ＩＳＳＮ １８８２－８９５７　第11回科学と人間シンポジウム論文集（抄録集：第6巻）

抄録

Samawa Language Project: Its Policy and Approach

Trienani Hariyanti[1,2], Saori Aida[3], Hiroyuki Kameda[3]
[1]*Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology*
[2]*Department of Computer Science, Sumbawa University of Technology*
[3]*School of Computer Science, Tokyo University of Technology*
E-mail: trienanihariyanti@gmail.com

Abstract

Samawa language is a local language in Sumbawa Island, Indonesia. It is a symbol and represents the culture of Sumbawa ethnic. Recently, Samawa has received less attention. There is an only small number of physical documented evidence towards the transfer of Sumbawa cultural knowledge through Samawa. Besides, the absence of Sumbawa local government's policy related to an effort to preserve the language against further reducing the current state of the language resources. As a consequence, Samawa can be categorized as under-resourced language. This paper aims to present an integrated planning as the first effort to collect and develop more resources related to Samawa and Natural Language Processing (NLP) tools in the near future. We also describe the description, ideas and challenges for building this project based on the local government policy and technical approach. Then, we design research schedule and provide an overview of two main subprojects regarded to the digital archive of Samawa.

1. Introduction

There are many natural languages on the Earth. The Ethnologue [1], a catalogue website of world languages reports the number of living languages are over 7000. However, many languages except Chinese, Spanish, English, and others are in danger disappearing. Most of them are extremely small amounts of available resources, getting dying and extinct. Samawa language, used in daily life on the Sumbawa Island in Indonesia, is one of such languages.

As a part of world cultural heritage, Samawa language (hereinafter simply "Samawa") belongs to the Malayo-Sumbawan subgroup which is a member of Malayo-Polynesian in the Austronesian language family [2]. Samawa as the language of Sumbawa ethnic is spoken for more than 300 years by around 582,575 people [3]. Since 1674, Samawa has been actively used in the level of Sumbawa sultanate and government. The speakers commonly used it for daily conversation, songs, poems, and cultural events.

Recently, Samawa has received less attention. There is only a small number of physical documented evidence towards the transfer of Sumbawa cultural knowledge through Samawa. Besides, the absence of Sumbawa local government's policy related to an effort to preserve the language against further reducing the current state of the language resource. Nowadays, many schools in Sumbawa from elementary to senior high school no longer provides Samawa subject in their class. Then, there are overwhelmingly few books written in Samawa. These are made Samawa be categorized as an under-resourced language. We are concerned that Samawa as the inseparable part of a culture which degrades local wisdom values in the past cannot survive for long periods of time. As a consequence, the next generation may lose Samawa as their mother tongue. Creating Samawa documentation in the form of collecting, analyzing and making

抄録

readily accessible to interested people is obviously the most useful to keep the existence of Samawa.

In this paper, we describe our integrated planning as the first effort to collect and develop more resources related to Samawa and Natural Language Processing (NLP) tools in the near future. Section 2 talks about several studies related to Samawa in the past. We present the description, ideas and challenges for building this project based on the government policy and technical perspectives in section 3. Also, we design our research schedule, discuss two mainly subprojects regarded to digital archive and summarize our conclusions in section 4, 5, 6 and 7, respectively.

2. Related Work

Some previous works have been addressed to Samawa and its resources. A researcher team of Teaching Faculty, Udayana University initiated research regarded to the structure of Samawa in 1980. At the same year, Arifin proposed general of Samawa morphology in Sumbawa Besar [4]. Six years later, Sumarsono et al. [5] provided more detail about phonology, morphology and grammatical rules of Samawa. In 1990, Mahsun in [6] focussed on Jereweh dialect morphology and tried to solve some affixes problem based on their correlation and meaning. He also presented a mapping Samawa based on four main dialects, i.e. Sumbawa Besar, Taliwang, Jereweh and Tongo [7]. Then, Kasman explored the position and function of the lingual form of Tongo dialect [8]. Interestingly, Samawa as an object research also has been done by Asako Shiohara, a researcher from Tokyo University of Foreign Studies, Japan. He tried to explore voice system in Samawa [9], deixis in Sumbawa Besar dialect [10] and some research related to grammatical rules of Samawa. Also, he collected some monologues, legends and folktales as language resources of Sumbawa in the form of voice recording at [11].

