NAZIEF AND ADRIANI'S STEMMING ALGORITHM IMPLEMENTATION ON INDONESIAN SCIENTIFIC WRITING ERROR IDENTIFICATION AND CORRECTION SOFTWARE

Sunda Ariana[1)], Hadi Syaputra[2)], Margareta Andriani[3)], Suheriyatmono[4)]

[1)]Language and Literature Faculty, BinaDarma University

[2)]Computer Science Faculty, Bina Darma University

[3)]Teacher Training and Education Faculty, Bina Darma University

E-mail: [1)]sunda@binadarma.ac.id, [2)]hadisyaputra@binadarma.ac.id, [3)]m.andriani@binadarma.ac.id

Abstract: Scientific writings in Indonesian language must follow the rules of the Perpected Spelling System (PSS) of Indonesian language especially in terms of writing derivative or affixed words as these kinds of words might contain prefixes, infixes, and suffixes. Therefore, there is a need of a softwarefor identifying and correcting errors in accordance with the Perpected Spelling System of Indonesian language to help the writers when writing scientific papers. In the process of identifying and correcting derivativeor affixed words, the first thing to do is identifying the errors of the root words. The root words found will then be verifiedwith a corpus of basic root words. If the root words arein accordance with the corpus, the next step is to check whether or not the writing or affixed words matches the PSS. If the writing of the affixed words does not match the PSS, the software will mark the errors and suggest corrections. The algorithm used for finding root words in affixed wordsis Nazief and Adriani's stemming algorithm**.**

Keywords**:** software, Indonesian language, scientific writing, Nazief and Adriani's stemming algorithm

I. Introduction

Bahasa Indonesia is the offficial language in education in Indonesia. Article 36 of 1945 Constitution states that *Bahasa Indonesia* functions as the official language, the language of instructions in educational institutions, means of communication in running the government and development throughout the nation, means of cultural development and use of science, arts, and modern technology. Unfortunately, not all of its users use *Bahasa Indonesia* properly.

Indonesian people, in general, master at least two languages; their mothertongue and *Bahasa Indonesia*. The mothertongue is acquired informally in family environment, while *Bahasa Indonesia* is learned and acquired in formal environments, e.g. in schools. This often causes a language interference. Language interference is the errors in the language use resulting from the use of two languages alternately. Language interference also occurs in scientifc writings that should be written in good structures and grammar especially when it comes to writing derivative and affixed words.

The thing explained above turns out to be the cause of writing errors that commonly occur in Indonesian scientific papers. Meanwhile, when writing scientific papers, a writer should follow the standardized writing rules. Andriani (2007) conducted a researchon misspellings commited by the students in writings. The result showed that the students still made mistakes in PSS, i.e. in lettering, wording, using punctuation, and using loanwords. Ariana's (2011) result of the research even showed that erros of *BahasaIndonesia* use also appeared on papers written by lecturers.

A Indonesian lecturer should understand, comprehend, and master the use of structure and grammar and PSS of *Bahasa Indonesia* especially when he/she is writing a scientific paper or when supervising students in writing academic papers. This, however, also make the lecturers spend more times on checking the grammar, often neglecting the content. Meanwhile, in scientific papers, the contenct and the metodhology should come first. Therefore, based on the things explained above, the objective of this research is to build a software for detecting and correcting errors of derivative and affixed words.

II. PROBLEM IDENTIFICATION

Through this research, it will be built a software which can automatically detect and correct use of root words and affixed words of Bahasa Indonesia in scientific papers. The problem of this research is how to identify and correct writing errors of scientific papers in *Bahasa Indonesia*.

In order that this paper does not come out of the subject matter defined, the scope of the discussion is limited to the algorithm used in detecting errors of derivative or affixed words based on Perfected Spelling System (PSS).The algorithm used is Nazief and Adriani stemming algorithm. In designing the software, the researchers used C # programming language and Microsoft SQL Server databases. Documents used in the process of derivate and affixed word error detection are those in (.doc) or (.docx) type.

From the above problems, the objective of this research is to build a software for identifying and correcting errors of writing in scientific papers in Bahasa Indonesia by using stemming algorithm of Nazief and Adriani so that writing errors can be detected quickly and corrected in accordance with the PSS.

This research is expected to provide the efficiency of time and work efficiency for researchers to detect errors in their research or scientific writings in accordance with the PSS.

