REKAYASA
Journal of Science and Technology

Rekayasa, 2023; 16(1):107-112
ISSN: 0216-9495 (Print)
ISSN: 2502-5325 (Online)

The Ngoko Javanese Stemmer uses the Enhanced Confix Stripping Stemmer Method

Shevia Ilfa Melia [1], Jamiatus Sholihah [2], Dianatin Nisak [3], Intan Sukma Juniaristha [4], Ana Tsalitsatun Ni'mah [5*]
[1,2,3,4,5] Pendidikan Informatika, Universitas Trunojoyo Madura, Indonesia
Jl. Raya Telang Kamal, Bangkalan 69162
*Email: ana.tsalits@trunojoyo.ac.id

DOI: https://doi.org/10.21107/rekayasa.v16i1.19308

ABSTRACT

Stemming is vital in text processing. The stemming that is most often encountered is Indonesian and English stemming. This is because more articles are processed in text processing in English and Indonesian. Indonesia has several regional languages , especially local school content, often used in learning. Therefore, research is needed to process Javanese language texts to make it easier for education practitioners, especially in Ngoko Javanese. Ngoko Javanese stemming, which still uses the affix removal stemmers method (rule-based approach) in previous research. Has a problem, namely the lack of success of this method when returning the root words of Ngoko Javanese, so it is necessary to check the Ngoko Javanese dictionary so that the results of the root words obtained are maximized. This study aims to conduct stemmer research on Ngoko Javanese using the Enhanced Confix Stripping (ECS) method. This stemmer is designed to do word splitting according to the Enhanced Confix Stripping algorithm and through checking the dictionary adapted to the Ngoko Javanese language. The results of this study are the ability to extract essential words in Javanese Ngoko to achieve a level of truth in returning root words reaching 97%.

Keywords: Javanese Ngoko, Stemming, Enhanced Confix Stripping

INTRODUCTION

Indonesia is a country with various tribes, religions, and languages. Indonesian residents generally use regional languages when carrying out daily communication [1]. Material on Regional Languages is also found in Elementary Level Education to deepen the characteristics of the use of the language. Local language learning at elementary school sometimes uses material, questions, and answers using the local language [2]. Documents using the Regional Language will, of course, be found on the internet. Therefore, it is necessary to process regional language texts to facilitate the digitization of regional languages in Indonesia.

Javanese is one of the regional languages in Indonesia. Javanese is the language commonly used by the people of the island of Java who lives in East Java, Yogyakarta, and others [1]. Javanese is very well known in society because it is more polite and gentle in its communication compared to other languages in Indonesia. The speech level of Javanese consists of three levels Javanese Ngoko, Middle Javanese, and Javanese Krama [3]. Many people are interested in learning Javanese because of the soft delivery brought by the Javanese. But the obstacle is the unique morphology of the Javanese language.

Morphology is a process of forming words from lexemes where lexemes are lexical units and words are grammatical units [4]. The difference in the morphology of the Javanese language with Indonesian or other national languages is that it has its uniqueness. Seselan (inset), ater-ater (prefix), and penambang (suffix) are some of the components used to form Javanese words. If the Javanese root word Ngoko has received inserts, prefixes, suffixes, or a combination thereof, the term is called tembung andhahan (invented word) [8]. Tembung andhahan is a derived word composed of basic words with prefixes, infixes, and suffixes. Inserts in Javanese are called seselan consisting of; intermittent (-in, -el, -er, -um). A prefix is called ater-ater, which is divided into three types of ater-ater. Hanuswara ater (n, m, ny, ng) tripusara ater (di, ko, not), and other ater (a, pa, ma, pra, ke, ka, pi, sa, kapi, kuma, we, tar). Suffixes in Javanese called miners consist of: miners (-i, -ake, -e, -ane, -ke, -a, -en, -ana, -na, -ku, -mu) [5].

The morphology of the Javanese language has its uniqueness and difficulties that are different from Indonesian. Terms and wordings in Bahasa have their characteristics, so they have meanings and word changes from Indonesian. The stemming process is a process used to find out root words by removing the Javanese affixes ngoko [5].

