LONTAR KOMPUTER VOL. 12, NO. 2 AUGUST 2021 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2021.v12.i02.p02 **e-ISSN 2541-5832**
Accredited Sinta 2 by RISTEKDIKTI Decree No. 30/E/KPT/2018

KEBI 1.0: Indonesian Spelling Error Detection System for Scientific Papers using Dictionary Lookup and Peter Norvig Spelling Corrector

Tresna Maulana Fahrudin[1], Ilmatus Sa'diyah[2], Latipah[3],
Ibnu Zahy' Atha Illah[4], Cagiva Chaedar Bey Lirna[5], Burhan Syarif Acarya[6]

[1,2,4,5,6]Department of Data Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Raya Rungkut Madya Street, Gunung Anyar, Surabaya, Indonesia
[1]tresna.maulana.ds (Corresponding author)
[2]ilmatus.sisfo@upnjatim.ac.id

[3]Department of Informatics Engineering, Universitas Narotama
Arief Rahman Hakim 51 Street, Surabaya, Indonesia

Abstract

Many Indonesian spelling errors occur in research papers published to the public, closely related to academics in all institutions such as research institutions, government, schools, and universities. The spelling errors usually writing punctuation, writing letters, writing words, writing words originating from foreign or regional languages (uptake words), using affixed words, and writing ineffective sentences. The mistakes made by the academics then become a cycle in the academic environment. They usually provide guidance for writing an undergraduate thesis, thesis, dissertations to students, or the other forms of documents and scientific papers. Therefore, the research proposed the application to facilitate all authors of scientific papers in producing quality scientific works based on the General Guidelines for Indonesian Spelling published by the Agency for Development and Language Development. The application is named KEBI 1.0 Checker (Indonesian Spelling Error 1.0 Checker), a web-based application with a built-in algorithm to detect and correct Indonesian Spelling in scientific papers. The experiment result shows that the application has given the best accuracy performance to correct the non-standard words, and typographical errors reached 100% and 55,52%, respectively. The application also has been detected 209 meaningless words. The application processing time is relatively low, the average time needed to correct non-standard words is 0.016 seconds, and typo words are 14.58 seconds. KEBI 1.0 Checker is helpful for the end-user in academics but needs to improve the vocabulary of the large corpus in various fields of science for correcting typo words.

Keywords: *KEBI 1.0 Checker, Spelling Error, Scientific Paper, Dictionary Lookup, Peter Norvig Spelling Corrector*

1. Introduction

Writing in Indonesian is closely related to the academic field in all institutions such as research institutions, government, schools, and universities. Unfortunately, this field is not supported by understanding good and correct Indonesian Spelling in writing, especially writing scientific papers. Therefore, many Indonesian spelling errors still occur in language written works published to the public.

Even in the academic field at universities, lecturers still made many spelling errors in Indonesian [1]. The errors include all elements of language, namely writing punctuation, writing letters, writing words, writing words originating from foreign or regional languages (uptake words), using affixed words, and writing ineffective sentences. This fact shows that the quality of Indonesian writing at the university level is very concerning. Whereas the standard variety of Indonesian language is used by educated people in formal education and is also learned from formal education [2]. The mistakes made by the lecturer then become a cycle in the academic environment because the

lecturer needs to provide guidance on writing undergraduate thesis, thesis, and dissertations to students. Finally, the students also made the same spelling mistakes in Indonesian.

Murtiningsih [3] stated that as many as 69.2% of students made mistakes in writing words that were not in the context of writing, and 17.4% of students made mistakes in writing affixes and prepositions. This error is called a typographical error [4] and is often caused by students when writing a paper or other written work. Students also need to make repeated revisions related to Indonesian Spelling, print out the corrections and report back to their lecturers. This makes the use of paper for printing increase depending on the number of revisions made by students.

In this case, several other factors cause students to make mistakes in spelling Indonesian [4]. First, the lack of understanding of students in using good and correct Indonesian Spelling, the most errors are in writing punctuation and capital letters. Second, the lack of understanding of the standard variety vocabulary and the vocabulary of the scientific field that students have is related to the student's habit of reading relevant reference activities. Third, the lack of students' ability to combine their ideas with other ideas from references that have been read.

The above opinion was also confirmed by [5] , who stated that the Indonesian spelling errors that occurred could be in the form of intentional or unintentional mistakes. Spelling errors in scientific writing are primarily determined by the level of understanding of Indonesian spelling principles. There are several factors that the effect spelling errors, such as typing error, lack of knowledge in writing standard language and punctuation, the habit of using non-formal daily language. All of these factors have a direct impact on the quality of scientific work produced by the authors.

Applications for spelling correction in English scientific writings can be corrected using a *Grammarly checker*. The application will provide a sign, even suggesting the right replacement word if there is a spelling error in a sentence or paragraph [6]. However, there are not many spelling correction applications in Indonesian scientific writing, and some other available applications have limited features. For example, Ejaan.id only provides non-standard word detection, Typoonline.com only provides typo word detection, and other applications are limited to detection without giving suggestions for correction.

Therefore, this research took the initiative to facilitate all authors of scientific papers in producing quality scientific works based on the General Guidelines for Indonesian Spelling published by the Agency for Development and Language Development. The application is named KEBI 1.0 Checker (Indonesian Spelling Error 1.0 Checker), a web-based application with a built-in algorithm to detect and correct Indonesian Spelling in scientific papers.

The KEBI 1.0 Checker framework features will be developed periodically to enrich its ability to detect spelling errors in Indonesian. This research study proposed and discussed KEBI 1.0 Checker that provides several features in one platform that detect non-standard Indonesian word-based and typographical errors in a scientific paper. The application is expected to have a positive impact on the editing process. The work of editing spelling errors in Indonesian scientific work has also become easier because the editing work takes a long time [7]. In addition, the application can also facilitate the work of editors at publishers in carrying out their work. The publishing world can also be more active and productive, especially in the field of self-publishing.

