International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:2, 2013

Parallel Text Processing: Alignment of Indonesian to Javanese Language

Aji P. Wibawa, Andrew Nafalski, Neil Murray, Wayan F. Mahmudy

Abstract**—**Parallel text alignment is proposed as a way of aligning *bahasa Indonesia* to words in Javanese. Since the one-to-one word translator does not have the facility to translate pragmatic aspects of Javanese, the parallel text alignment model described uses a phrase pair combination. The algorithm aligns the parallel text automatically from the beginning to the end of each sentence. Even though the results of the phrase pair combination outperform the previous algorithm, it is still inefficient. Recording all possible combinations consume more space in the database and time consuming. The original algorithm is modified by applying the edit distance coefficient to improve the data-storage efficiency. As a result, the data-storage consumption is 90% reduced as well as its learning period (42s).

Keywords**—**Parallel text alignment, phrase pair combination, edit distance coefficient, Javanese-Indonesian language.

I. Introduction

LANGUAGE is recognized as one of a nation cultural identity. It is used to communicate with others as well as media in art. The language is categorized as endangered language when the use of it is uncommon. For example in Indonesia, with the number of their users decreasing, the preservation of traditional languages is becoming a national issue. Javanese, for example, a traditional language with the largest population of users (75 million) [1], is no longer considered fashionable among Indonesian teenagers [2] who see it as irrelevant in an era of globalization. Moreover, difficulties arise when youngsters try to use spoken Javanese using appropriate levels of politeness.

Javanese has a complex sub system which is called levels of speech. The levels are focused on how to communicate with others based on the attributes of subject and object of the utterance. The attributes are age difference, social status and relationships between speaker and interlocutor. Javanese speech levels can be related to the traditional performance which used the language to deliver goodness. People may unable to learn morality from the art, unless they understand the language. Consequently, improper use of speech levels is not only dangerous to the language existence but is also harmful to Javanese culture.

Finding technological solutions may be one means of saving endangered languages. In means of saving dying out languages by using technology, an Artificial Javanese Intelligent Tutor (AJI-Tutor) is currently in development [3]. In the first stage of AJI-Tutor development, a simple bilingual translator was created as the foundation of the politeness agent of the system. However, that translator utilized a single (one-to-one) word translation system which was unable to translate Javanese properly since the language has pragmatic meaning realized through particular combinations of words.

The pragmatic competence, an ability to understand the accurate expression and interpretation of intended meaning [4], is a complicating factor in translation. In order to cope with that factor, The next development of the translation is based on the research that phrase based alignment as an enhanced option to form bilingual corpora for translation knowledge base because of its accuracy and flexibility [5], [6]. The results produced are better than the single word alignment algorithm as all pair combinations are recorded. However, the knowledge base produced by the algorithm may become inefficient due to a rich corpus which consists of large structured sets of texts [7]. The greater trained parallel text consumes more space in the database as make the searching process become slower. In the last part of this paper, we modify the Javanese automatic parallel text alignment to increase the efficiency of the database.

II. Javanese Translation Modeling

Javanese linguists [8]-[11] classify the speech levels into three levels *krama*, *madya* and *ngoko. Ngoko* (Ng) is a casual speech, used between friends and close family. *Madya* (Md) is everyday speech used between villagers, and *krama* (Kr) is refined, formal speech used with and by high-status people. The three levels of politeness effectively constitute different dialects, albeit with similar grammars.

A complicating factor in translation is the need to accurately represent pragmatic meaning. For example, when someone wants to translate bapakku turu (my father is sleeping) from *ngoko* into *krama*,both are different speech levels in Javanese, this can be translated as either *bapak kula tilem* or *bapak kula sare*. Both forms are grammatically correct but only the second translation is pragmatically acceptable – or felicitous – for it allows the speaker to adopt a position of humility and incorporates honorific. words used for elder people with higher social status [8]. Furthermore, some individual *ngoko* words cannot be translated into individual

A. P. Wibawa is a PhD student at School of Electrical Information and Engineering of University of South Australia and a lecturer at Department of Electrical Engineering , State University of Malang (UM) Indonesia (e-mail: aji.wibawa@ mymail.unisa.edu.au).

A. Nafalski is a professor of electrical engineering at School of Electrical Information and Engineering of University of South Australia (phone: 061-883-023932; e-mail: Andrew.Nafalski@unisa.edu.au).