Stemming is the process of changing an affixed word into a root word using specific rules [6]. Stemming algorithms have been implemented in various fields of information retrieval, such as translation, summary, and document classification. The stemming process needs to study the correct morphology of a language [7]. Morphology is the process of forming words from words, main words, and phrases [8]. Research on the stemming of several regional languages has been carried out, namely research on stemming Balinese texts using the enhanced confix algorithm [9]. This research was conducted to correct errors made by the Rule-Based Approach method in previous stemmer research in Balinese, which showed that this research could reduce errors in previous studies to 3.06%. Another regional language research is Stemmer for the Madurese language with modified enhanced confix stripping stemmer [10]. This research was conducted for Madurese language stemmers by making adjustments to the rule base according to the morphology of the Madurese language. The Javanese language research that has been done is the Javanese Ngoko stemmer with the affix removal stemmers method (Rule-Based Approach). The research was able to correctly make basic words in Javanese Ngoko reach 62%. Therefore, there is a need for research on Javanese Ngoko stemming from providing the best results for the formation of appropriate root words. This study aims to develop a Ngoko Javanese Stemmer with the Enhanced Confix Stripping Stemmer Method to improve the Ngoko Javanese stemmer to get more precise base word results.

METHOD

Figure 1 shows the research process from start to finish. This research process is the flow of what will be done during the study. The first process in this research is problem analysis. Previous research on stemming in Javanese Ngoko still used the affix removal stemmers method (rule-based approach), only being able to make root words in Javanese Ngoko with correct results reaching 62%. This stemmer's ability must be increased until it reaches a level of truth in returning root words coming 100% [8]. This research develops and modifies the hyphenation rules from the previous algorithm, namely the rule-based approach with the ECS stemmers algorithm. The next stage is a literature study. At this stage, the researcher conducted a literature review, namely, studied journal articles, literature books, and others obtained from various sources. The next stage is needs analysis. The needs analysis process is carried out by collecting the materials needed to

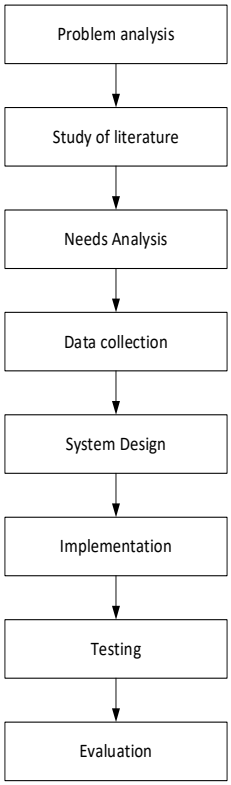

Figure 1. Research Stages

strengthen the research data, then fulfilling the needs as gathering needs as the basis for data collection. This analysis collected requirements regarding the beheading of prefixes in the Javanese Ngoko language. The next stage is data collection. This stage aims to obtain relevant and accurate data. The next stage is the general architecture for developing the Javanese Ngoko language document stemmer system. This stage has several processes that are the same as other stemming; the first stage is the input of data or documents in Javanese Ngoko.

The second stage is preprocessing. This preprocessing stage is processing raw data in the text that has yet to be structured and processed into structured text. The stages carried out in preprocessing are the first: Filtering, which is the process of removing some things that are not important, such as deleting characters, numbers, and punctuation marks. The next stage in