2. Related Works

This section will discuss several other studies that have been carried out by the other researcher regarding the detection of spelling errors for text documents in the Indonesian language. Mawardi et al. introduced a spelling correction application for text documents in Bahasa Indonesia using Finite State Automata (FSA), Levenshtein distance, and N-gram methods. They have been stated that one of the causes of misinformation due to an error in writing a document. To overcome the issue, a system is needed to make spelling corrections. The text document that has been used as input is an Indonesian language document, and the output is a *.txt file. They tested 5000 news articles in their study. The study reported the smallest perplexity value is unigram with a value of 1.14. Bigram has a higher correction rate than unigram and trigram, with adjoined word error of 75%, while non-adjoined word error for bigram and trigram has the same correction level of 71.20%. The percentage of the total false positive rate for unigram, bigram, and trigram is the

same percentage of 4.15%. The study also indicates that using the FSA Method can shorten the processing time of spelling corrections [8].

Irmawati et al. proposed a method to correct preposition errors in Indonesian sentences written by second language learners. Their method includes two components: training word embeddings using large native sentences and copying preposition errors from learners' sentences to native sentences. The proposed method calculates a syntactic similarity score between the native sentence and the learners' phrase before copying a preposition mistake from a learner sentence to a native sentence. The learner sentence's preposition mistake with the highest syntactic similarity score to the native sentence is then chosen to replace the original preposition in the native sentence. On the same training data size, depWE-Head has higher precision, recall, and F1 Score than CSRnd and depWE-HeadObj. The results show that their method outperforms the correction model trained on the similar size of native data [9].

Gunawan et al. carried out the microtext normalization process to convert these features into well-written text or, in other words, convert non-standard words into standard words according to the Indonesian dictionary and the Longest Common Subsequences (LCS) algorithm. They stated that social media had become a significant need for communication. Short message services (SMS), Facebook statuses, Twitter posts, chat messages, and comments are mediums for expressing ideas in short texts. However, these media limit the space for writing characters to be shorter than a sentence. The writing is informal and includes acronyms, pictograms, and hashtags, or what is known as microtext. These attributes can cause inaccuracies and cannot be directly processed like traditional word processors. The amount of data tested is 400 tweets. For normalizing the Indonesian-language Twitter text, they get an accuracy value of 82% for the application of Dictionary Based, 74% for the application of the Longest Common Subsequences Algorithm only, and 84% for the combined application of the Longest Common Subsequences and Dictionary Based Algorithm so that it can be concluded that the implementation of the Dictionary based is proven help the normalization process and maximize the normalization results in the Longest Common Subsequences Algorithm [10].

According to the related works above, it can be concluded that spelling error detection can be applied to Indonesian text documents, correcting prepositions in second language learner writing, and normalizing microtext from messages, posts, and comments from social media. In contrast, the most widely used methods include Finite State Automata (FSA), Levenshtein distance and N-gram methods, syntactic similarity score, dictionary-based, and Longest Common Subsequences (LCS) algorithm. We have not seen their research on whether it has fallen into detecting spelling errors in Indonesian scientific papers. Therefore, our research tries to contribute to the development of web-based applications with non-standard and typographical error detection features in one platform that users can access directly. This application will be developed periodically both from the features offered and the corpus it has.

3. Research Methods

The proposed system design of KEBI 1.0 research is illustrated in Figure 1, consisting of 4 processes, namely preprocessing, error spelling detection, error spelling correction, and the final result is getting correct words. The scientific paper submitted in the system will detect non-standard words using Dictionary Lookup and typographical error using Peter Norvig Spelling Corrector.

KEBI 1.0 Checker has stored 1,453 pairs of standard words and non-standard words, 80,762 words in Indonesian, and 109,926 words in English. In addition, we are also preparing for the detection of punctuation errors in the future, namely conjunctions before the comma and after the comma, connecting expressions between sentences, looping words and pre-, post-, and inter-particles. The word list will be added periodically to the corpus and validated by Indonesian language experts. The corpus in KEBI refers to the KBBI (Indonesia Dictionary) and PUEBI (General Guidelines for Indonesian Spelling) published by the Indonesian language agency and then validated by language experts of the research group. However, this paper focuses more on discussing and evaluating non-standard spelling errors and typographical errors in scientific papers.

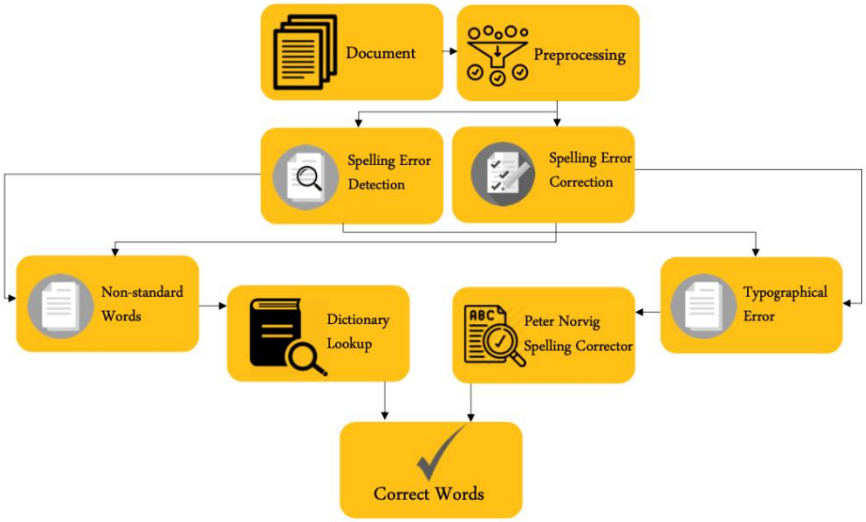

Figure 1. Proposed System Design of KEBI 1.0 Research

3.1. Indonesian Spelling Error in Scientific Paper

Indonesian spelling errors are included in language errors. Nurwicaksono et al. (2018) state that language errors are a form of error in using language that deviates from the language rules set by a language [11]. In this case, an Indonesian spelling error means an error that occurs due to the use of language that deviates from the rules in the General Indonesian Spelling Guidelines. Dulay et al. (1982) distinguish the types of language errors into four levels: linguistic category errors, performance strategy category errors, comparative category errors, and communication effects category errors [12]. However, this study only refers to linguistic category errors that focus on non-standard words and typographical errors.