3. Samawa Language Project

In this section, we present the goal and describe the ideas and challenges related to project based on political and technical viewpoint. Also, we show the current and ideal structure of staff for supporting us to solve this project as well.

3.1. Project Goal and Objectives

The goals of Samawa language project is to collect and develop more resources as many as possible in digital archive format and create NLP tool sets for Samawa. In achieving these goals, we will address four primary objectives such as:

Objective 1: To collect and store manuscripts, textbooks, magazines, song, poem, text from website, audio and video in Samawa as language resources in digital format.

Objective 2: To build text corpora, i.e., text corpus, tagged corpus and parallel corpus.

Objective 3: To deliver Samawa language resources on a website and make available online.

Objective 4: To develop NLP tool sets, i.e., part of speech tagger and machine translation.

Moreover, we limit our focus on Sumbawa Besar dialect which is the standard dialect of Samawa. Besides, it has a large number of the speaker which spread in average in the western part of Sumbawa Island.

抄録

3.2. Approach Adopted from the Political Viewpoint

Bahasa Indonesia as the national language of Indonesia has an enormous impact in every level start from education, government and media. It was noted in the constitution, chapter XV and article 36 that the language of the state is Bahasa Indonesia, after independence day on August 17th, 1945. Furthermore, the law number 24 year 2009 has arranged it as an official language, whereas local and foreign language as a support function. It means people might be used local and foreign language to communicate with each other when Bahasa Indonesia could not used effectively [12].

On the other hand, Indonesia also guaranteed that the local language that is well preserved by the people would be protected and maintained by the state in the constitution. Even the state obliges to local government to develop, maintain, and protect the local language in order to sturdy the position and function in the society as a part of cultural heritage (see the law number 24 year 2009 article 42 in [13]). Some province in Indonesia has been implemented these laws in the form of local regulations, i.e., Central Java, West Java, and Bali.

For instance, Central Java province with their local regulations in [14] mandated the allocation of Javanese subject at least two hours each week at every level in the school. The government also encouraged the people to use Javanese once a week, communicate each other in a formal meeting, and support electronic media to provide a program using Javanese. In another province, Bali stipulated the Balinese, script and philology taught at all levels of elementary and senior high school as subjects. They appointed the teacher of Balinese as a professional teacher and provided funding to preserve the language (see more detail in their regulations in [15]). In fact, we have not found any references to Sumbawa government's policy related to Samawa.

In December 2016, Lembaga Adat Tana Samawa (LATS) which is a customary institution of Samawa people held a big forum namely Mudzakarah Rea LATS. Many humanist attended it included the Sultan of Sumbawa and determined 40 recommendations related to cultural and language preservation [16]. We totally agreed with all recommendations, especially for Samawa preservation. They suggest to the local government to emphasize a policy in the field of education that enforce a curriculum based on Samawa and its philology in all levels of school. Besides, they encouraged the government to use Samawa as a language of instruction in learning process once a week, organization activity, company and administration.

Reflected to local regulations of another province, it is the best action when Sumbawa local government can follow them. We also propose some essential points which need in building Samawa language policy. These are:

1. To organize competition as a stimulant to the younger speaker to pride and not averse to use Samawa.
2. To support and manage electronic media, i.e., radio, website and newspaper to allocate a particular space in their place.
3. To collaborate with some institutions in Sumbawa to present training workshop, study and research in depth.

The importance of the language policy is designed to protect and promote local language systematically to the society by the local government. Also, the government can provide punishment appropriate with its portion when the society does not implement it. Finally, the

抄録

policy will speed up the growth of Samawa resources, and it can be used for many applications of human language technology.

3.3. Approach Adopted from the Technical Viewpoint

Another way to keep a language alive is to make it as a research object, primarily related to human language technology. Natural Language Processing (NLP), a branch of artificial intelligence which intersects with linguistic has played an important role in most aspects of a language. The ultimate goal of research in NLP is to make computers understand human language. There are many exciting researches in NLP such as information retrieval, question answering, text summarization and machine translation. All they need is a language resources as the raw material. In this subsection, we provide the current and some possible tasks in NLP which conducted shortly.