N. Murray is an associate professor of applied linguistics at University of Warwick, Centre for Applied Linguistics (e-mail: N.L.Murray@warwick.ac.uk).

W. F. Mahmudy is a lecturer at Department of Computer Science, Brawijaya University (UB) Indonesia (e-mail: wayanfm@ ub.ac.id).

krama words. For instance, the phrase bapak kula (my father) in *krama* is the literal translation of bapakku in *ngoko*. Since aligning a single word with only one word in another level is often unfeasible, the language must be aligned in a pair combination; that is, a lexical and pragmatic combination.

The lexical combination is created based on the unique translation of Javanese words. When transforming the language literally, one *krama* word may change into two words (1:2) at other speech levels; the same applies to *bahasa Indonesia*, and vice versa (2:1). Furthermore, the reverse combination must be provided if the translation direction is changed. However, a single word alignment (1:1) is still covered to accommodate the translation of one single word to another single word, as shown in Table I.

TABLE I
EXAMPLES OF JAVANESE LEXICAL TRANSLATION

bahasa Indonesia	Ngoko	krama	English meaning
ibuku	*ibuku*	*ibu kula*	*my mother*
mengapa	*geneya*	*kenging menapa*	*why*
hari libur	*prei*	*prei*	*a holiday*
kita	*awake dhewe*	*kita*	*us*
kuda	*jaran*	*turangga*	*a horse*
sama	*padha*	*sami*	*equal*

The pragmatic combination is modeled to align particular pairs. Referring to Javanese subject-verb agreement (SVA) [12], the verb may change based on who is the subject of the action. The example in Table II shows how a verb employed in speech may be totally different depending on the social status or age of the subjects. This fact needs to be accommodated.

TABLE II
EXAMPLE OF PRAGMATIC PAIR ALIGNMENT

Action (*ngoko*)	Higher status subject (father)	Lower status subject (young brother)
sleeping (*turu*)	*bapak turu (Ng)* *bapak sare (Kr)*	*adik turu (Ng)* *adik tilem (Kr)*
talking (*omong*)	*bapak omong (Ng)* *bapak ngendika (Kr)*	*adik omong (Ng)* *adik matur (Kr)*
eating (*mangan*)	*bapak mangan (Ng)* *bapak dhahar (Kr)*	*adik mangan (Ng)* *adik nedha (Kr)*
taking a bath (*adus*)	*bapak adus (Ng)* *bapak siram (Kr)*	*adik adus (Ng)* *adik adus (Kr)*

Both pragmatic and lexical pair combinations (P) are then modeled as follows.

$$P \begin{cases} (n_i, 0, k_j, 0): single\ word\ pair\ (1:1) \\ (0, n_i, k_j, k_{j+1}): one - two\ pair\ (1:2) \\ (n_i, n_{i+1}, 0, k_j): two - one\ pair\ (2:1) \\ (n_i, n_{i+1}, k_j, k_{j+1}): two - two\ pair\ (2:2) \end{cases} \quad (1)$$

III. THE JAVANESE AUTOMATIC PAIRED BI-TEXT ALIGNMENT

Basically, the Javanese automatic paired bi-text alignment consists of two processes, text parsing and phrase combination alignment. The results then recorded into the database as shown in Fig. 1.

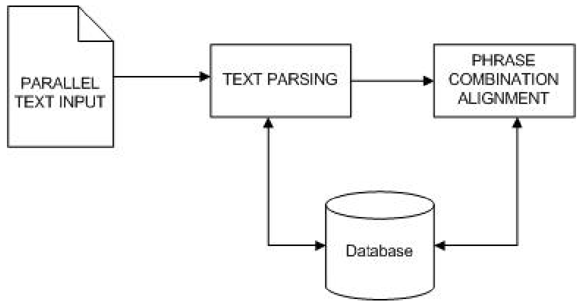

Fig. 1 Javanese automatic paired bi-text alignment

A. The Database

The database is designed to store and manage the data (Fig. 2). It consists of two tables of words (WINA and WKrm) and two phrase tables (PINA and PKrm) to record unique words and phrases in both languages. The primary key of word tables is the ID_word (ID_INA and ID_ Krm) that is an auto increment integer. The combination of two ID_words in phrase tables can be considered as primary key of such tables because of its uniqueness. The contents of these four tables are the results of parsing process. Afterwards, the parsing data will be stored as pairs in the Table of pairs. Furthermore, the frequency of word, phrase and pair is recorded in the database during the related procedures.