The common Indonesian spelling errors in scientific papers, such as punctuation and letter errors. In this case, punctuation errors which include a dot (.), comma (,), semicolon (;), colon (:), hyphen (-), dash (--), question mark (?), exclamation point (!), ellipsis (…), quotation marks (“…”), single quotation mark ('…'), brackets ((…)), square brackets ([…]), slash (/) and apostrophes ('). In contrast, punctuation errors include capital letters, italics, and bold letters. The word writing errors include writing prepositions, standard words, particles, abbreviations and acronyms, numbers, pronouns, articles a/an/the. All spellings refer to the fourth edition of the General Guidelines for Indonesian Spelling published by the Language Development and Development Agency in 2016. Typing errors in vocabulary are also included in word spelling errors. As we know, vocabulary is a list of words used by authors to write science and technology [13].

Many works can be done to build applications that can detect Indonesian spelling errors in scientific papers but the features discussed in KEBI 1.0 in this study are about detecting non-standard words and typographical errors in writing standard words, including standard spelling forms because the word is commonly used in formal or official situations whose writing is in accordance with standardized rules. Whether or not a word is standardized can be seen in its pronunciation, Spelling, grammar, and nationality. Setiawati (2016) states that the standard rules in question can be in the form of spelling guidelines (EYD), standard grammar, and dictionaries [14]. While word typos are the technical error of an author in ensuring that the words were types is correct and can be read by readers.

3.2. Word Preprocessing

Preprocessing is the first step in processing input data before moving on to the main stage of Indonesian text document spelling correction [8]. The preprocessing that has been carried out in this study is tokenization and case folding. Formulas, characters other than alphabet letters, and images are eliminated because they are not needed in the spelling correction.

The explanation of the preprocessing stage is as follows:

a. Case folding is the process of change all of the letters in a document to lowercase. The letters 'a' to 'z' are the only acceptable ones. Case folding is used to handle the possibility the input might be a capitalized or non-capitalized letter. Figure 2 shows the input of the sentence “Spelling Error” will be changed in lower case to be “spelling error”.

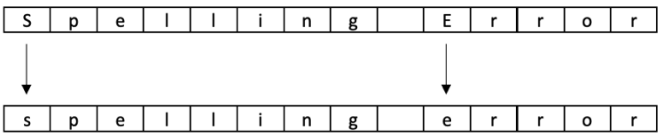

Figure 2. Case Folding in Preprocess Stage

b. Tokenization is the process of splitting text into individual words. This is useful for computational processes related to calculating the frequency and measuring the similarity of sentences based on words. Figure 3 shows the input of the sentences “Spelling Error Correction System” will be split into individual words to be “Speeling”, “Error”, “Correction”, and “System”.

Scientific papers consist of several paragraphs, especially in this study. To detect non-standard words and typos, it is necessary to partition paragraphs based on the "\n" sign as a new line. The next step is that each partition of paragraphs will be detected word by word iteratively containing spelling errors.

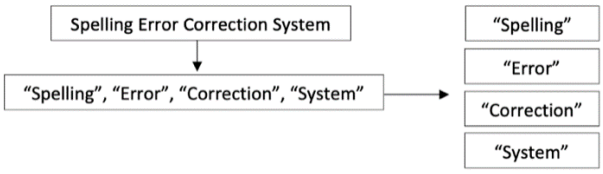

Figure 3. Tokenization in Preprocess Stage

Tokenization in the partition of paragraphs is illustrated in Figure 4. For example, there are one paragraph consists of 4 lines. After the paragraph is split into several lines based on the new line sign (“\n”), each line will be split again into individual words, and the system will detect the non-standard word and typos in the corpus.

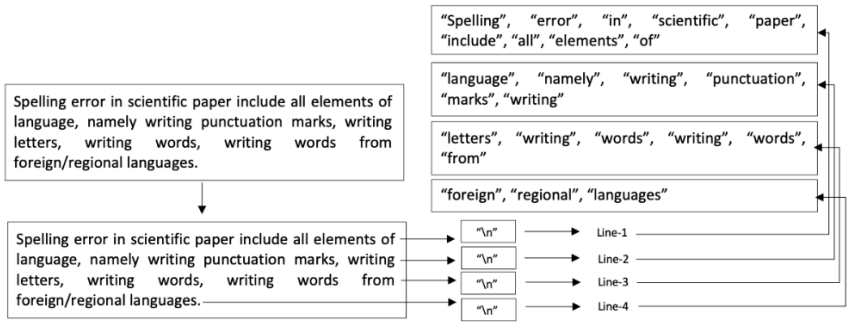

Figure 4. Tokenization in Partition of Paragraphs

3.3. Spelling Error Detection and Spelling Error Correction

Spelling error detection and spelling error correction are important parts of this research. The functional capabilities of both are needed to improve the performance of the Indonesian spelling error detection system. Spelling Error Detection is checking the validity of a word in a language. A word is considered to be valid when it can be found in a lexical resource [6]. In this case, the lexical resource is a data center in which there is a corpus, lexicon, word list, or other forms. The main error detection process is to compare the Indonesian Spelling in the text with the Indonesian Spelling contained in the lexical resource. Many methods are used for the error detection process. However, Dictionary Lookup and N-Gram Analysis are among the most commonly used.

After an Indonesian spelling error is detected, it will be corrected using the spelling error correction method. When the system has determined that the Spelling in the text has errors, the next step is to make suggestions for correcting the errors. The next step is to make suggestions for correcting the errors. To overcome spelling errors requires additional knowledge from other sources to extract sentence context [15]. Rule-based methods, similarity key techniques, Peter Norvig, and Minimum Edit Distance consisting of Levenshtein, Hamming, Damerau-Levenshtein, and Longest Common Subsequence (LCS) methods are widely used in similar research.