Text Corpus. A text corpus is a collection of written text which has been selected and represents a language in use for linguistic studies. It is the first language resource we have now. We collected manuscripts, books, magazines, folktales from Sumbawa archive office and language center office of NTB at various times.

Tagged Corpus. Tagged corpus is a kind of corpus which each word or token in a sentence labelled with its part of speech information. It is a fundamental task which needs for high-level NLP task such as machine translation. Currently, we have designed 24 part of speech and tagged a corpus which consists of 11,799 tokens.

Parallel Corpus. Since Samawa resources are strongly limited in size and variation, we can extend our corpus by translating Indonesian corpus into Samawa. The collection of the result of the translation and the source (Bahasa Indonesia and Samawa) are called parallel corpus.

Automatic Tagger. Manually assigning part of speech information needs a time-consuming and challenging process. We have to use an automatic tagger to speed up the tagging process.

Machine Translation. A machine translation is used to translate Samawa into Bahasa Indonesia automatically. We will use our parallel corpus to training our machine translation based on statistical approach.

All NLP task above will create more resources for Samawa although it is categorized as an under-resourced language. By collaboration with linguistics, we can realize these projects into account as well as possible.

3.4. Structure of the Staff

We conducted this project in Thought and Language laboratory under the supervision of Professor Hiroyuki Kameda and Professor Saori Aida. Currently, a team consists of 2 staffs, i.e. one staff as an annotator and programmer at once. The other one charged to type all raw materials manually.

Since NLP is an intersection between artificial intelligence and linguistics, then we propose in advance the ideal team to build this project. Table 1 below shows the ideal team and their task description.

Table 1. The propose NLP team of Samawa project.

Member of Team	Number of People	Task Description

Supervisor	2	Provide guidance and assistance to all member of team on their respective roles and responsibilities
Project Manager	1	Manage the planning, executing monitoring, and controlling the project
Annotator	3	To type manually and annotate the corpus
Linguist for Bahasa Indonesia	1	Evaluate grammatical structure of the result of the translation corpus for Bahasa Indonesia
Linguist for Samawa	1	Evaluate grammatical structure of the result of the translation corpus for Samawa
Translator	2	Translate the document from Samawa to Indonesia and vice versa
NLP engineer	2	Create language model and implement algorithms.

4. Project Timeline

We propose research schedule which spans across a 36-months period. Samawa project will be conducted in six phases, each of six months long. It was inspired by previous project schedule on language resources of Bahasa Indonesia [17]. Then, we modify the needs and adjust to the number of members who conducted this project. Table 2 shows our proposed work for 36 months.

Table 2. Samawa project for 36 months.

Phases	Research Activities
First Phase	▪ Initial research report on an overview of Samawa project, data collection, metadata description, corpus design, and cleaning tools ▪ 100,000 words in Samawa corpus with its part of speech information ▪ Research report on Samawa tagset and tagged corpus
Second Phase	▪ Store our resources languages into Samawa archive in XML format and make available online ▪ Translate 100,000 words into Bahasa Indonesia from Samawa corpus ▪ 100,000 words translated from Indonesian corpus into Samawa ▪ Develop part of speech tagger ▪ Research report on parallel corpus and part of speech tagger
Third Phase	▪ Additional 150,000 words translated from Indonesia Corpus into Samawa ▪ Initial design of Statistical Machine Translation (SMT) framework for Samawa ▪ Research report the challenges and issues related to parallel corpus development.
Fourth Phase	▪ Store our resource languages into a website and make available online

	▪ Evaluate the result of automatic tagger of 150,000 words in the third phase ▪ Final design of SMT framework for Samawa ▪ Research report on evaluation of automatic tagger and SMT framework
Fifth Phase	▪ Additional 150,000 words translated from Indonesia corpus into Samawa ▪ Initial prototype of SMT for Samawa
Sixth Phase	▪ 500,000 tagged corpus of Samawa ▪ SMT from Samawa to Bahasa Indonesia ▪ Research report on SMT for Samawa to Indonesia

5. Subproject 1: Samawa Archive

Storing and long-term preservation of Samawa is one of the keys to keeping the language survive. This subproject aims to digitalize all documents we have collected before in digital format. First of all, request for copyright permission must be sought from the owner first. All owners have to sign a declaration which allows their materials to be used for the creation of Samawa archive and only for academic research purpose.