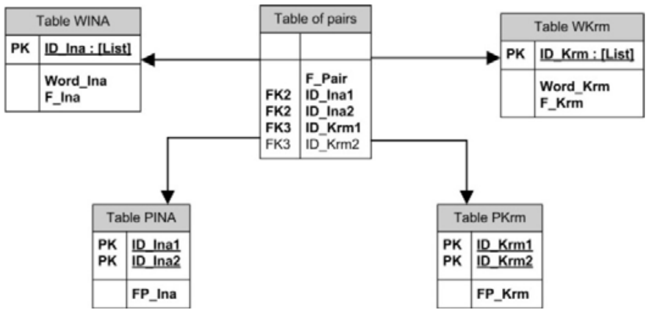

Fig. 2 Database design

B. The Text Parsing Algorithm

The text parsing stage is a monolingual process, where every text divided into sentences then parsed into a list of words. Based on definitions and models explained before, punctuation still used as the separator between sentences as well as the space to distinct words. The training data is the parallel text which parsed using the modified parsing algorithm as shows in Fig. 3.

International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:2, 2013

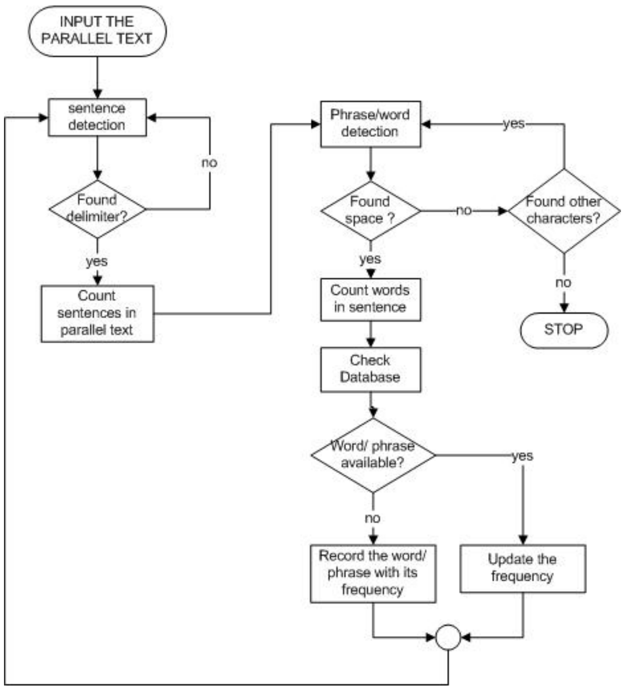

Fig. 3 The parsing process

Considering that every sentence is made of words, the next stage consists of parsing the parallel sentences into a set of words defined by the equation below.

$$s_i = \{n_j : a \text{ indonesia } word \mid 0 \le j < w_s\} \quad (2)$$

$$t_i = \{k_j : a\ krama\ word \mid 0 \le j < w_t\} \quad (3)$$

Space (j) acts as separator between words in a sentence. If a space is found on either side of characters, the system recognizes characters before the space as a word and thus indexes and stores it in the database.

After words identified, the availability of the words will be checked in the database. If the database is empty, the word will be classified as a new one, then stored and automatically indexed. The frequency of such a word also recorded because the value will be applied in the translation process. However, when the system found that the specific word had been inserted, the frequency will be updated without any changing to the words' order in the database.

Two arrays created to substitute and simplify the bilingual text. The rows indicated the ith sentence in the monolingual text while columns represented jth word in every sentence. The content of arrays is the word's index as substitution of the related word in the database. For example, the sentence *ayahku tidur* recorded as 1 and 2 in an array of *bahasa Indonesia* which will be implemented in the automatic alignment process. Tables III A and B show the array of each dialect which will be implemented in the automatic alignment process.

TABLE III A
MAPPING SENTENCE INTO ARRAY OF BAHASA INDONESIA

Sentence	Array			
	w1	w2	w3	...wn
ayahku tidur.	1	2		
aku tidur.	3	2		
ayahku makan.	1	4		
aku makan.	3	4		

TABLE III B
MAPPING SENTENCE INTO THE ARRAY OF *KRAMA*

Sentence	Array			
	w1	w2	w3	...wn
bapak kula sare.	1	2	3	
kula tilem.	2	4		
bapak kula dhahar.	1	2	5	
kula nedha.	2	6		

C. The Phrase Combination Alignment

Providing knowledge for the bilingual translation in form word or phrase pairs is the aim of parallel text alignment. The pairs are bilingually composed from arrays of *krama* (K) and *bahasa Indonesia* (B). The first stage is checking the equality of the number of sentences in both arrays. If the number is unequal, the minimum value procedure will choose the lowest array as a reference of the alignment. When the number of sentences is equal, matrix are ready to arrange into pairs.