The spelling error correction can be divided into three tasks such as:

a. Autocorrect: the system directly corrects Spelling
b. Suggest a correction: the system will provide suggestions for improvements to the Spelling but waiting for action from the author
c. Suggestion list: the system will provide several spelling correction suggestions, and authors are allowed to choose the appropriate one.

3.4. Non-standard Words and Dictionary Lookup

The use of standard words in scientific papers is very important according to the rules of writing words in Indonesian languages. Standard words are generally often used in official sentences or a variety of standard languages, whether spoken or written. The standard word in Indonesian also has the following characteristics. **First**, both orally and in writing, standard words are used in official situations, such as official correspondence, legislation, scientific essays, research reports, etc. The standard language variety is not interfered with by certain dialects or accents. **Second**, both orally and in writing, standard words use the applicable provisions in the General Indonesian Spelling Guidelines. **Third**, both orally and in writing, the standard variety fulfills grammatical functions such as subject, predicate, and object wholly and explicitly [16].

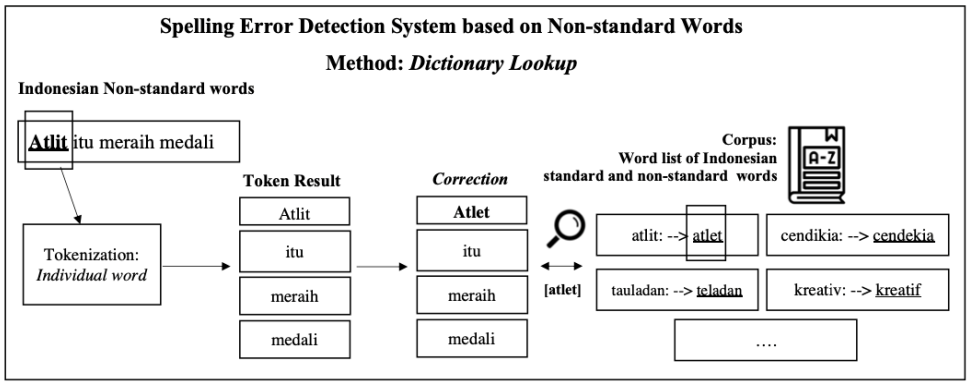

Figure 5. System Design of Spelling Error Detection Process Based on Indonesian Non-standard Words Using Dictionary Lookup

A word can be called a non-standard word if the word used is not in accordance with the Indonesian language rules. The non-standard of a word is caused by writing errors and can also be caused by incorrect pronunciation and incorrect preparation of a sentence. This non-standard word often appears in our daily life [17].

Some examples of Indonesian spelling errors in writing non-standard words are as follows:

a. Non-standard words: *atlit*, standard words: *atlet* (in English: athlete)
b. Non-standard words: *tauladan*, standard words: *teladan* (in English: role model)
c. Non-standard words: *cendikia*, standard words: *cendekia* (in English: scholar)
d. Non-standard words: *kreativ*, standard words: *kreatif* (in English: creative)

Figure 5 shows how the process of detecting spelling errors in sentences containing non-standard words is carried out. For example, a sentence contains the word “atlit”, after tokenization. The word "atlit" will be checked and matched with a keyword that contains non-standard Indonesian words in the corpus. Then, this keyword will go to a value containing the standard word. The standard word can replace the non-standard word in the previous input word. The non-standard word “atlit” will be replaced with the standard word “atlet”. This is a process called *key: value* that can be applied to the dictionary lookup method.

3.5. Typographical Error and Peter Norvig Spelling Corrector

Typographical error, or called a typo, is often done by authors when typing letters in digital documents. The more pages of the typed document, the greater the chance of the author's error in typing the words in the document. It can be avoided if the author checks the document page earlier before moving on to the next document page. The typographical error can occur when students type answers to essay questions and short answers [18]. In other cases, people sometimes mistype the URL when trying to access a website, leading to the wrong website page [19]. People sometimes make a typographical error when filling out their profiles in name, date of birth, address, and affiliation.

The typographical error can be categorized into two types, namely, non-word errors and real-word errors. Non-word error has no meaning in the Dictionary, while in real-word error, the written word is correct or has meaning in the Dictionary but is not intended in a sentence and has a different meaning [20]. The study focused on non-word errors because people often make these mistakes.

Some examples of a typographical error in non-word error are as follows:

a. The correct word: “Saya”, typo word: “Sya” (in English: i)
b. The correct word: “Indonesia”, typo word: “Indonsia” (in English: Indonesia)
c. The correct word: “adalah”, typo word: “adlah” (in English: is/am/are)

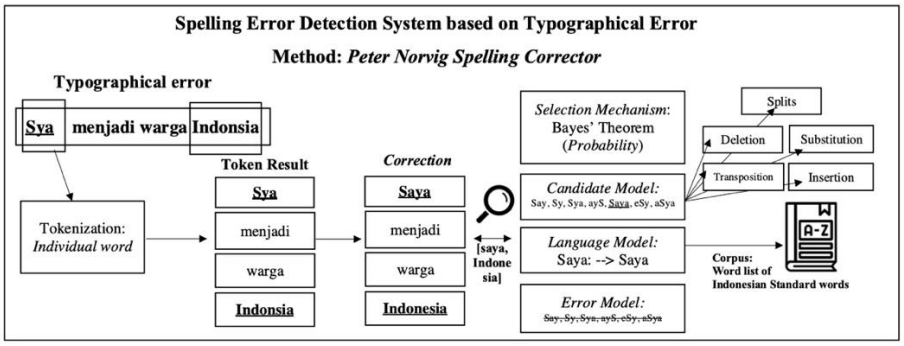

Figure 6. System Design of Spelling Error Detection Process Based on Typographical Error Words Using Peter Norvig Spelling Corrector

The spelling error detection system that can be applied to the typographical error words detection is Peter Norvig Spelling Corrector, as shown in Figure 6. The method uses probability to predict the possible closeness between the typo word and the words available in the corpus. For example, there is a typo "Sya" (the correct word: “Saya”, in English: “i”), the method will look for word candidates that are close to the actual word using candidate models such as splits, deletion, transpostion, substitution and insertion which will generate: “Say”, “Sy", “Sya”, “ayS", “Saya”, "eSy", and "aSya". Even the word will be combined with the characters "a" to "z". When the word candidates have calculated the probability in the corpus, then the typo word "Sya" is the word that

is closest to the word "Saya" in the corpus. The computational process in this method will continue to run in the background process until the closest word correction can be found.

