Indigenous Languages of Indonesia: Creating Language Resources for Language Preservation

Hammam Riza
IPTEKNET
Agency for the Assessment and
Application of Technology (BPPT)
Jakarta, Indonesia
hammam@iptek.net.id

Abstract

In this paper, we report a survey of language resources in Indonesia, primarily of indigenous languages. We look at the official Indonesian language (Bahasa Indonesia) and 726 regional languages of Indonesia (Bahasa Nusantara) and list all the available lexical resources (LRs) that we can gathered. This paper suggests that the smaller regional languages may remain relatively unstudied, and unknown, but they are still worthy of our attention. Various LRs of these endangered languages are being built and collected by regional language centers for study and its preservation. We will also briefly report its presence on the Internet.

1 Introduction

It is not hard to get a picture of just how linguistically diverse Indonesia is. There are 726 languages in the country; making it the world's second most diverse, after Papua New Guinea which has 823 local languages (Martí et al., 2005:48). Indonesia also has a high ratio of languages to speakers in each major region in Indonesia (see Figure 1). Diversity is the outcome of processes of language change (Schendl, 2001). The loss of language is itself a process that will logically result in monolingualism.

It is not uncommon to find the attitude among the general public and even among some Indonesian linguists that the process of language endangerment or language extinction is not something that needs worried about, that it is part of a natural process that should be left to take its course. This paper suggests otherwise. The smaller regional languages may remain relatively unstudied, and unknown, but they are still worthy of our attention (Lauder, 2007). This paper puts forward a number of claims that have been made in favour of linguistic diversity and how we can preserve this diversity.

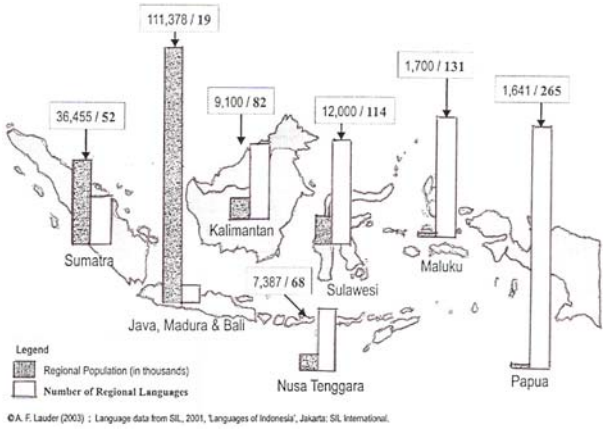

Figure 1. Ratio of Population to Languages across Indonesia

The languages of Indonesia are part of a complex linguistic situation that is generally seen as comprised of three categories: Indonesian language, the regional indigenous languages, and foreign languages (Alwi and Sugono, 2000). Most of these regional languages have not received attention for computerization; they are less privilege languages that need to be brought into digitalization.

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 113–116,
Hyderabad, India, January 2008.

If we were to create NLP system for these languages, we will face one of the major obstacles, i.e. the amount of linguistic knowledge. Language analysis and generation require a complete set of lexical, grammatical, semantic and world knowledge to carry out accurate function. On the other hand, these types of knowledge bases are hard to acquire and considerable attention has to be paid to the role that corpus and lexical resources can play.

2 The Indigenous Languages and Its Endangerment

The indigenous languages of Indonesia - also referred to as vernaculars or provincial languages, collectively called as Bahasa Nusantara - exhibits great variation in numbers of speakers. Thirteen of them have a million or more speakers, accounting for 69.91% of the total population. These languages are Javanese (75,200,000 speakers), Sundanese (27,000,000), Malay (20,000,000), Madurese (13,694,000), Minangkabau (6,500,000), Batak (5,150,000), Bugisnese (4,000,000), Balinese (3,800,000), Acehnese (3,000,000), Sasak (2,100,000), Makasarese (1,600,000), Lampungese (1,500,000), and Rejang (1,000,000) (Lauder, 2004).

Figure 2. Major Indigenous Languages

The remaining 713 languages have a total population of only 41.4 million speakers, and the majority of these have very small numbers of speakers. For example, 386 languages are spoken by 5,000 or less; 233 have 1,000 speakers or less; 169 languages have 500 speakers or less; and 52 have 100 or less (Gordon, 2005). These languages are facing various degrees of language endangerment (Crystal, 2000).

