Resources Report on Languages of Indonesia

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 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).

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).

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 – 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: 3-4). Of these 13 languages, only 7 languages have presence on the Internet (Riza 2006).

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).

2 Language Resources

Many language centers in Indonesia have embarked in various research and development in creation of language resources (LRs). Unfortunately, this development mainly only focused on creating LRs for the official language Bahasa Indonesia. In the followings, we describe the present and ongoing LRs research projects with emphasis on the indigenous languages.

2.1 Indonesian Electronic Dictionary System (KEBI)

Our laboratory has worked to enlarge and improve the quality of Indonesian electronic dictionaries. Starting 1987, it took us at least 4 years to develop all necessary components for CICC-MMTS, resulting with many first-ever Indonesian language resources, primarily electronic dictionaries and grammar rules for language analysis and generation. This extension have resulted in a collection of 500,000 word entries and more than 2 million derivational and inflected words. As part of this research, we built an online access to the dictionaries (http://nlp.inn.bppt.go.id/kebi) enabling users to add new words and definition. KBBI electronic dictionary is scheduled to be launched in 2008, during 100 years celebration of the official Bahasa Indonesia.

2.2 BPPT-ANTARA Corpus

This parallel corpus was developed as extension to Indonesia National Corpus Initiative (INCI) which was earlier created to support the development of a hybrid stochastic-symbolic system BIAS-II. Currently, a pure statistical MT system based on Pharaoh is developed by BPPT and National News Agency (ANTARA) using 500K sentences pair, expected to have better accuracy and robustness and could enhance the quality of translation (current BLEU score 0.72).

2.3 Regional Languages Mapping (National Language Center)

For the past 15 years, the Indonesia National Language Center have been collecting information regarding all indigenous languages. By the end of this year, this project will be completed and all result and findings will be open to public.

2.4 Dictionaries of Bahasa Nusantara, Indonesian Linguistics Association (MLI)

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. MLI also help members to use the specification for 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 to Asian Language Resource (ALR) through making the results of collaborative endeavors available throughout the members of the group and wider associations; provide training, awareness and educational events and share with each other their work on related issues.

2.5 Speech Corpus

In a mission to improve the quality of automatic speech recognition (ASR), a collaboration of Telkom RDC and ATR-Japan has constructed speakers' corpus (40 speakers, 2000 sentences) which is expected to improve the accuracy of ASR to 90% level.

2.6 Other Corpus

Other monolingual corpus is found online. The major news articles corpora on the web is Tempointerakif.com (56,471 articles). Kompas corpus (71,109 articles) can be found at http://ilps.science.uva.nl/Resources/BI.

References

Alisjahbana, S. T. 1984. The problem of minority languages 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

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

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.

Martí, Fèlix, et.al. eds. 2005. Words and Worlds : World Languages Review. vol. 52. Bilingual Education and Bilingualism. Clevedon [England] ; Buffalo [N.Y.]: Multilingual Matters.

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