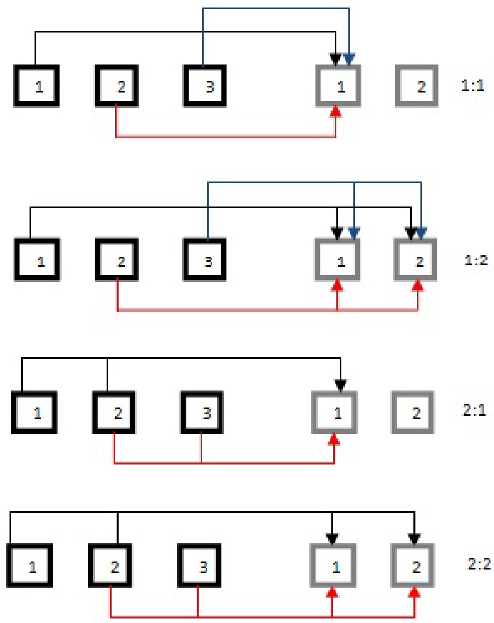

Fig. 4 Pair combinations illustration

The next step is pairing the matrix elements based on lexical and pragmatic pair combinations (1). For example, one element of K will be coupled with single element of B (1:1). The pair will be recorded as a new data as well as its frequency (F_pair). Contrary, if existed pair found, the system will only update the frequency of the related pair. Those procedures are repeated from the beginning to the end of the particular row of the matrix, represent a sentence. The iteration for one type pair combination will stop after processing the last sentence in the parallel matrix. Afterwards, the alignment process will be repeated for the rest combinations (Fig. 4). The flowchart in Fig. 5 illustrates the whole process of the automatic bilingual alignment.

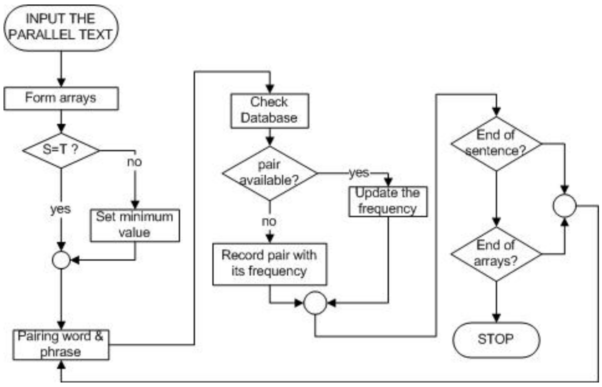

Fig. 5 Pair alignment process

D. Testing Results

The result covers all possible combinations based on the highest frequency of the pair as the accepted alignment. The other combinations are not deleted because they may come to constitute a proper alignment when the quantity of training data is increased. The alignment results should be more accurate if the number of examples is extended since it is based on the occurrence frequency of the word and phrase pair in the sentence. Table IV illustrates the 15 combinations of matrix elements (*ayahku tidur/ bapak kula sare*) extracted from the first row of Tables IIIA and IIIB.

TABLE IV
PAIR COMBINATIONS OF "*AYAHKU TIDUR*" WITH "*BAPAK KULA SARE* "

ID_Ngk1	ID_Ngk2	ID_Krm1	ID_Krm2	F_pair
1	0	1	0	2
2	0	1	0	1
1	0	2	0	2
2	0	2	0	2
1	0	3	0	1
2	0	3	0	1
1	2	1	0	1
1	2	2	0	1
1	2	3	0	1
1	0	1	2	2
2	0	1	2	1
1	2	1	2	1
1	0	2	3	1
2	0	2	3	1
1	2	2	3	1

However, when the number of examples is increased (Table V); the number of pair combinations (#P) increase as well as the learning time. Furthermore, the length and complexity of the sentence may prolong the learning iteration. As presented by Table IV, a huge difference between total words in the training data and the corresponding pair combinations shows that the proper alignments are mixed with the unintended ones.