Table 1. Peter Norvig Spelling Correction Process to Find The Word Correction using Splits, Deletion, Transposition, Substitution, and Insertion

Incorrect words (typo)				‘kmarin’			
Splits	'', 'kmarin'	'k', 'marin'	'km', 'arin'	'kma', 'rin'	kmar', 'in'	'kmari', 'n'	'kmarin', ''
Deletion	'marin'	'karin'	'kmrin'	'kmain'	'kmarn'	'kmari'	
Transposition	'mkarin'	'kamrin'	'kmrain'	'kmairn'	‘kmarni'		
Substitution	‘amarin’	‘kaarin’,	‘kmarin’	‘kmaain’	‘kmaran’	‘kmaria’	
	‘bmarin’	‘kbarin’	‘kmbrin’	‘kmabin’	‘kmarbn’	‘kmarib’	
	…	…	…	…	…	…	
	…	…	…	…	…	…	
	‘zmarin’	‘kzarin’	‘kmzrin’	‘kmazin’	‘kmarzn’	‘kmariz’	
Insertion	‘akmarin’	‘kamarin’	‘kmaarin’	‘kmaarin’	‘kmarain’	‘kmarian’	‘kmarina’
	‘bkmarin’	‘kbmarin’	‘kmbarin’	‘kmabrin’	‘kmarbin’	‘kmaribn’	‘kmarinb’
	…	…	…	…	…	…	…
	‘ekmarin’	‘kemarin’	‘kmearin’	‘kmaerin’	‘kmarein’	‘kmarien’	‘kmarine’
	…	…	…	…	…	…	…
	‘zkmarin’	‘kzmarin’	‘kmzarin’	‘kmazrin’	‘kmarzin’	‘kmarizn’	‘kmarinz’
Correct words				‘kemarin’			

For another example, there is a typo word “kmarin” (the correct word is “kemarin”, in English: “yesterday”). Peter Norvig Spelling Corrector will look for the combination of characters in its word to get the correct words and match the words in the corpus-based on probability. Table 1 shows Peter Norvig Spelling Corrector process to find the word correction using splits, deletion, transposition, substitution, and insertion. The typo word will be separated into two parts at the splits stage: the left and right words. The number of characters on the left word will increase by taking the frontmost character on the right word until it is empty.

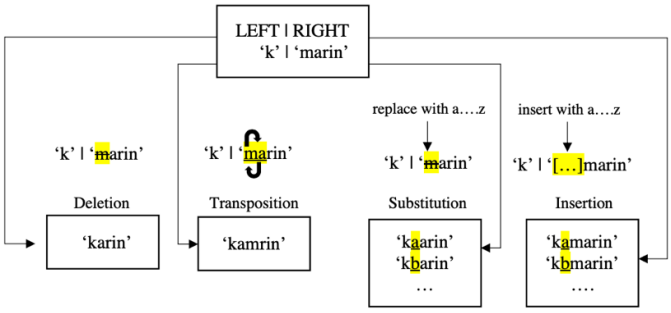

Figure 7. The Shifting Character Process of Typo Words using Peter Norvig Spelling Corrector

At the **deletion** stage, this will take advantage of the results of the splits. The left word is retained, but the first character in the right word is deleted, and then the left and right words are merged again. At the **transposition** stage, this will also take advantage of the result of the split. The left word is retained, but the first character's position in the right word is swapped with the second character in the right word, and then the left word and right word are merged again. At the **substitution** stage, the split results are still used, the left word is retained, but the position of the first character in the right word is replaced with 26 letters (from a to z), then the left word and right word are combined again. At the **insertion** stage, the split results are also used. The left word is retained, but the first character's position in the right word will be inserted with 26 letters (from a to z) then the left word and right word are combined again. The shifting character process of typo words using Peter Norvig Spelling Corrector is illustrated in Figure 7.

After all stages of getting a combination of words, calculate each candidate word's probability that matches the corpus. Peter Norvig Spelling Corrector has the word corrector function to choose

the closest spelling correction c for word w. None of the word candidates is absolutely selected because it is only a suggestion to use probability. The algorithm tries to find the correction c of all possible candidate correction which maximizes the probability that c is the addressed correction with the original word w, following the equation (1).

$$argmax_{c \in candidates} P(c|w) \quad (1)$$

By Bayes' Theorem, it is equivalent to the following equation (2):

$$argmax_{c \in candidates} \frac{P(c)P(C|W)}{P(w)} \quad (2)$$

Since P(w) is the same for every possible candidate c, we can factor it out following equation (3):

$$argmax_{c \in candidates} P(c)\ P(w|c) \quad (3)$$

The word candidate with the highest probability value has a high chance of being selected as the correct word.

