Separation of Basic Words in Angkola Batak Text Documents using Enhanced Confix Stripping Stemmer
Case: Mandailing Ethnic

M A Muchtar1*, I Jaya1, Manguji Nababan2, U Andayani1, Lisa Noprianti Siregar1, Erna B Nababan1, Opim S Sitompul1

1Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
2Center of Documentation and Assessment of Batak Culture, Universitas HKBP Nommensen, Medan, Indonesia

*Correspondence Email: anggi.muchtar@usu.ac.id

Abstract. Stemming is the process of separating essential words from an affixed word. Stemming works by eliminating morphological variations which attached to a word by removing affixes in a word with a dictionary as a reference to the stemming process. One effective algorithm for resolving stemming in Indonesian is an Enhanced Confix Stripping Stemmer algorithm. In this case, the improved Confix stripping stemmer algorithm is implemented for the separation of essential words in the Angola-Mandailing Batak language document that has different phonologies and morphologies with Indonesian. The documents used are Latin character documents in the Angkola-Mandailing language that goes through the stages of filtering, folding cases, and tokenization before stemming. The stemming process is done by removing particles, ownership, suffixes, and prefixes. After testing this research, it was concluded that this algorithm was able to separate essential words in the Angkola-Mandailing Batak language with accuracy 87.05%.

1. Introduction
Indonesia is a country consisting of many tribes and languages. Based on the 2010 Central Bureau of Statistics census results, there are 1340 tribes and 2500 regional languages in Indonesia. The Batak tribe is one of the three tribes with the largest population in Indonesia. The Batak tribe itself is divided into five sub-groups, namely: Toba Batak, Karo Batak, Pakpak Batak, Simalungun Batak, and Angkola-Mandailing Batak. The Angkola-Mandailing Batak Population is a community living in the Mandailing Natal area, South Tapanuli, Padang Lawas, and Padang Sidimpuan. The language used in people's daily lives in this region is the Angkola Batak language - Mandailing. Angkola Batak Language - Mandailing itself has its own rules in pronunciation, sentence structure, and grammar. Grammar itself includes essential words and affixed words in a sentence.

One process for transforming essential words from words in a document is the process of stemming [2]. Stemming is a process or way of finding the essential words of Scripture. Stemming itself serves to eliminate the morphological variations that are attached to a word by removing the affixes to the word. The stemming process for each language is different because each style has different grammar rules in the use of affixed words [6]. Example sentence in Indonesian sentence "Saya berlari ke sekolah". The word that needs to be stemmed in the sentence is the word "berlari" because in the word "berlari", there
is a ber-/ affix. The stemming process is done to get the word "lari" from the word "berlari". If the phrase "Saya berlari ke sekolah" is translated into the Angkola Batak language - Mandailing becomes the phrase "au marlojong tu sikola". The word that needs to be stemmed in the sentence is the word "marlojong" because the word contains the prefix mar-. The stemming process is done to separate the word "lojong" from the word "marlojong."

One algorithm for stemming is an enhanced Confix stripping stemmer (ECS) algorithm. The stemming algorithm that was used for the first time to stimulate Indonesian was the Nazief-Adriani algorithm [1]. The enhanced Confix stripping stemmer (ECS) algorithm is an algorithm developed by Mahendra to correct the lack of the Confix Stripping Stemmer algorithm [6]. This algorithm previously used in Tahitoe research [8]. Furthermore, to solve the stemming of Indonesian absorption words [4] or the improvement of summarization by using TextTeaser[9].

Additionally, Maulidi completed the stemming of the Madura language by implementing a modified, enhanced Confix stripping stemmer (ECS) algorithm [7]. In previous research, stemming was done for Indonesian. Research for regional languages is still very little. Research for deriving local languages was previously conducted for Madurese language stemming. Based on this, the author is compelled to implement an enhanced Confix stripping stemmer (ECS) algorithm for stemming processes in the Angkola Batak language - Mandailing in obtaining the exact base words of a sentence that is affixed by doing the right prefix and suffix. The proper stemming process for the separation of essential words from the sentences in a text document is Angkola-Mandailing Batak Language as in Indonesian.

In 2012, Ayyidar et al. implemented an enhanced Confix stripping stemmer for absorption in Indonesian [4]. Enhanced Confix stripping stemmer (ECS) algorithms can well complete the stemming process for absorption words by adding additive rules but require a longer time due to the addition of stemming rules to overcome absorption words in Indonesian. Furthermore, a study was conducted to compare the stemming process using the Enhanced Confix Stripping Stemmer algorithm and the Porter stemmer algorithm. The study was conducted by testing 300 documents with the results: Enhanced Confix Stripping Stemmer produced 81% accuracy in 11.8311 seconds and a stemmer porter with 61% accuracy in 0.0291 seconds.