TABLE V
ALIGNMENT RESULTS FOR VARIOUS PARALLEL TEXTS

#W		#S	AvgW/S	#P	t (s)
Kr	Ina				
8	10	4	2	43	7
253	235	36	7	7800	344
828	855	105	7	31612	4805

IV. EDIT DISTANCE COEFFICIENT TO INCREASE THE EFFICIENCY OF BI-TEXT ALIGNMENT

The automatic alignment records all combinations in the knowledge base. However, when the sentences become long and complex, recording of all possible pairs of the phrase is not efficient. The process may too slow and consume more space in the database [13]. In order to increase the learning efficiency and to reduce data storage consumption, shifting distance coefficient (D) is proposed. The shifting distance is set in every alignment combination of *krama* (K) and *bahasa Indonesia* (B) to limit the iteration, as shown in the next pseudo code.

```
for each sentence in bahasa Indonesia
for c:= i - D  to  i + D do
   if (c>0) and (c<=K[j,0])
      then
      P{(b_i,0,k_j,0 ):one-one,(0,b_i,k_j,k_j+1): one-two,
      (b_i,b_i+1,0,k_j):two-one, (b_i,b_i+1,k_j,k_j+1):two-two}
      check the database
            if the combination is unavailable in database
      then
            record the pair combination with its
      frequency
            else update the frequency of the pair
            end if
      end for
end for
```

Iteration limit will be changed by adjusting the shifting distance (Fig. 6). While the coefficient is set to zero, any word or phrase in *krama* is paired with another equally indexed chunk in *bahasa Indonesia*. As pictured by Fig. 3, if D=0, the result of the 1:2 pair combination is P $\{(b_2,0,k_2,k_3)\}$. When the coefficient tuned to one, the iteration will start from one word before (i-1) then ended in one after the reference word (i+1). As a result, the pair extended to P $\{(b_2,0,k_1,k_2), (b_2,0,k_2,k_3), (b_2,0,k_3,k_4)\}$ for just the 1:2 combination. Therefore, the original combination, for instance, P $\{(b_2,0,k_1,k_2), (b_2,0,k_2,k_3), (b_2,0,k_3,k_4),..., (b_2,0,k_{n-1},k_n)\}$ is simplified to speed up the process and to save the data space.

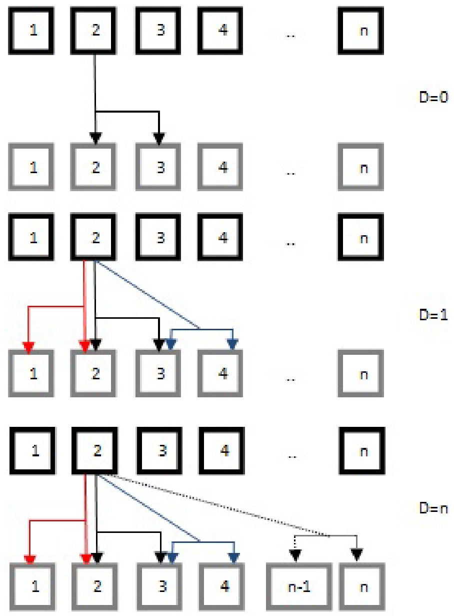

Fig. 6 Pairing bilingual words (2:1) in various shifting distance (D)

A bilingual short story is used to evaluate the efficiency of implementing shifting distance coefficient in the automatic alignment algorithm. The parallel text consists of long and complex sentences, 25 words per sentence. Table VI shows that the number of words between texts is unequal which means some single word may be translated into more than on in another language. The total words and the number of unique words are measured then stored in the database. The average word’s frequency and the average number of words in a sentence are calculated by dividing the total words with the number of unique words and the number of sentences correspondingly.

TABLE VI
DETAILS OF PARALLEL TEXT

Text details	***Krama***	***Bahasa Indonesia***
total words	253	235
the number of unique words	164	159
the average of word’s frequency	2	1
the number of sentences	36	36
the maximum words in a sentence	25	25
the average number of words in a sentence	7	7

$$E_d = \frac{\max PC - PC_d}{\max PC} \quad (4)$$

Afterwards, the bilingual text is automatically aligned by tuning various values of shifting distance coefficient (Table VII). The formed pair candidates (PC_d) and the learning period (t_d) are measured during the alignment process. The maximum quantity of PC is used as reference to measure the storage efficiency (E). The efficiency calculates by dividing the difference between pair candidates with the value of the reference number (2).