4. Result and Discussion

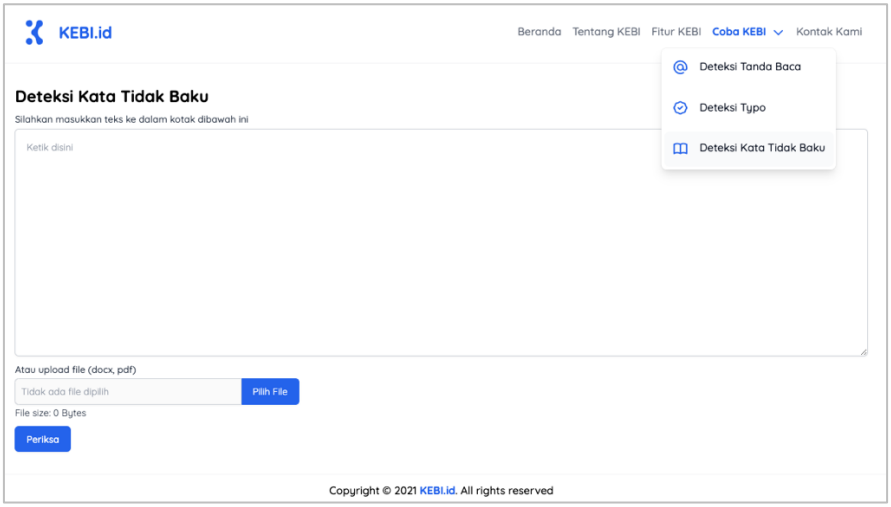

Figure 8. The User Interface (UI) of KEBI 1.0 Checker Web-based Application

The documents used in this study were obtained from a paper assignment written by ten students in Indonesian text. The paper consists of the title, author profile, abstract, introduction, literature review, method, conclusion, and references. There are two scenarios in this study. KEBI 1.0 will check for non-standard words and check for typos in the ten documents. The experimental results were evaluated by counting the number of words detected and the accurate spelling correction. In addition, the inaccurate spelling correction also is calculated.

4.1. Experiment Result of Spelling Error Detection based on Non-standard Words and Typographical Errors

Figure 8 shows the KEBI 1.0 Checker web-based application user interface, which has provided several features. They are spelling error detection based on punctuation, non-standard words, and typographical errors, but this study discussed two primary features. The application detected 209 meaningless words such as name, city, country, airport, abbreviation, and information technology terms (HTTP, www, etc.). Table 2 shows the percentage of spelling error correction which the application has given the best accuracy performance to correct the non-standard words reached 100%, while to correct the typographical error only reached 55.52%.

Table 2. The Percentage of Spelling Error Correction based on Non-standard Words and Typographical Errors

Student	Number of Pages	File Size	Spelling Correction based on Non-standard Words		Spelling Correction based on Typographical Errors		
			Accurate	Inaccurate	Accurate	Inaccurate	Meaningless Words
Student-1	5 pages	25.82 KB	14 of 14	0	6 of 11	5 of 11	22
Student-2	5 pages	17.96 KB	12 of 12	0	15 of 22	7 of 22	15
Student-3	14 pages	863.9 KB	14 of 14	0	16 of 24	8 of 24	31
Student-4	16 pages	5.85 MB	16 of 16	0	18 of 27	9 of 27	47
Student-5	8 pages	590.12 KB	8 of 8	0	7 of 15	8 of 15	4
Student-6	9 pages	1.82 MB	8 of 8	0	4 of 13	9 of 13	39
Student-7	6 pages	29.11 KB	13 of 13	0	10 of 15	5 of 15	14
Student-8	6 pages	21.34 KB	14 of 14	0	9 of 27	18 of 27	3
Student-9	8 pages	3.93 MB	8 of 8	0	14 of 21	7 of 21	19
Student-10	6 pages	54.1 KB	15 of 15	0	11 of 20	9 of 20	15
Average (%)			100%	0%	55.52%	44.48%	Total = 209

The features provided in KEBI are currently located in a separate menu for users because it is necessary to know how the performance of each of these features is. Document testing for non-standard word detection does seem to work well because the dictionary lookup algorithm only needs to find pairs of non-standard words with their standard words in the database. The result of the spelling correction will certainly be quite accurate, it only needs to increase the number of *key: values* in the corpus.

The challenge is how accurate the spelling error correction on the typo words is. Peter Norvig Spelling Corrector algorithm is only tasked with finding word candidates through splits, deletion, transposition, substitution, and insertion. In contrast, the results of word correction are highly dependent on the word list in the corpus. If the candidate word that is close to the typo has not been found in the corpus, it is possible to an error in the correction of the selected word. The problem is not in the algorithm's ability but the availability of word lists in the corpus.

For further research, we need to make a word extractor that can extract documents from various fields into terms that are stored in the corpus to detect spelling errors in scientific works, for example, in the fields of computer science, natural science, engineering, language and literature, social, economics, education, medical, and others.

4.2. The Word List of Spelling Error Correction based on Non-standard Words and Typographical Errors

The paper assignments that have been written by 10 students have many spelling errors such as writing active verbs, passive verbs, conjunctions, prepositions and placements, and standard words for general terms. That means, KEBI 1.0 Checker is useful for correcting spelling errors of the student paper assignment. Tables 3 and 4 show the word list of spelling error correction based on non-standard words and typographical errors that KEBI 1.0 Checker has detected.

Table 3. The Word List of Spelling Error Correction based on Non-standard Words

Non-standard Words (1)	Standard Words (1)	in English (1)	Non-standard Words (2)	Standard Words (2)	in English (2)
‘dimana’	‘di mana’	*where*	‘bis’	‘bus’	*Bus*
‘disamping’	‘di samping’	*besides*	‘service’	‘servis’	*service*
‘praktek’	‘praktik’	*practice*	‘pada hal’	‘padahal’	*even though*
‘ditempat’	‘di tempat’	*in place*	‘bertanggung jawab’	‘bertanggungj awab’	*to be responsible*
‘dulu’	‘dahulu’	*past*	‘komplek’	‘kompleks’	*complex*
‘tapi’	‘tetapi’	*but*	‘presentase’	‘persentase’	*percentage*
‘jaman’	‘zaman’	*era*	‘respon’	‘respons’	*response*
‘ditengah’	‘di tengah’	*in the middle*	‘seksama’	‘saksama’	*carefully*
'mempengaruhi'	‘memengaruhi’	*influence*	‘survey’	‘survei’	*survey*

Non-standard Words (1)	Standard Words (1)	in English (1)	Non-standard Words (2)	Standard Words (2)	in English (2)
'sumatera'	'sumatra'	*Sumatra (name of an island)*	'efektifitas'	'efektivitas'	*effectiveness*
'diatas'	'di atas'	*in above*	'kerjasama'	'kerja sama'	*cooperation*
'antri'	'antre'	*queue*	'berfikir'	'berpikir'	*think*
'jendral'	'jenderal'	*General*	'ibukota'	'ibu kota'	*Capital city*
'presepsi'	'persepsi'	*perception*	'shalat'	'salat'	*Salat*
…	…	…	…	…	…