Furthermore, we will store our language resources in two formats. Firstly, we will type them manually in notepad++ with Unicode encoding standard of UTF-8 format. This format is flexible and portable across platform [18]. Secondly, we develop a program to convert the result in the first step in eXtensible Markup Language (XML) based on Text Encoding Standard (TEI) guidelines [19]. Then, we will be made searchable on a website or a repository under Creative Commons license.

6. Subproject 2: NLP tool sets specified to Samawa

In this section, we present the overview of two main NLP tools we are going to develop. There are automatic tagger and machine translation of Samawa.

6.1. Automatic Tagger

Part of speech tagging is a process to assign a word or a token with its part of speech information. It is hard to handle manually and much time-consuming for tagging a corpus. Automatic tagger can be used for speeding up the process. In principle, there are two main approaches for building automatic tagger, i.e., statistical and rule-based approach.

Statistical approach finds the best tag in the unannotated text with the most frequently used tag in annotated training data. The weakness of this approach sometimes appears sequences of tags for sentences that inappropriate to the grammatical structure of a language. Conditional Random Fields, Maximum Entropy and Hidden Markov Model are commonly used for statistical approach.

On the other hand, rule-based approach determines the tag of a word based on handwritten rules list which are the lexical and grammatical information. However, for unknown words, rule-based cannot find the precise tag. Transformation-Based Error-Driven Learning (TEL) in [20] which is presented by Eric Brill in 1995 is one of the popular rule-based approaches.

抄録

We specify to use the statistical and rule-based approach in our project and try to compare both of them. We will see which one gives the best accuracy and performance for modelling of Samawa. Then, we will choose the best one and implement it to extend our corpus.

6.2. Statistical Machine Translation

Machine translation is an NLP tool which uses to translate text or speech from one natural language to another one using computers. Fundamentally, the process of translation has two levels, i.e., metaphrase and paraphrase. Metaphrase means word-to-word translation, which is translated version will have “literal” translation of each word in the text. Paraphrase means the translated text would contain the gist of the original text but may not necessarily contain the word-to-word translation [21].

Statistical Machine Translation (SMT) is one of the famous approaches and dominates machine translation research in last a few decades. It had introduced by Warren Weaver in 1949. SMT is classified into corpus-based which generates translation using statistical methods. According to Lopez in [22], there are four main points which must be focussed to build functioning SMT system, i.e., translational equivalence model, a parameterization of the model, parameter estimation, and decoding process. In Samawa, we will use these points to develop our SMT and train 500,000 words as parallel corpus of Sawama-Indonesia.

7. Conclusions

This research focuses on integrated planning to collect, store and develop more language resources of Samawa. We proposed some ideas and recommendations to the Sumbawa local government to follow the success stories of some provinces that have been implemented the policies related to their local language. Also, we provided planning to create Samawa archive in digital format and NLP tool sets (automatic tagger and statistical machine translation) which can be used to increase Samawa resources. In the future, we would like to implement our project based on this guidance.

Acknowledgment

Mrs. Trienani Hariyanti would thank to Indonesia Endowment Fund for Education (LPDP) for acknowledge gratefully funding on her study.