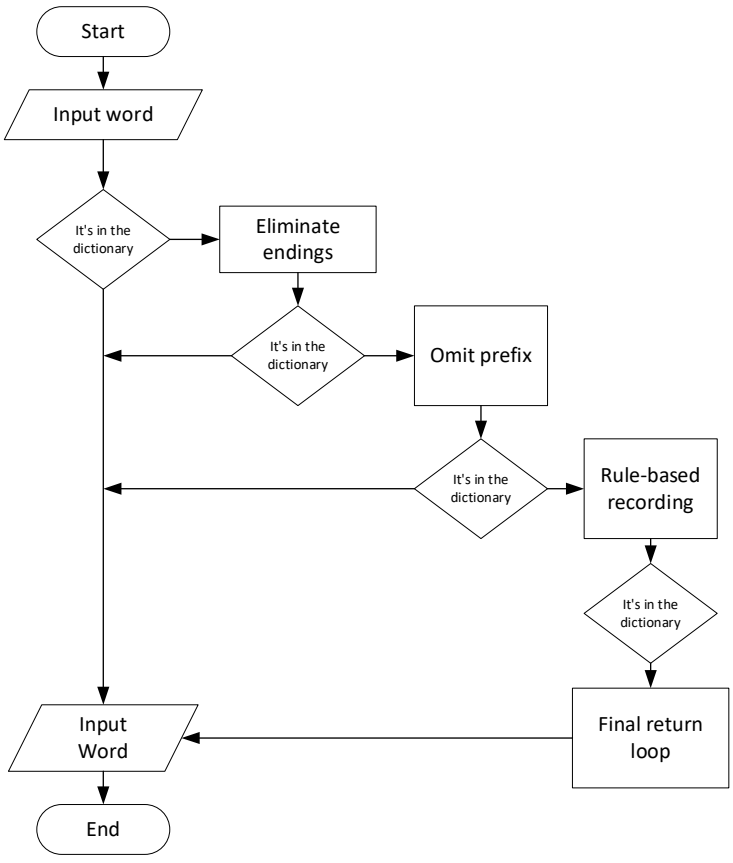

Figure 2. Flowchart of Enhanced Confix Stripping Stemmer in Javanese Ngoko

preprocessing is Case Folding. The case folding process converts the letters in the document into all lowercase characters. The next stage is Tokenization. The tokenization process aims to form a term index. The next stage after preprocessing is the Stemming process. An enhancement algorithm is a technique for processing images to make the results visually clearer than the original image. The ECS algorithm is an Indonesian stemming development algorithm that researchers often use. This algorithm has several advantages: returning endings that should not be omitted. Therefore, this study uses ECS to be applied to the Ngoko Javanese stemming process.

The stemming process of Enhanced Confix Stripping (ECS) is shown in Figure 2, namely:

a) Input word data to be processed.
b) ECS checks the word input data and matches it to the dictionary. If the input word is found in the dictionary, it is considered the root word. If not, then the process of deleting the suffix is carried out first
c) ECS removes the word ending in the word input data. After removing the suffix, continue to check the dictionary. If it is found in the dictionary, it will be stated as a base word. If not, the following process will be carried out, namely removing the prefix.
d) ECS removes the prefix. After the process of deleting the prefix then, continue checking the dictionary again. If there is a word resulting from the removal of the prefix in the dictionary, then it is returned as a base word. If it is not in the dictionary, the next process will be carried out, namely the recording process. If the process of recording the word is in the dictionary, then the process is stopped. The recording process is breaking words or rearranging words that have undergone an excessive stemming process [15].
e) If it still fails, continue with the final return process or the final return loop.
f) If it still fails, the word will be returned to its original form and considered as a root word.

The next stage is the output of the Javanese Ngoko language stemming. In the testing section, all existing datasets will be used, and corrected incorrect hyphenation in the previous algorithm. For evaluation, use the accuracy of the ECS Stemmer.

RESULTS

This study designed a stemmer to remove affixes to find the base word. This stemmer has several stages, namely tokenization, filtering, and stemming. Using the Ngoko Javanese stemmer begins with collecting documents and then entering the tokenization stage, which produces a term index, filtering, removing stopwords, then entering the stemming phase.

Interface

The results of the Java language streamer are in the form of views per word that have been processed by separating prefixes and suffixes, for example (figure 3).

Testing the results of the Ngoko Javanese Stremmer

The testing process goes through a comparison stage of the results of the

TITLE	ATER-ATER
KEYWORDS	DISEPELEKE
FREQUENCY	1
PREFIX	DI
ROOT WORD	SEPELE
SUFFIX	KE

Figure 3. Interfaces

stemming process with the Javanese Ngoko dictionary. There are three stages of deletion of affixes, namely the deletion of prefixes (ater-ater), insertions (seselan), and suffixes (miners).

a. Hanuswara Ater-Ater Result Test

Ater-ater hanuswara is a prefix in Javanese, which is included in ater-

Hanuswa ater include: m, n, ng, ny (table 1)

Table 1. After Hanuswara

m	buka	mbuka
n	tuku	nuku
ng	kanggo	nganggo
ny	susu	nyusu

b. Tripurasa Ater-Ater Results Test

Tripurasa ater are Javanese prefixes, which are included in the tripusa ater, including: dak, ko, di (table 2)

Table 2. Tripusara aters

dak	payu	dakpayu
ko	jaluk	kojaluk
di	roso	diroso

c. Test Results in Ater-Ater Liyane

Ater-ater Liyane is a Javanese prefix, which is included in the ater-ater Liyane, among others; a, ma, ka, ke, sa, pa, pi, pra, kuma, kami, kapi, tar (table 3)

Table 3. Ater-ater Liyane

a	buntut	abuntut
ma	guru	maguru
ka	jupuk	kajupuk
ke	sandhung	kesandhung
sa	kantor	sakantor
pa	emut	paemut
pi	tuduh	pituduh
pra	thanda	prathanda
kuma	lancang	kumalancang
kami	tegeng	kamitegeng
kapi	temen	kapitemen
tar	waca	tarwaca

d. Sesel Results in Test

Seselan is an insert in the Javanese language, which is included in Seselan, among others; -um, -in, -el and -er (table 4)