Table 4. The Word List of Spelling Error Correction based on Typographical Errors

Typo Words (1)	Correct Words (1)	in English (1)	Typo Words (2)	Correct Words (2)	in English (2)
'tranportasi'	'transportasi'	*transportation*	'menengakkan'	'menegakkan'	*uphold*
'swasto'	'swasta'	*private*	'seharunya'	'seharusnya'	*should*
'baahasa'	'bahasa'	*language*	'indoonesia'	'indonesia'	*Indonesia (name of a country)*
'berkomnikasi'	'berkomunikasi'	*communicate*	'penumpanng'	'penumpang'	*passenger*
'penillaian'	'penilaian'	*assessment*	'perppindahan'	'perpindahan'	*Transfer of*
'internasioal'	'internasional'	*international*	'perekinomian'	'perekonomian'	*economy*
'mengikutu'	'mengikuti'	*follow*	'berwennang'	'berwenang'	*authorized*
'kenyataanya'	'kenyataannya'	*reality*	'peregerakan'	'pergerakan'	*movement*
'penggunaaan'	'penggunaan'	*use*	'variebel'	'variabel'	*variable*
'yogjakarta'	'yogyakarta'	*Yogyakarta (name of a city)*	'mamudahkan'	'memudahkan'	*make it easy*
'kuantiatif'	'kuantitatif'	*quantitative*	'permasalahn'	'permasalahan'	*problem*
'sebagaian'	'sebagian'	*part*	'berfikir'	'berpikir'	*think*
'perskriptif'	'preskriptif'	*prescriptive*	'menggurangi'	'mengurangi'	*reduce*
'disepakai'	'disepakati'	*agreed*	'suautu'	'suatu'	*a/an*
…	…	…	…	…	…

4.3. Document File Size and Processing Time

Table 5. Document File Size and Processing Time of KEBI 1.0 Checker Web-based Application

Student	Number of Pages	File Size	Processing Time	
			Non-standard Words	Typographical Errors
Student-1	5 pages	25.82 KB	0.11 second	18.74 seconds
Student-2	5 pages	17.96 KB	0.08 second	14.44 seconds
Student-3	14 pages	863.9 KB	0.14 second	21.12 seconds
Student-4	16 pages	5.85 MB	0.23 second	12.7 seconds
Student-5	8 pages	590.12 KB	0.07 second	19.5 seconds
Student-6	9 pages	1.82 MB	0.58 second	8.97 seconds
Student-7	6 pages	29.11 KB	0.08 second	6.42 seconds
Student-8	6 pages	21.34 KB	0.14 second	10.8 seconds
Student-9	8 pages	3.93 MB	0.09 second	13.41 seconds
Student-10	6 pages	54.1 KB	0.07 second	19.72 seconds
	Average		0.016 second	14.58 seconds

To know the performance of KEBI 1.0 for end-users, we evaluated the application's performance during the process of checking documents based on the number of pages, document file size, and processing time. The paper assignments that students have written consist of 5 to 16 pages, while the file size of 3.93 KB to 5.85 MB. Table 5 shows the processing time of KEBI 1.0 is relatively low. The average time needed to correct non-standard words is 0.016 seconds, and correcting typo words is 14.58 seconds.

4.4. The Comparison Analysis with Previous Research

We compared the objective, methods, and accuracy between the proposed research with the previous research following Table 6. The accuracy of the proposed methods is quite competitive with the other research, but KEBI 1.0 Checker has different objectives and alternative methods.

Table 6. The Comparison Analysis with Previous Research

Objective	Methods	Accuracy
Proposed the spelling correction application for text documents in Bahasa Indonesia [8].	▪ Finite State Automata (FSA) ▪ Levenshtein distance ▪ N-gram methods	▪ With FSA: 75.63% ▪ Without FSA: 83.75% ▪ Modification: 85.80%
Proposed a method to correct preposition errors in Indonesian sentences written by second language learners [9].	▪ Native ▪ CSRnd ▪ depWE-HeadObj ▪ depWE-Head	▪ depWE-Head: 70.59% ▪ Native: 76.07%
Proposed a method to microtext normalization process to convert these features into well-written text or in other words, convert non-standard words into standard words [10].	▪ Indonesian Dictionary-based ▪ Longest Common Subsequences (LCS) algorithm	▪ Dictionary-based: 82% ▪ LCS: 74% ▪ Dictionary-based + LCS: 84%
The comparison methods of edit distance	▪ Jaro-Winkler Distance	▪ Jaro-Winkler Distance: 21.65%
KEBI 1.0 Checker: Indonesian Spelling Error Detection System for Scientific Papers based on non-standard words and typographical errors	▪ Dictionary Lookup ▪ Peter Norvig Spelling Corrector	▪ Dictionary Lookup: 100% ▪ Peter Norvig Spelling Corrector: 55.52%

In addition to comparing the performance of the previous research, we have also compared with other edit distance methods such as Jaro-Winkler. However, the Jaro-Winkler method gives a relatively poor performance to correct spelling errors reached an accuracy of 21.65%. This method has not been able to correct the words in the document even and adequately stuck to provide corrections for long characters.

5. Conclusion

KEBI 1.0 Checker provides several features that can detect and correct non-standard Indonesian words and typographical errors in a scientific paper. The documents used in the experiment were obtained from a paper assignment written by ten students in Indonesian text. The experiment result shows that KEBI 1.0 Checker has given the best accuracy performance to correct the non-standard words reached 100%, while to correct the typographical error only reached 55,52%. The application has been detected 209 meaningless words such as name, city, country, airport, abbreviation, and information technology terms. The paper assignments that ten students have written have many spelling errors, such as writing active verbs, passive verbs, conjunctions, prepositions and placements, and standard words for general terms. We used ten student paper assignments consisting of 5 to 16 pages to test the application, while the file size was 3.93 KB to 5.85 MB. The processing time of KEBI 1.0 is relatively low. The average time needed to correct non-standard words is 0.016 seconds, and correcting typo words is 14.58 seconds. For further research, KEBI 1.0 Checker requires a large corpus to deal with typo words from various fields of scientific paper such as computer science, engineering, economics, social, medical, and other vocabularies.