There is evidence from census data over three decades that the growth in the numbers of speakers of Indonesian is reducing the numbers of speakers of the indigenous languages (Lauder, 2005). Concerns that this kind of growth would give Indonesian the potential to replace the regional languages were aired as early as the 1980s. (Poedjosoedarmo, 1981; Alisjahbana, 1984).

These languages tend to be spoken in the eastern or more remote parts of the country. Their small populations of speakers make them vulnerable to processes of unhealthy language change and language endangerment. Greatest language diversity is found in eastern part of Indonesia (Papua)

A language that does not have official status, but which has a large enough number of speakers and which are being safely transmitted to new generations can usually be classified as either NOT ENDANGERED (SAFE or VIABLE), or POTENTIALLY ENDANGERED. This would include the 13 largest local languages and perhaps a few dozens of others.

However, this does not apply to the majority of the remaining 700 or so languages. Among them, there should also be many which could be classified as VIABLE BUT SMALL or ENDANGERED because they have small numbers of speakers, are socially or economically disadvantaged and they are not being transmitted to younger generations of speakers. There will also be many of these regional languages which can be classified as SERIOUSLY ENDANGERED or MORIBUND (NEARLY EXTINCT) because the speaker populations are very small and these few remaining speakers are mostly old.

When trying to estimate the degree of endangerment of the regional languages in Indonesia, it becomes apparent that there is a singular lack of focused and comprehensive research. However, in spite of this, based on a consideration of the various possible causes, there are good reasons to suspect that many of the smaller languages in Indonesia are indeed endangered.

3 Preserving Endangered Languages

Within Indonesia, and globally, we are currently experiencing a massive and rapid loss of language and culture. In particular, the languages and cultures of communities with very few speakers have practically no chance of survival beyond the end of

this century and many will disappear much sooner, perhaps within the next 10 to 20 years.

The loss of these languages is largely because of linguistic and cultural assimilation with the majority group, with migration to the cities and lack of support for these languages in state education being important factors. This is particularly true in Indonesia, where Bahasa Indonesia is being taught in school and the indigenous languages are loosing their ground in the daily life.

Each language is part of patterns of diversity that have evolved over millennia. There are a number of reasons why diversity is beneficial. For example, by learning from the original languages we increase our stock of human wisdom. Diversity breeds diversity; the seeding of insights in the fields of science, art and literature.

Meanwhile, the problem is urgent. A language is being lost on average every two weeks worldwide. When an oral language is lost, it takes with it all the knowledge that the people possessed. When the last speaker dies, there is likely to be no trace at all of their existence. There will be no artifacts or physical record to reconstruct the language or the knowledge it encoded. As each language dies, we lose data for philosophers, anthropologists, folklorists, historians, psychologists, linguists, and writers. The loss of one is a tragedy; what do we call the loss of a large proportion of the 6,000 existing languages? (Crystal, 2000: 53). The loss of diversity is something that we need to do something about.

Two kinds of action can be taken, depending on the status of the language. But to know what the status of languages is, a survey needs to be made to gather information for all the regional languages concerning the factors that are usually the causes of language loss or language maintenance, such as numbers of speakers, language attitudes, and so on. As a result, estimates can be made about which are likely to survive and which not. From this, an action plan can be set up based on priorities.

Of the 13 major indigenous languages, there only are 7 languages presence on the Internet under the ccTLD .id (Riza 2006). We need to explore furthermore to map the remaining regional languages that probably exist on the Internet. The issue of 'digital language divide' has shown that many of the indigenous language do not have access to Information and Communication Technology (ICT) in general; hence they are lacking the process of digitalization. The relationship between languages on the Internet and diversity of language within a country indicates that even with a globalize network, nation states have a role to play in encouraging language diversity in cyberspace. Language diversity can be viewed as much within a country as within the Internet as a whole.

For languages which are not seriously endangered or moribund, language maintenance and language revitalization programs should be put in place. These programs include creating LRs that would involve the people themselves to provide them with NLP toolkit and language computerization to help keep the language alive. For seriously endangered languages, those that cannot possibly be saved, LRs creation should be set up. These programs would involve study, documentation and the assembly of a rich archive of materials that will help to preserve as much as possible of the language and way of life in digital and other formats.

We have identified three important tasks in language preservation. The first is the exploitation of current techniques from computational linguistics to permit a multidimensional view of the LRs. The second is the increasing orientation of the regional research centers towards the creation and use of resources of various sorts, either to extract useful information or directly as components in systems. The third, related, trend is towards statistical or empirical models of language especially if the language is near extinction and found only as spoken language.