TABLE VII
RESULTS OF BI-TEXT ALIGNMENT WITH VARIOUS SHIFTING DISTANCE COEFFICIENT

D	**Pair candidates (PC_d)**	**Learning period (t_d)**	**Storage efficiency (E_d)**
0	783	42	0.9
1	2164	105	0.72
2	3345	155	0.57
N	7800	344	0

From Table VII, we can conclude that the standard coefficient D=0 can align the parallel text fastest (42 seconds) and increase (0.9) the data-storage efficiency. In other words, the smallest coefficient D provides the quickest learning period and the most efficient data storage. However, setting D to its minimum value should be avoided in order to prevent the loss of potential pairs recorded during the training progress. The suggested value of the shifting distance coefficient is one due to the rule of the Javanese translation that one Javanese word may be translated into two words in different levels of speech.

V. CONCLUSION

This algorithm can be considered the first development of bilingual text of Javanese speech levels given that other alignments focus on English or other languages [5]-[7], [14]. While the single word alignment algorithm is unable to align Javanese properly, the automatic alignment algorithm can extract the parallel texts into pairs of languages based on Javanese rules of translation. However, the frequency-based algorithm needs to be further developed in order to differentiate the proper alignment of pairs with low frequency of occurrence. Experiments to define the optimal frequency threshold should be established to classify the low-frequency occurrence pairs as incorrect alignment.

The original Javanese parallel text alignment is modified in order to increase the data-storage efficiency. As a result, the storage efficiency is increased (90%) by applying shifting distance coefficient to the alignment algorithm. In consequence, the application of the coefficient can boost the speed of the learning process by cutting the number of iteration in each sentence. The modified algorithm may contribute to the next development of the Javanese pragmatic translation which will be consisting of at least five languages; *bahasa Indonesia* , *ngoko*, *ngoko alus*, *krama* and *krama alus* [15].

REFERENCES

[1] W. Wedhawati, *et al.*, *Latest structure of Javanese language* Yogyakarta: Kanisius, 2006.
[2] D. E. Subroto, *et al.*, "Endangered krama and krama Inggil varieties of the Javanese language," *Linguistik Indonesia*, vol. 26, pp. 89-96, 2008.
[3] A. P. Wibawa and A. Nafalski, "Intelligent tutoring system: a proposed approach to Javanese language learning in Indonesia," *World Institute for Engineering and Technology Education* vol. 8, pp. 216-220, 2010.
[4] N. Murray, "Pragmatics, awareness raising and the cooperative principle.," *E:T Journal*, pp. 1-9, 2009.
[5] J. Zhao, *et al.*, "Two-phase base noun phrase alignment in Chinese-English parallel corpora," in *Natural Language Processing and Knowledge Engineering*, Wuhan, 2005, pp. 360-365.

[6] L. Ahrenberg, *et al.*, "A simple hybrid aligner for generating lexical corespondences in parallel text," in *36 th Annual Meetingof the Association for Computational Linguistics* Montreal, Quebec, Canada., 1998, pp. 29-35.
[7] R. Terashima, *et al.*, "Learning method for extraction of partial correspondence from parallel corpus," in *International Conference on Asian Language Processing*, Singapore, 2009, pp. 293-298.
[8] S. Poedjosoedarmo, "Javanese Speech Levels," *Indonesia,* pp. 54-81, 1968.
[9] S. Poedjosoedarmo, "Wordlist of Javanese Non-Ngoko Vocabularies," *Indonesia,* vol. 7, pp. 165-190, 1969.
[10] P. Purwadi, *et al.*, *Javanese language structure*. Yogyakarta: Media Abadi, 2005.
[11] A. B. Setiyanto, *Parama Satra: Javanese Language*. Yogyakarta: Panji Pustaka, 2010.
[12] Sukarno, "The Reflection of the Javanese Cultural Concepts in the Politeness of Javanese," *k@ta,* vol. 12, pp. 59-71, 2010.
[13] A. P. Wibawa, *et al.*, "Edit Distance Algorithm To Increase Storage Efficiency Of Javanese Corpora," in *International Conference on Computer, Electrical, and Systems Sciences, and Engineering (ICCESSE)*, Singapore, 2012, p. 5.
[14] D. Sotiropoulus, *et al.*, "Application of a word alignment algorithm to bilingual Greek-Latin documents," in *International Conference on Applied Computer Science*, Venice, Italy, 2007, pp. 238-241.
[15] S. Wibawa, "Efforts to maintain and develop Javanese language politeness," in *International Seminar of Javanese Language*, Paramaribo,Suriname, 2005, pp. 1-10.