III. PREVIOUS RELATED STUDIES

Stemming algorithms for multiple languages have been developed, such as Nazief&Adriani’s algorithm for Indonesian-language text [4]. The algorithm created by Bobby Nazief and MirnaAdriani has the following stages:

1. Find a word that will be stemmed in the dictionary. If it is found then it is assumed that it is the root word word. Then the algorithm will stop.

2. Inflection Suffixes ("- lah", "-kah", "ku", "mu" or "nya") are removed. If in the form of particles ("- lah", "-kah", "-tah" or "-pun") then this step is repeated to remove possesive pronouns ("- ku," "mu," or "nya") , If any.
3. Remove Derivation Suffixes ("- i", "an" or "-kan"). If the word is found in the dictionary, then the algorithm stops. If not then it will proceed to step 3a.
a. If the "-an" has been removed and the last letter of the word is "-k", hence the "-k" is also deleted. If the word is found in the dictionary then the algorithm stops. If not found then it will go to step 3b.
b. Suffix removed ("-i", "-an" or "-kan") is returned, go to step 4.
4. Delete Derivation Prefix. If in step 3 the suffix is removed then go to step 4a, if not go to step 4b.
a. Check the table of prefix-suffix combination that is not permitted. If found then the algorithm stops, otherwise go to step 4b.
b. For i = 1 to 3, specify the type of the prefix and then remove the prefix. If the root word has not been found proceed to steps 5, if so, then the algorithm stops. Note: if the second prefix is the same as the first prefix, the algorithm stops.
5. Recoding.
6. If all steps have been completed but were not successful then the initial word is assumed to be the root word. The process is complete.

The type of prefix is determined through the following steps:

1. If the prefix is: "di-", "ke-", or "se-" then the type of the prefix in a row is "di-", "ke-", or "se-"
2. If the prefix is "te-", "me", "be-" or "pe-" then it takes an additional process to determine the type of the prefix.
3. If the first two characters are not "di-", "ke-", "se-", "te-", "be", "me-", or "pe" then stop.
4. If the type of the prefix is "none" then stop. If the type of the prefix is not "none", the prefix can be seen in Table 2. Remove the prefix if found.

Tabel 1.Unallowed Prefix-Suffix Combination

Prefix	Uallowed Suffix
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se-	-i, -kan

Tabel 2. Techniques for Determining Types of Prefix for Words with Prefix “te-“

Following Characters				Prefix Type
Set 1	Set 2	Set 3	Set 4	
“-r-“	“-r-“	-	-	none
“-r-“	Vowel	-	-	ter-assimilated
“-r-“	not (vowel or “-r-”)	“-er-“	Vowel	ter
“-r-“	not (vowel or “-r-”)	“-er-“	Not Vowel	ter-
“-r-“	not (vowel or “-r-”)	not “-er-“	-	ter
not (vowel or “-r-”)	“-er-“	Vowel	-	none
not (vowel or “-r-”)	“-er-“	Not Vowel	-	te

Tabel 3.Prefixes Based on Their Types

Prefix	Prefix that should be deleted
di-	di-
ke-	ke-
se-	se-
te-	te-
ter-	ter-
ter-assimilated	Ter

To overcome the limitedness in the above algorithm, the following rules apply

1. Rules for Reduplication

- If the two words are linked by connecting words are the same words, the root word is the singular one, eg "booksthe root word is the "book".

1. Additional Rules for Prefix and Suffix Forms.
   - For prefix “mem-“, words with prefix “memp-” has “mem-” prefix type.
   - For prefix “meng-“, words with prefix “mengk-” has “meng-” prefix type.

IV. RESEARCH METHODOLOGY

a. Software Designing

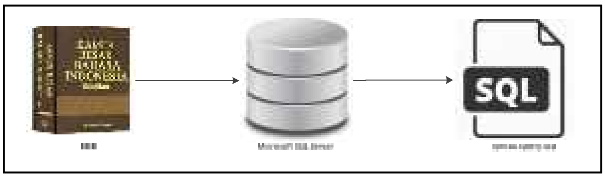


Image 1. Corpus Collection Process

b. Software design is based on creating a standard database that contains a collection of root words from the Great Dictionary of Indonesian language (*KBB*I) that is stored on a Microsoft SQL Server database system.