Table 4. Seselan

-um-	gantung	gumantung
-in-	nulis	tinulis
-el-	awu	kelawu
-er-	ogel	kerogel

e. Miner Yield Test

Miner is a suffix in Javanese, which includes miners, among others; -i, -ake, -e, -ane, -ke, -a, -ana, -na, -ku, -mu, -en (table 5)

Table 5. Penambang

nurut	-i	nuruti
sekolah	-mus	sekolahmu
deleh	-en	delehen

Ngoko Javanese Stremmer Test Results

Table 6. Results of the Java streamer test

No	Ater-ater	Results
1	Awalan/hanuswara n-, ng-, ny-, m-	true
2	Tripursa dak-	false
3	Tripurasa di-, ko-	true
4	Liyane ka, sa, ke, ma, pi, pa, pra, kami, kuma, tar	true
5	Seselan/sisipan -um, -cl, -cr, -in	true
6	Penambang -ke, -e, dan -a	true
7	Penambang -ake, -ane, -mu	true

CONCLUSION

This study uses a document dataset in the Javanese Ngoko language. This study aims to adapt the Enhanced Confix Stripping (ECS) algorithm in Javanese Ngoko to improve the quality or performance of the previous method, namely the Affix Removal Stemmers method. Which is only able to extract root words in Javanese Ngoko with a correct result of 62%, whereas, with modification using the Enhanced Confix Stripping (ECS) algorithm, the ability to extract root words in Javanese Ngoko achieves a correctness level in returning root words reaching 97%.

BIBLIOGRAPHY

[1] K. Saddhono and W. Hartanto, "Heliyon A dialect geography in Yogyakarta-Surakarta isolect in Wedi District : An examination of permutation and phonological dialectometry as an endeavor to preserve Javanese language in Indonesia," *Heliyon*, vol. 7, no. September 2020, p. e07660, 2021, doi: 10.1016/j.heliyon.2021.e07660.

[2] A. Rahman, U. Islam, and N. Alauddin, "Pengaruh Bahasa Daerah Terhadap Hasil Belajar Bahasa Indonesia Peserta Didik Kelas 1 SD INPRES Maki Kecamatan Lamba-Leda Kabupaten Manggarai," vol. 3, no. 2, pp. 71–79, 2016, doi: 10.24252/auladuna.v3i2a3.2016.

[3] Suharyo, "Nasib Bahasa Jawa dan Bahasa Indonesia dalam Pandangan dan Sikap Bahasa Generasi Muda Jawa," *NUSA J. Ilmu Bhs. dan Sastra*, vol. 13, no. 2, pp. 244–255, 2018.

[4] A. T. Ni'mah, D. Ari, and A. Z. Arifin, "Autonomy Stemmer Algorithm for Legal and Illegal Affix Detection Use Finite-State Automata Method," vol. 2, no. 1, pp. 46–55, 2019, doi: 10.25042/epi-ije.022019.09.

[5] M. Indriani, "Penanda Morfologi Bahasa Jawa Dialek Rembang," *Sutasoma J. Javanese Lit.*, vol. 3, no. 1, pp. 64–72, 2014.

[6] A. Z. Ni'mah, Ana Tsalitsatun; Arifin, "Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis," *Rekayasa J. Sci. Technol.*, vol. 13, no. 2, pp. 172–180, 2020, doi: https://doi.org/10.21107/rekayasa.v13i2.6412.

[7] A. T. Ni'mah and F. Syuhada, "Term Weighting Based Indexing Class and Indexing Short Document for Indonesian Thesis Title Classification," *J. Comput. Sci. Informatics Eng.*, vol. 6, no. 2, pp. 167–175, 2022, doi: 10.29303/jcosine.v6i2.471.

[8] P. Gede, S. Cipta, N. W. Wardani, P. T. Informatika, and R. B. Approach, "Stemming Dokumen Teks Bahasa Bali Dengan Metode Rule Base Approach," vol. 7, no. 3, pp. 510–521, 2020.

[9] N. W. Wardani, P. Gede, and S. Cipta, "Stemming Teks Bahasa Bali dengan Algoritma Enhanced Confix Stripping," vol. 4, pp. 103–113, 2020.

[10] R. Maulidi, "Modifikasi Metode Enhanced Confix Stripping," *Pros. Semin. Nas. FDI*, no. December, pp. 12–15, 2016.