References

[1] Simons, G. F., and C. D. Fennig, “Ethnologue: Languages of the World,” *Ethnologue*, 2017. [Online]. Available: https://www.ethnologue.com/. [Accessed: 09-Feb-2018].
[2] A. Shiohara, “Information Structure and Information Status in the Sumbawa Language of Indonesia,” in *Proceedings of the International Workshop on Information Structure of Austronesian Languages*, 2014, pp. 221–228.
[3] Badan Pusat Statistik, “Provinsi Nusa Tenggara Barat Dalam Angka 2017,” *BPS Provinsi Nusa Tenggara Barat*, 2017.
[4] Margono, I. G. Ngurah Bagus, A. Meko Mbete, I. N. Sulaga, and I. N. Sudipa, *Fungsi Bahasa Sumbawa*. Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan, 1986.
[5] Sumarsono, M. Nadera, S. Basuki, N. Merdhana, and N. Mertha, *Morfologi dan Sintaksis Bahasa Sumbawa*. Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan, 1986.
[6] Mahsun, “Morfologi Bahasa Sumbawa Dialek Jereweh,” Universitas Gadjah Mada, Yogyakarta, 1990.
[7] Mahsun, “Penelitian Dialek geografis bahasa Sumbawa,” PhD Thesis, Universitas Gadjah Mada, 1994.
[8] N. F. N. Kasman, “Kedudukan dan Fungsi Satuan Lingual/q/,/n/,/èn/, /èng/, dan /s/ dalam Bahasa Samawa Dialek Tongo Subdialek Lebangkar,” *Kandai*, vol. 12, no. 2, pp. 241–254, 2016.
[9] A. Shiohara, “Voice system in Sumbawa,” in *Workshop on Indonesian-type Voice System, Tokyo University of Foreign Studies, http://lingdy. aacore. jp/en/activity/indonesiatam/docs. html*, 2010.

抄録

[10] A. Shiohara, “Deixis in the Sumbawa Besar Dialect of Sumbawa,” *NUSA: Linguistic studies of languages in and around Indonesia*, vol. 56, pp. 139–152, 2014.
[11] A. Shiohara, “Language Resource of Sumbawa,” 2015. [Online]. Available: http://id-lang-rc.aa-ken.jp/?page_id=118. [Accessed: 17-Feb-2018].
[12] Sugiyono, “Pelindungan Bahasa Daerah dalam Kerangka Kebijakan Nasional Kebahasaan.” [Online]. Available: http://badanbahasa.kemdikbud.go.id/lamanbahasa/content/pelindungan-bahasa-daerah-dalam-kerangka-kebijakan-nasional-kebahasaan. [Accessed: 18-Feb-2018].
[13] P. R. Indonesia, “Undang-Undang Republik Indonesia Nomor 24 Tahun 2009 Tentang Bendera, Bahasa, dan Lambang Negara, serta Lagu Kebangsaan,” 2009. [Online]. Available: http://badanbahasa.kemdikbud.go.id/lamanbahasa/sites/default/files/UU_2009_24.pdf. [Accessed: 19-Feb-2018].
[14] P. P. Jawa Tengah, “Peraturan Gubernur Jawa Tengah No 57 Tahun 2013 tentang Bahasa, Sastra dan Aksara Jawa,” 2013. [Online]. Available: http://jdihukum.jatengprov.go.id/download/produk_hukum/pergub/pergub_tahun_2013/pergub_57_th_2013.pdf. [Accessed: 18-Feb-2018].
[15] P. P. Bali, “Peraturan Gubernur Bali Nomor 20 Tahun 2013 tentang Bahasa, Aksara dan Sastra Daerah Bali pada Pendidikan Dasar dan Menengah,” 2013. [Online]. Available: http://jdih.baliprov.go.id/uploads/produk-hukum/peraturan/2013/PERGUB/pergub-20-2013.pdf. [Accessed: 18-Feb-2018].
[16] LATS, “40 Rekomendasi Mudzakarah Rea Lembaga Adat Tana Samawa,” 2016. [Online]. Available: http://pulausumbawanews.net/index.php/2016/12/19/ini-40-rekomendasi-mudzakarah-rea-lembaga-adat-tana-samawa/. [Accessed: 11-Feb-2018].
[17] M. Adriani and H. Riza, “Research Report on Local Language Computing: Development of Indonesian Language Resources and Translation System,” *Ref. No: PANL10n/Admn/RR/001, PAN Localization Project*, 2008.
[18] C. Chungku, J. Rabgay, and P. Choejey, “Dzongkha Text Corpus,” in *Conference on Human Language Technology for Development*, 2011, pp. 34–38.
[19] T. E. I. Consortium, *TEI P5: Guidelines for electronic text encoding and interchange*. URL: http://www. tei-c. org/release/doc/tei-p5-doc/en/Guidelines. pdf [September, 2011], 2008.
[20] E. Brill, “Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging,” *Computational linguistics*, vol. 21, no. 4, pp. 543–565, 1995.
[21] S. Tripathi and J. K. Sarkhel, “Approaches to Machine Translation,” *Annals of Library and Information Studies*, vol. 57, pp. 388–393, 2010.
[22] A. Lopez, “Statistical machine translation,” *ACM Computing Surveys*, vol. 40, no. 3, pp. 1–49, Aug. 2008.