In cases where the indigenous languages exist only in the form of spoken language, there should be a collection efforts similar to the work carried out by ELRA on the Basic Language Resource Kit (ELDA, 2007) and LDC on Less Commonly Taught Languages (LCTL, 2007). Both initiatives focus on the minimal sets of LRs required developing basic research for a given language. It is crucial to connect the preservation work to this language kit in order to be shared with the language research community.

In Indonesia, over the last few years, there has been an increasing awareness of the importance of corpus resources in language preservation. As regional leaders begin to consider the implications of loosing their indigenous assets, considerable attention is being aid to the role that corpus and lexical resources can play.

Masyarakat Linguistik Indonesia (MLI) is a group of institutions, organizations and corporation, working together on mutually defined goals and projects that seek to provide a specification of LRs of all languages of Indonesia. It is currently in the process of mapping indigenous written languages of Indonesia (540 of languages).

MLI also help members to use the specification for NLP tools and applications; find the best means to disseminate the specifications, tools and applications and encourage an open standard-based approach to the creation and interchange of LRs. It also demonstrate how MLI can be applied through making the results of collaborative endeavors available to wider associations; provide training, awareness and educational events and share with each other their work on related issues.

4 Conclusion

A perspective on preservation of the languages of Indonesia is given together with a brief overview of some of the indigenous languages, which are being actively researched today by national language centers throughout Indonesia.

Culture and language are fundamental human rights; it is our right and duty to preserve and develop them. This is an ethical choice, not simply a scientific one or one based on political or economic expediency. Total lack of concern and inaction may seem to some to be a rational choice but it represents an ethical failure. In addition, research which merely documents an endangered language but does nothing to help the community of the informants is like the photographer who takes a picture of someone in difficulty but do nothing to help them. Any delay now will mean that many of the languages which are still around now won't be there for them to do something about. Diversity will have been lost.

We have identified three important tasks in language preservation, which is the exploitation of computational linguistics, increasing orientation of the regional research centers towards the creation and use of resources and using towards statistical or empirical models of language.

The current effort of documenting the indigenous languages will be shared with the rest of the world, to close 'digital language divide'.

References

Alisjahbana, S. T. 1984. The problem of minority lanfguages in the overall linguistic problems of our time. *In Linguistic Minorities and Literacy: Language Policy Issues in Developing Countries*, ed. F. Coulmas. Berlin: Mouton.

Alwi, Hasan, and Sugono, Dendy. 2000. From National Language Politics to National Language Policy. *Procedings of the Seminar on Language Politics*, Jakarta

ELDA. 2007. Basic Language Resource Kit (Blark). ELRA Project, http://www.elda.org/blark/index.php

Crystal, David. 2000. *Language Death*. Cambridge: Cambridge University Press.

Gordon, Raymond G., Jr. ed. 2005. *Ethnologue: Languages of the World, Fifteenth edition*. Dallas, Tex.: SIL International.

Lauder, Multamia RMT. 2005. Language Treasures in Indonesia. In Words and Worlds : *World Languages Review*, eds. Fèlix Martí et al., 95-97. Clevedon [England] ; Buffalo [N.Y.]: Multilingual Matters.

Lauder, Allan F. 2007. Indigenous Languages in Indonesia: Diversity and Endangerment. *In Proceedings of Kongres Linguistik Nasional XII*, Surakarta, 3-6 September.

LCTL, 2007. Less Commonly Taught Language Project. http://projects.ldc.upenn.edu/LCTL/index.html

Martí, Fèlix, et.al. eds. 2005. *Words and Worlds : World Languages Review*. vol. 52. Bilingual Education and Bilingualism. Clevedon. England.

Mikami, Y., Zavarsky, et.al. 2005. The Language Observatory Project (LOP), www2005, *Proceedings, Chiba*, Japan, 990-991.

Poedjosoedarmo, S. 1981. Problems of Indonesian. *In Language and Nation Building*, ed. Amran Halim. Jakarta: Center for Language Development.

Riza, H, et. al. 2006. Indonesian Languages Diversity on the Internet, Internet Governance Forum (IGF), Athens.

Schendl, Herbert. 2001. *Historical linguistics*. Oxford Introductions to Language Study. Oxford: Oxford University Press.

Wurm, S. A. 1998. Methods of language maintenance and revival, with selected cases of language endangerment in the world. *The International Symposium on Endangered Languages*, Tokyo, 18-20 November 1995), ed. Kazuto Matsumura, 191-211. Tokyo: Hituzi Syobo.