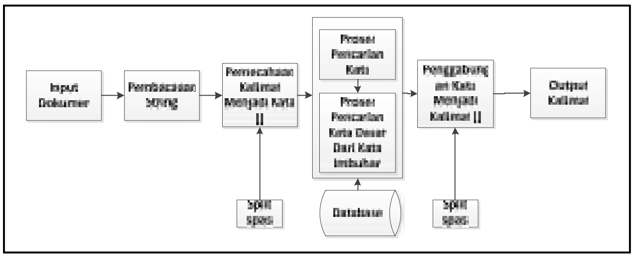

Input Dokumer
Pembacaan String
Pemecahaan Kalimat Menjadi Kata []
Proses Pencarian Kata
Proses Pencarian Kata Dasar Dari Kata Imbuhan
Penggabung an Kata Menjadi Kalimat []
Output Kalimat
Split spasi
Database
Split spasi

Image 2. Sentence Indetifying and Correcting Sentences

a. Sentence to words splitting process Flowchart

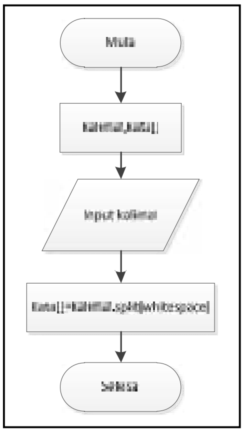


Image3. Sentence to word Flowchart

b. Nazief and Adriani’s Stemming Algorithm Flowchart

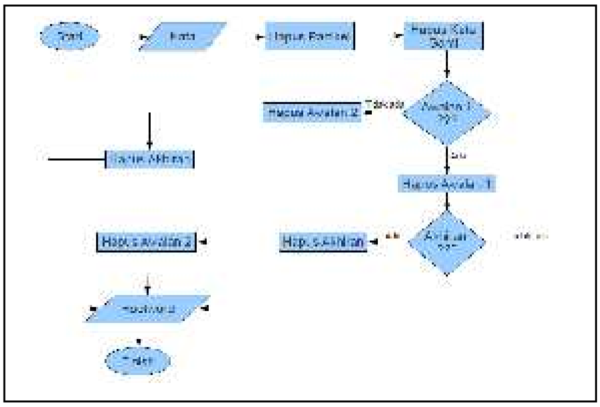
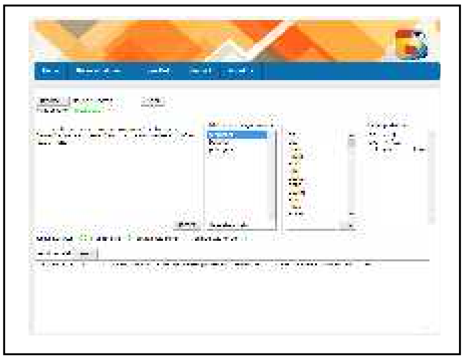


Image 4. Word Stemming Flowchart

V. RESULTS

Gambar 5. Software form



Gambar 6. Result of correction form

This software provides a function to open the (.doc) or (.docx) document type then do the process of splitting the sentence into words. The word derived from the breakdown of the sentence is verivied with the corpus of data, if the word is in acoordance with the corpus of data then the word will be made right standing and words that are not in accordance with the corpus will be processed to see if the word is affixed words or not and will be given status as a wrong word. Words detected affixed words will be processed with Nazief and Adriani’s stemming algorithm. If the results of these will generate root words that will be verified with the corpus of the data. If the word is not listed in the corpus of data, then it is marked wrong.

VI. CONCLUSION

From the process of the software implementation and testing, the researchers conclude that:

1. The software is made to detect errors that occur in the documents of scientific papers.

2. This software can display options for improvements to words that are not in accordance with the corpus ofdata and PSS.

REFERENCES

[1] Andriani, Margareta. 2007. AnalisisKesalahanEjaankaryaIlmiah: StudiKasusMahasiswaNonbahasa2007/2008 UniversitasBinaDarmaPalembang.*DalamJurnalBinaEdukasi*,vol.1 No.1juni 2008

[2] Ariana, Sunda. 2011. KesalahanPenggunaanEjaan yang DisempurnakandalamKaryaIlmiahDosenUniversitasBinaDarma.*DalamJurnalBinaEdukasi*, vol.5 No.2 Desember

[3] Ariana, Sunda, dkk..2012.PrototipePerangkatLunakKesalahanBerbahasauntuk MeningkatkanKualitasPenulisanKaryaIlmiah.*DalamProsiding Seminar NasionalTeknologiInformasidanKomunikasiTerapan, ISBN: 979-26-0276-3.* Semarang 2014

[4] Tarigan, Henry Guntur. 2011. *PengajaranPemerolehanBahasa*.Bandung: Angkasa

[5] Nazief, Bobby dan Mirna Adriani, Confix-Stripping: Approach to Stemming Algorithm for Bahasa Indonesia, Fakulty of Computer Science University of Indonesia.