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

Trienani Hariyanti¹², Saori Aida¹ and Hiroyuki Kameda¹

¹ School of Computer Science, Tokyo University of Technology, 1404-1 Katakuramachi, Hachioji, Tokyo
² Department of Computer Science, Sumbawa University of Technology, Batu Alang, Sumbawa, Indonesia

E-mail: trienanihariyanti@gmail.com

Abstract. Samawa language is one of the local languages in Sumbawa Island, Indonesia. It was categorized as an under-resourced language since there is only a small number of books written and studies reported in the literature. The availability of Samawa corpus is an alternative to preserve Samawa as a cultural heritage. In this work, we describe our first effort to build the first Samawa corpus and assign 24 part of speech information to each token based on lexical and grammatical rules of Samawa. The raw data used was collected from the manuscripts, textbooks, magazines, and text from websites. It was cleaned and normalized by UTF-8 Unicode standard scheme and converted into XML format with TEI guidelines. The result is a Samawa tagged corpus of 739 sentences that contain 11,799 tokens and can be used for developing tools in many NLP applications.

1. Introduction
Samawa language is one of the local languages in Sumbawa Island, Indonesia. It is used as the second language in daily communication besides Bahasa Indonesia as the official language in most Indonesian areas. Also, it was spoken by around 500,000 speakers spread in average from western (Seteluk sub-district) to the east (Empang sub-district) in Sumbawa Island. Despite the relatively large number of speakers, Samawa was categorized as an under-resourced language where there is only small number of books in Samawa published to date, and studies have been reported in the literature. Moreover, in recent year, the subject of Samawa does not exist in curriculum’s policy in the school. It may have severe consequences for the existence and sustainability of Samawa for the next generation in the future. The availability of corpus might be an alternative to preserve Samawa as a cultural heritage and making it more accessible. A corpus represents a documentation of language in use and provides linguistic diversity. It can be useful for developing many tools in Natural Language Processing (NLP). Therefore, we initiate a new NLP research to build a Samawa corpus with its part of speech tagset.

The raw data used in Samawa corpus was collected from the manuscripts, textbooks, magazines, and text from websites. Then in the preprocessing phase, it was cleaned and normalized by Unicode in plaintext format. Regarding standard of the text corpus in digital form, the plaintext was converted to XML (eXtensible Markup Language) appropriated with Text Encoding Initiative (TEI) guidelines [1]. We also added 24 part of speech information in the corpus which consists of 11,799 tokens arranged into 739 sentences.
The rest of this paper is organized as follows. In section 2 we describe related work regarded to morphology and grammatical rules of Samawa. We generally present Samawa language and its morphology, and introduce our methods to create the corpus from preprocessing the raw text to manually tagging in section 3 and 4, respectively. In the last section, we summarise our conclusions and present future works.

2. Related Work
Several studies have already been done related to Samawa language. In 1986, Sumarsono et al. [2] proposed a general phonology, morphology of Samawa such as affixation, reduplication, composition and grammatical structure of Samawa language. Besides, they presented kind, function and the meaning of morphemes. In other research, Seken et al. in 1990 focused on Samawa morphology. They studied some of the rules that have not been described in Sumarsono [3]. In the same year, Mahsun [4] corrected the lack of previous studies in the form of separation and grouping of some affixes based on their correlation and meaning in Jereweh dialect. Mahsun also proposed a morphology of Jereweh dialect in [5] and mapped the dialects of Samawa in diachronic dialectological studies [6]. Then, Kasman in 2016 studied about the position and function about lingual form /q/, /n/, /én/, /èng/, and /s/ in Tongo dialect [7]. The last, Asmadi [8] provided morphology process in Samawa based on descriptive approach.

3. Samawa Language and its morphology
Samawa language has four dialects, i.e. Taliwang, Jereweh, Tongo, and Sumbawa Besar. One of the most critical linguistic characteristics of four dialects is the realisation of the high, front of the ultima syllabic vocals which end with the dorso-velar consonant, i.e., /s/, /t/ and /n/ [9]. Sumbawa Besar dialect was chosen as the standard of Samawa language because it has a relatively wide spread of speakers. On the other hand, the difference between each dialect is not too noticeable, lies only in the lexical and grammatical sides in some affixes [5].

