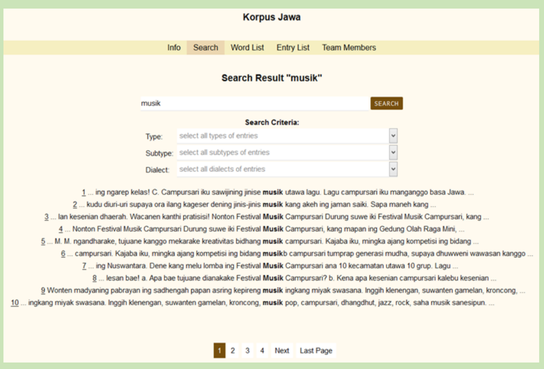
This research was supported by The Directorate of Higher Education, The Ministry of Education Republic of Indonesia with Research Grant No. 1/E1/KP.PTNBH/2019

DESIGN AND ARCHITECTURE

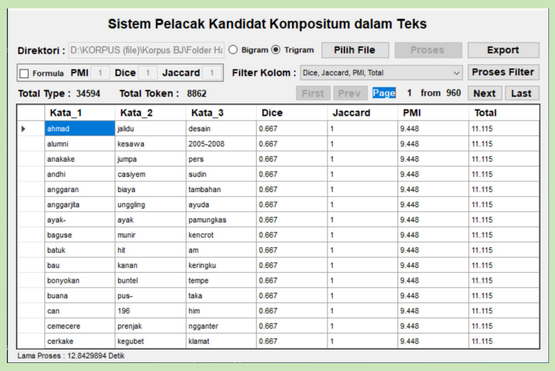


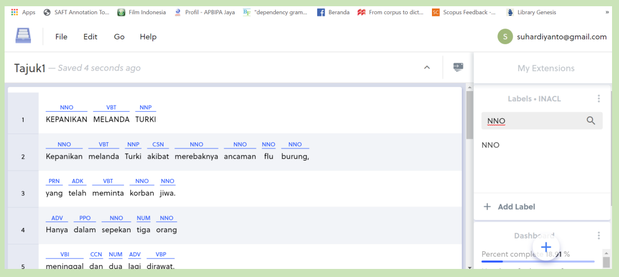
FEATURES AND FUNCTIONALITIES

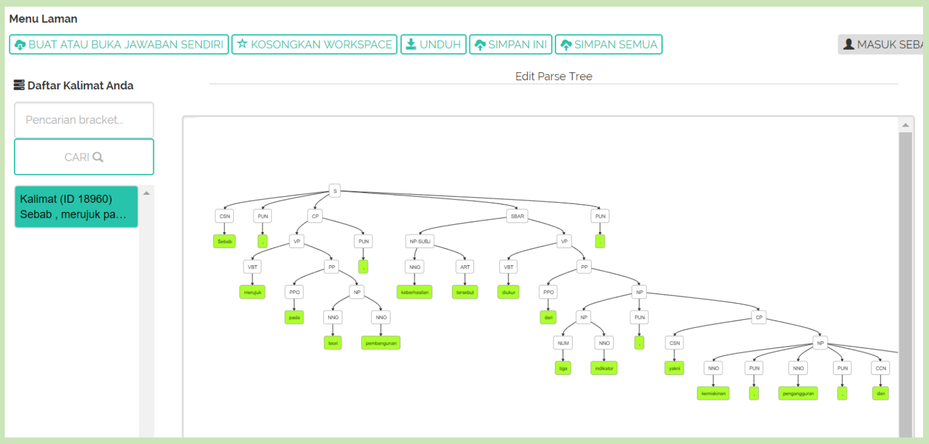




1. Select Corpus
2. View Word List
3. View Concordance Search Result
4. View Entry List
5. Generate MWE
6. Annotate Texts







REFERENCES

Simons G F and Fennig C D (eds) 2018 Ethnologue: Languages of the World 21st ed (Dallas: SIL International)
Ishida T, Murakami Y, Lin D, Nakaguchi T and Otani M 2018 Computer 51 72-81 ISSN 0018-9162
Nasution A H, Murakami Y and Ishida T 2018 ACM Trans. Asian & Low-Resource Lang. Inf. Process. 17 9:1-9:29 URL http://doi.acm.org/10.1145/3138815
McEnery T, Xiao R and Tono Y 2006 Corpus-based Language Studies: An Advanced Resource Book (London/New York: Routledge)
Sinclair J 2005 Developing Linguistic Corpora: a Guide to Good Practice ed Wynne M (Oxford: Oxbow Books) pp 1-16
Atkins S B T and Rundell M 2008 The Oxford Guide to Practical Lexicography (Oxford: Oxford University Press)
Kilgarriff A and Kosem I 2012 Electronic Lexicography ed Granger S and Paquot M (Oxford: Oxford University Press) pp 83-106
Anthony L 2006 Proceedings of the JACET 45th Annual Convention pp 218-219
Kilgarriff A, Baisa V, Buta J, Jakubek M, Kov V, Michelfeit J, Rychl P and Suchomel V 2014 Lexicography 7-36
Scott M 2016 WordSmith Tools version 7 (Stroud: Lexical Analysis Software)