Furthermore, in 2013, Iqramitha Annifa applied the enhanced Confix stripping stemmer algorithm to the stemming process for the dictionary of Indonesian language cues with results [5]. The enhanced Confix stripping stemmer algorithm can be applied in a sign language learning system. This algorithm is able to chop words with a combination of Prefix 1 + Prefix 2 + Basic Words + Suffix 3 + Suffix 2 + Suffix 1 where suffix 3 = suffix, suffix 2 = pronoun belongs, and suffix 1 = particle. However, the stemming process cannot be carried out because in Indonesian sign language there are no rules that explain the insertion while for the homonym word in sign language there is no difference in form, like snake can with what I can do. Subsequent research in 2016, conducted a study to complete the stemming process in the Madurese language by modifying the rule base on Enhanced Confix Stripping Stemmer according to the morphology of the Madura language and getting good results [3].

2. Methodology

The system was built to complete the stemming process of the Angkola batak language document - Mandailing that had never been made before. The system will process stemming from a document and then decompose it into word words which are then stemmed using an enhanced Confix stripping stemmer (ECS) algorithm. The general architecture carried out in this study is shown in Figure 1. The description are as follows:

2.1 Input

In the first stage, the system is inputting documents with the extension .docx. As for documents in the form of folk tales or oral transcripts of the traditional ceremony of the Angkola Batak language - Mandailing.
2.2 Pre-processing

At this stage, the process of changing the form of unstructured data into structured data is carried out according to system requirements. There are three stages to be carried out, namely:

- **Filtering**
  This process aims to take the alphabet character of the document. In this process, the numbers, punctuation, and characters are removed from the letters of the alphabet.

- **Case Folding**
  Case folding is the process of converting all letters in a document to lowercase. Only the letters 'a' up to 'z' are accepted. Non-letter characters are omitted.

- **Tokenization**
  This process aims to separate each word that composes a document into words per word by marking a space character. Each word, the tokenization result will be saved in the array.

![Figure 1. General Architecture](image)

2.3 Processing

At this stage, a stemming process is carried out with an enhanced Confix stripping stemmer algorithm. One by one, the word will be checked, whether it is in the dictionary or not. If the word is included in the dictionary of the basic words, the stemming process for the word will not be confirmed. Then proceed to the next word, if the word is not in the basic word dictionary, a stemming process will be carried out with an enhanced Confix stripping stemmer algorithm. The process of stemming this algorithm has the following stages, namely:

a. Check the dictionary. If the word is in the dictionary, the basic word has been found.
b. Remove inflectional suffixes (particles) (-do, -ma, -pe)
c. Remove possessive pronoun (ownership) (you, me, -na)
d. Check the dictionary
e. Remove the derivation prefix (prefix) (mar- /, ma-/, tar-/, pa-, di-/, paN-/, par-/, saN-/, um-/, si-/)f. Delete infix
g. Check the dictionary
h. Remove the derivation suffixes (khiran) (-an, -on, -kon, -hon, -i)
i. Check the dictionary
j. Replenishment returns

The stemming process for the Angkola-Mandailing Batak language refers to the rule table 1:

| Rule | Word Format | Syllabification |
|------|-------------|-----------------|
| 1    | marV        | mar-V           |
| 2    | m-arC       | mar-C           |
| 3    | mam{p|b}    | mar-{p|b}       |
| 4    | man{ji|di|c} | man-{ji|di|c}   |
| 5    | mang{kl|g}  | mang-{kl|g}     |
| 6    | mang{1|ų}   | manga-{1|ų}     |
| 7    | many{s}     | ma-ny{s}        |
| 8    | man{t}      | ma-{t}          |
| 9    | tarV        | tar-V           |
| 10   | tarC        | tar-C           |
| 11   | diV         | di-V            |
| 12   | diC         | di-C            |
| 13   | paC         | pa-C            |
| 14   | paV         | pa-V            |
| 15   | pang{h}     | pang-{h}        |
| 16   | pan{cl|dl|l|t} | pan-{cl|dl|l|t} |
| 17   | parV        | par-V           |
| 18   | parC        | par-C           |
| 19   | saV         | sa-V            |
| 20   | saC         | sa-C            |
| 21   | siC         | si-C            |
| 22   | sal{l}      | sal-{l}         |
| 23   | pang{h}     | pa-ng{h}        |

2.4 Output

After all the words in the document have been processed, the results of the stemming process will be displayed in the form of files with the extension .docx that can be stored on the user's computer. All words in the original document will be displayed, and the words that are stemmed will be separated by prefix, essential words, and endings with space characters.