Phonology system of Samawa language divided into vowels and consonants. Vowels consists of 8 phonemes such as /i/, /ɛ/, /ɛ/, /ɔ/, /o/, /ɔ/, /u/, and /a/. Consonants consists of 19 phonemes such as /p/, /b/, /t/, /d/, /ɛ/, /j/, /k/, /q/, /s/, /h/, /m/, /n/, /ŋ/, /n/, /l/, /w/, and /y/. Samawa language has three kinds of affixation, i.e., prefix, confix, and combination of them. Affixation means a morphology process of adding affixes to the base form of words. It can change the form, part of speech and the meaning of a word. Table 1 shows kinds of affixes in Samawa language.

| Prefix | Confix | Combination of Affix |
|--------|--------|----------------------|
| ba-    | bakaN- | baka-               |
| gaN-   | basaN- |                      |
| ka-    | YagaN- |                      |
| kaN-   | yaka-  |                      |
| N-     | yapa-  |                      |
| pa-    | yasaN- |                      |
| paN-   |        |                      |
| ra-    |        |                      |
| raN-   |        |                      |
| sa-    |        |                      |
| saN-   |        |                      |
| ya-    |        |                      |
4. Samawa Corpus Creation

Building Samawa corpus is a quite challenging task. First, the amount of document and textbook in Samawa language is strongly limited in size and variation since the speakers only used the Samawa language in daily communication and cultural event. Second, almost documents are particularly difficult to handle. So we have to type them manually. In the following subsection, we describe each phase of building Samawa corpus we made.

4.1. Data Collection

Raw data for corpus was collected from Sumbawa humanists, archive office of Sumbawa and language office centre of West Nusa Tenggara from the various time that the data has published. It is also essential that the data must be compiled from various domain or genre in order to represent a corpus with language variety. Table 2 shows the frequency distribution of genre in our collection.

| Genre     | Words (%) |
|-----------|-----------|
| History   | 4.66      |
| Culture   | 3.98      |
| Relationship | 4.37   |
| Religion  | 25.20     |
| General   | 61.79     |
| **Total** | **100**   |

4.2. Preprocessing

There is one main process in the preprocessing phase, i.e. text cleaning. In text cleaning process, the data was standardized into Unicode character encoding scheme of UTF-8. This encoding standard was chosen because of its flexibility and portability across platform [10]. Since some of the documents were written using original phonemes that follow phonology system above, then we normalized them with existing Latin characters. For instance, a word “Tøngøngø”, we normalized the text to become “Tøngøngø”. Later, we corrected some spelling mistakes based on Samawa dictionary in [11, 12] and deleted white spaces.

4.3. Part of Speech Tagset

Recently, there are many lists part of speech or tagset for English. Penn Treebank, Brown and CLAWS are commonly used for English tagset. A Tagset for language processing gives information about a word and its neighbours. It can be useful to build a language model [13].

Sumarsono et al. [2] classified Samawa tagset into 11 categories. These are noun, pronoun, verb, adjective, particle, number, question, coordinate conjunction, negate and modal. We used it as the base of our design to define our original tagset. Then, we analyzed the needs tagset of Samawa and proposed modification tagsets in Table 3.

| No. | Tagset | Description       | Example      |
|-----|--------|-------------------|--------------|
| 1   | CC     | Coordinate Conjunction | atawa, tapi, ke |
| 2   | CD     | Cardinal Number   | telu, satenga, saratis, peno |
| 3   | DT     | Determiner        | Datu, Dea, Lalu, Lala |
4.4. Manually Tagging of Samawa Corpus

In this subsection, we manually assigned a tagset of 24 tags to Samawa corpus based on our principle guidelines, i.e., three books of Samawa morphological and grammatical rules [2, 3, 4]. The process of annotating the corpus was divided into two major steps. First, we manually tagged 100 sentences using part of speech tagset to each token with a slash character. Each token is tagged mainly by checking its grammatical category to see if it is either word or punctuation. During this manual tagging process, we have noted the primary issue that the token have more than one tag (ambiguous). We listed, for disambiguation every rules that can be useful for tagging the next sentences with the same pattern. Moreover, it also useful as grammatical rules to develop rule-based tagger in the future.