Gmail Trienani Hariyanti <trienanihariyanti@gmail.com>

Acceptance Notice and Peer-review of A029

AIACT <aiact@smehk.org> 23 February 2018 at 19:17
To: Trienani Hariyanti <trienanihariyanti@gmail.com>

Acceptance Notification of AIACT 2018

2018 2nd International Conference on Artificial Intelligence, Automation and Control Technologies (AIACT 2018)

Osaka, Japan

April 26-29, 2018

http://www.aiact.net/

Dear Trienani Hariyanti, Saori Aida and Hiroyuki Kameda,

We are pleased to inform you that, after our double-blind peer review (please refer to the attached files), your manuscript identified below has been accepted for publication and oral presentation by 2018 2nd International Conference on Artificial Intelligence, Automation and Control Technologies (AIACT 2018) to be held in Osaka, Japan during April 26-29, 2018.

Paper ID: **A029**

Paper Title: **Samawa Language: Part of Speech Tagset and Tagged Corpus for NLP Resources**

All the registered and presented papers will be published in the volume of Journal of Physics: Conference Series, which will be indexed by EI Compendex, Scopus, Thomson Reuters (WoS), Inspec, and other databases.

Selected papers with great extension will be recommended to publish in international journals.

Yours sincerely,

AIACT 2018 Organizing Committees

Feb. 23, 2018

在2018-02-01，Trienani Hariyanti <trienanihariyanti@gmail.com> 写道:

> -----原始邮件-----
> 发件人: Trienani Hariyanti <trienanihariyanti@gmail.com>
> 发送时间: 2018年2月1日 星期四
> 收件人: AIACT <aiact@smehk.org>
> 主题: Re: Re: Submission confirmed of A029
>
> Dear Ms. Linda Lee,
>
> Thank you for your information.

RANCANGAN ANGGARAN BIAYA
ARSIP DIGITAL BASA SAMAWA

No.	Uraian Item	Volume	Unit	Harga Satuan	Jumlah
Transportasi					
1	Jepang - Sumbawa PP	4	Tiket	Rp6,500,000	Rp26,000,000
2	Sumbawa - Sumbawa Barat PP	4	orang	Rp2,500,000	Rp10,000,000
3	Pemateri PP	4	orang	Rp2,000,000	Rp8,000,000
Workshop					
1	*Stand Banner*	2	pcs	Rp300,000	Rp600,000
2	Spanduk	2	pcs	Rp350,000	Rp700,000
3	Backdrop	1	pcs	Rp300,000	Rp300,000
4	ATK	45	orang	Rp10,000	Rp450,000
5	*Snack*	45	orang	Rp10,000	Rp450,000
6	Makan Siang	50	orang	Rp30,000	Rp1,500,000
7	Suvenir Pemateri	4	orang	Rp750,000	Rp3,000,000
8	Sertifikat	50	orang	Rp20,000	Rp1,000,000
Perlengkapan					
1	Sewa Studio Berjalan	4	Kali	Rp3,000,000	Rp12,000,000
2	Sewa gedung workshop	1	gedung	Rp3,500,000	Rp3,500,000
3	*Website Engineer*	2	orang	Rp3,000,000	Rp6,000,000
Publikasi					
1	*Proofreading*	1	orang	Rp3,000,000	Rp3,000,000
2	Konferensi Internasional + publikasi	1	jurnal	Rp17,500,000	Rp17,500,000
3	Hosting Website	1	hosting	Rp3,000,000	Rp3,000,000
4	*Wordpress Tools*	1	unit	Rp3,000,000	Rp3,000,000
Total					Rp100,000,000