3. Results and Discussion

System testing is done to find out whether the system has been built functions properly and runs as desired. This test will explain how much the accuracy of the implementation of Enhanced Confix Stripping Stemmer is applied to the separation of essential words in text documents Angkola Batak Language - Mandailing. The transliteration system interface can be seen in Figure 2.

On the transliteration page, there is the Select File button to enter the text document file in the Angkola-Mandailing Batak Language. The user will wait a few seconds until the document entered is displayed by the system. After the document is displayed, then the user selects the Stemming and Waiting button until the system displayed the stemming results. The last step is to choose the Save button to save the transliteration results in the form of files with the extension .doc. This stemming process can be seen in Figures 3, 4, and 5.
The test results on this system will use the following formula 1:

$$\text{Accuracy} = \frac{\text{Total Successful stemmed words}}{\text{Total stemmed words}} \times 100\%$$  \hspace{1cm} (1)

The calculation of the accuracy of the test results for each script document can be seen in table 2.

| Input Documents | Words in Documents | Stemmed Words | Successes | Accuracy  |
|-----------------|--------------------|---------------|------------|-----------|
| Halilian 1      | 246                | 24            | 22         | 91%       |
| Halilian 2      | 195                | 22            | 19         | 86%       |
| Halilian 3      | 225                | 38            | 35         | 92%       |
| Halilian 4      | 228                | 26            | 22         | 84%       |
| Halilian 6      | 230                | 34            | 30         | 88%       |
| Halilian 7      | 238                | 26            | 20         | 76%       |
| Halilian 12     | 474                | 55            | 41         | 74%       |
4. Conclusion

From the results of testing on the stemming system of the Angkola-Mandailing Batak language text document, outcomes are obtained as follows.

- Stemming the Angkola-Mandailing Batak language with the enhanced Confix stripping stemmer algorithm for the Angkola-Mandailing Batak language obtained an accuracy of 87.08%.
- The system cannot resolve to stem for decays, which are given the prefix t and p.
- The system cannot resolve to stem for words with the prefix manga-/ and panga-/.
- Suggestions that can be given by the author for the development of further research are as follows:
  - Further research is expected to be able to overcome the shortcomings in this study, namely the prefix that decays and the prefix manga// and panga/.
  - Further research is expected to be able to overcome more than one word of additive.
  - The next research is expected to be able to overcome the addition of letter characters to words ending in vowels and affixed with the word ownership.
  - Further research is expected to be able to process the Angkola-Mandailing Batak language with other algorithms to improve accuracy.
  - Stemming data can be developed to do translations into other languages.

Acknowledgments

This research was supported by the Universitas Sumatera Utara. All lecturers and staff members of the Information Technology Study Program. Mr. Manguji Nababan as lecturer and Pusat Dokumentasi Pengkajian Kebudayaan Batak (PD & PKB) Universitas Nommensen Medan who was willing to provide research data to the Author.

References

[1] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. ACM Transactions on Asian Language Information Processing (TALIP) Hal 1-33.
[2] Agusta L. 2009. Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief &Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. Proceeding Konferensi Nasional Sistem dan Informatika 2009, pp. 09-036.
[3] Alif M, Solihin F, and Husni H. (2013). Perbandingan Metode Enhanced Confix Stripping dan Porter Stemmer Untuk Stemming Konten Bahasa Indonesia. *Jurnal Buana Informatika* Vol. 4 No.2

[4] Ayyidar, N. H., Syukur, A., & Subagyo, H. Algoritma Stemming Untuk Kata Serapan Bahasa Indonesia. Jurnal Teknologi Informasi, Volume 8 Nomor 1, April 2012, ISSN 1414-9999

[5] Iqramita, Annifa. 2013. Implementasi Algoritma Enhanced Confix Stripping Stemmer Pada Kamus Sistem Bahasa Isyarat Indonesia. Skripsi. Universitas Sumatera Utara.

[6] Mahendra, I. P. A. K., Arifin, A. Z., dan Ciptaningtyas, H. T. 2008. Penggunaan Algoritma Semut dan Confix Stripping Stemmer Untuk Klasifikasi Dokumen Berita Berbahasa Indonesia. Skripsi. Institut Teknologi Sepuluh Nopember.

[7] Maulidi, Rakhmad. 2016. Stemmer Untuk Bahasa Madura Dengan Modifikasi Metode Enhanced Confix Stripping Stemmer. Prosiding Seminar Nasional FDI 2016 Hal 12–15. ISSN. 2460-5271

[8] Tahitoe, A. D. dan Purwitasari, D. 2010. Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia Dengan Metode Corpus Based Stemming. Tesis. Institut Teknologi Sepuluh Nopember

[9] Gunawan D, Pasaribu A, Rahmat R F, Budiarto R 2017 Automatic Text Summarization for Indonesian Language using TextTeaser *IOP Conference Series : Material Science and Engineering* 109(1) 012048