Second, we defined and implemented a hierarchy structure of the data files of Samawa corpus based on Text Encoding Initiative (TEI) guidelines for XML. It was inspired by previous work of XML based research on Myanmar National Corpus [14]. A simple program was developed to convert
the result of manually tagging in XML. The hierarchy structure of Samawa corpus file can be seen in Figure 1.

![Hierarchy structure of Samawa corpus file.](image)

**Figure 1.** Hierarchy structure of Samawa corpus file.

Furthermore, we used the script by Jacob Perkins in [15] to see the lexical diversity of our corpus. There are 11,799 words in total and 1,674 unique words. Figure 2 shows the distribution of 10 major tags mostly used in Samawa corpus.

![Part of speech tags coverage in Samawa corpus.](image)

**Figure 2.** Part of speech tags coverage in Samawa corpus.

5. Conclusion and Future Works
This work describes our initial effort to build a new Samawa corpus. It is available in plaintext and XML format with TEI standard. We designed a tagset of 24 tags that contains the information about part of speech. We noted some grammatical rules which could be used to speeding up our manually tagging. There are, however, some limitations in our corpus such as the absence of some domain of text in Samawa language and its size of data. We wish to extend our tagged corpus and its variation using automatic tagger by collaborating with linguistics that leads to developing many tools in NLP with the corpus.
Acknowledgments
This work is supported by Tokyo University of Technology, Japan. Mrs Trienani Hariyanti would also thank to the Indonesian Endowment Fund for Education (LPDP) for gratefully acknowledge funding on her study.

References
[1] Consortium T E I 2008 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])
[2] Sumarsono, Nadera M, Basuki S, Merdhana N and Mertha N 1986 Morfologi dan Sintaksis Bahasa Sumbawa (Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan)
[3] Seken I K, Basuki S, Gosong I M and Putu Antara I G 1990 Morfologi Bahasa Sumbawa (Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan)
[4] Mahsun 1990 Morfologi Bahasa Sumbawa Dialek Jereweh (Yogyakarta: Universitas Gadjah Mada)
[5] Mahsun 2006 Kajian Morfologis Bahasa Sumbawa Dialek Jereweh (Gama Media)
[6] Mahsun 1995 Dialektologi Diachronis - Sebuah Pengantar (Gadjah Mada University Press)
[7] Kasman N F N 2017 Kedudukan dan Fungsi Satuan Lingual/q/,/n/,/è n/, /è ng/, dan /s/ dalam Bahasa Samawa Dialek Tongo Subdialek Lebangkar Kandai 12 241–254
[8] Asmadi 2016 Proses Morfologis pada Kultur Bahasa Etnis Sumbawa RETORIKA: Jurnal Ilmu Bahasa 2 214–232
[9] Mahsun 1994 Penelitian Dialek geografis bahasa Sumbawa PhD Thesis (Universitas Gadjah Mada)
[10] Chungku C, Rabbay J and Choejey P 2011 Dzongkha Text Corpus Conference on Human Language Technology for Development pp 34–38
[11] Muslim N, Nuryati, Hartini and Yudianti N M 2016 Kamus Samawa - Indonesia (Kantor Bahasa NTB)
[12] Amin U and H.M. A H 2016 Kamus Bahasa Sumbawa - Indonesia (Lembaga Adat Tana Samawa (LATS): Paboat Adat Bidang Bahasa dan Sastra Daerah)
[13] Jurafsky D and Martin J H 2000 Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (Upper Saddle River, N.J: Prentice Hall)
[14] Ko W K and Phyo T Z 2008 Selection of XML tag set for Myanmar National Corpus IJCNLP pp 33–40
[15] Perkins J 2014 Python 3 text processing with NLTK 3 cookbook: over 80 practical recipes on natural language processing techniques using Python’s NLTK 3.0 (Birmingham: Packt Publ)