Acknowledgment
The author would like to thank the Ministry of Education, Culture, Research, and Technology, Universitas Pembangunan Nasional “Veteran” Jawa Timur which has funded this research based on the Assignment Agreement for the Implementation of the Batch I Internal Research Program for the Basic Research Scheme (RISDA) in 2021, Number: SPP / 12 /UN.63.8/LT/IV/2021.

References

[1] S. Ariana, "Kesalahan Penggunaan Ejaan yang Disempurnakan dalam Karya Ilmiah Dosen Universitas Bina Darma," *Jurnal Ilmiah Bina Edukasi,* vol. 5, no. 5, pp. 53-62, 2011.

[2] H. Alwi and dkk, Tata Bahasa Baku Bahasa Indonesia, Jakarta: Balai Pustaka, 1998.

[3] Murtiningsih, "Kesalahan Berbahasa Indonesia Mahasiswa S-1 PGSD STIKIP Nuuwar Fak-fak," *Peneliti Ilmu Pendidik,* vol. 6, no. 1, pp. 74-82, 2013.

[4] G. L. Y. Londo, Y. S. P. W.P. and M. Maslim, "Pembangunan Aplikasi Identifikasi Kesalahan Ketik Dokumen Berbahasa Indonesia Menggunakan Algoritma Jaro-Winkler Distance," *AKSIS: Jurnal Pendidikan Bahasa dan Sastra Indonesia,* vol. 2, no. 2, pp. 138-153, 2018.

[5] T. Hartina and A. Masri, "Pendeteksi Kesalahan Pengetikan Kata Nonbaku pada Karya Tulis Menggunakan N-Gram," *Jurnal Informatika,* vol. 7, no. 1, pp. 77-84, 2020.

[6] R. N. E. Anggraini, M. A. Zinni and S. Rochimah, "Kakas Bantu Pendeteksi Kesalahan Tanda Baca pada Karya Tulis Ilmiah," *JUTI,* vol. 14, no. 1, pp. 117-125, 2016.

[7] D. Surianto, D. Triyanto and U. Ristian, "Penerapan Algoritma Boyer Moore dan Metode N-Gram pada Aplikasi Penyunting Naskah Teks Bahasa Indonesia Berbasis Web," *Coding: Jurnal Komputer dan Aplikasi,* vol. 8, no. 3, pp. 50-60, 2020.

[8] V. C. Mawardi, N. Susanto and D. S. Naga, "Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method," *MATEC Web of Conferences,* vol. 164, no. 7, p. 1047, 2018.

[9] B. Irmawati, H. Shindoa and Y. Matsumotoa, "Exploiting Syntactic Similarities for Preposition Error Corrections on Indonesian Sentences Written by Second Language Learner," *Procedia Computer Science,* vol. 81, pp. 214-220, 2016.

[10] D. Gunawan, Z. Saniyah and A. Hizriadi, "Normalization of Abbreviation and Acronym on Microtext in Bahasa Indonesia by Using Directionary-based and Longest Common Subsequence (LCS)," *Procedia Computer Science,* vol. 161, pp. 533-559, 2019.

[11] B. D. Nurwicaksono and D. Amelia, "Analisis Kesalahan Berbahasa Indonesia Pada Teks Ilmiah Mahasiswa," *AKSIS: Jurnal Pendidikan Bahasa dan Sastra Indonesia,* vol. 2, no. 2, pp. 138-153, 2018.

[12] H. C. Dulay, M. K. Burt and S. D. Krashen, Language Two, New York: Oxford University, 1982.

[13] Supriadin, "Identifikasi Penggunaan Kosakata Baku Dalam Wacana Bahasa Indonesia Pada Siswa Kelas VII Di SMP Negeri 1 Wera Kabupaten Bima Tahun Pelajaran 2013/2014," *JIME,* vol. 2, no. 2, pp. 150-161, 2016.

[14] S. Setiawati, "Penggunaan Kamus Besar Bahasa Indonesia (KBBI) Dalam Pembelajaran Kosakata Baku dan Tidak Baku pada Siswa Kelas IV SD," *Jurnal Penelitian Bahasa dan Sastra Indonesia,* vol. 2, no. 1, pp. 44-51, 2016.

[15] A. R. N., M. Kamayani, R. Reinanda, S. Simbolon, M. Y. Soleh and A. Purwarianti, "Application of Document Spelling Checker for Bahasa Indonesia," in *2011 International Conference on Advanced Computer Science and Information Systems*, Jakarta, 2011.

[16] S. and S. Saudah, Buku Ajar Bahasa Indonesia Akademik, Yogyakarta: Pustaka Pelajar, 2015.

[17] V. S. Ningrum, "Penggunaan Kata Baku dan Tidak Baku di Kalangan Mahasiswa Universitas Pembangunan Nasional "Veteran" Yogyakarta," *Jurnal Skripta: Jurnal Pembelajaran Bahasa dan Sastra Indonesia,* vol. 5, no. 2, pp. 22-27, 2019.

[18] V. C. Mawardi, F. Augusfian, J. Pragantha and S. Bressan, "Spelling Correction Application with Damerau-Levenshtein Distance to Help Teachers Examine Typographical Error in Exam Test Scripts," *E3S Web of Conferences,* vol. 188, pp. 1-10, 2020.

[19] E. Hargittai, "Hurdles to Information Seeking: Spelling and Typographical Mistakes During Users' Online Behavior," *Journal of the Association for Information Systems,* vol. 7, no. 1, pp. 52-67, 2006.

[20] A. I. Fahma, I. Cholissodin and R. S. Perdana, "Identifikasi Kesalahan Penulisan Kata (Typographical Error) pada Dokumen Berbahasa Indonesia Menggunakan Metode N-gram dan Levenshtein Distance," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 2, no. 1, pp. 53-62, 2018.