ERRORS BY AUTO-MORPHOLOGICAL ANALYSIS IN A CHILDREN STORY CORPUS: AN EVALUATION OF MORPHIND PROGRAM

A FINAL PROJECT

In Partial Fulfillment of the Requirement

For S-1 Degree in Linguistics

In English Department, Faculty of Humanities

Diponegoro University

Submitted by:

Noveka Erviana Nur Alfiani

13020113120025

FACULTY OF HUMANITIES

DIPONEGORO UNIVERSITY

SEMARANG

2017

PRONOUNCEMENT

The writer honestly confirms that she compiled this final project entitled "Errors by Auto-Morphological Analysis in a Children Story Corpus: An Evaluation of Morphind Program" by me without taking any result from other researchers in S-1, S-2, S-3 and in diploma degree of any university. The writer ascertains that she did not quote any material from other publications or someone's paper except from the references mentioned.

Semarang, 24th August 2017

Noveka Erviana Nur Alfiani

MOTTO AND DEDICATION

"We cannot solve our problems with the same thinking,
we used when we created them".

– Albert Einstein

"You were given this hard life because you are strong enough to face it".

– Anonymous

"Janganlah kamu bersikap lemah, dan janganlah pula kamu bersedih hati, padahal kamulah orang-orang yang paling tinggi derajatnya, jika kamu orang-orang yang beriman."

– ***(Q.S. Al-Imran: 139)***

This final project is dedicated to my beloved father and mother and my family

APPROVAL

ERRORS BY AUTO-MORPHOLOGICAL ANALYSIS IN A CHILDREN STORY CORPUS: AN EVALUATION OF MORPHIND PROGRAM

Written by:
Noveka Erviana Nur Alfiani
13020113120025

is approved by the project advisor

on 24th August, 2017

Project Advisor,

Prihantoro, S.S., M.A.

NIP. 198306292006041002

The Head of the English Department,

Dr. Agus Subiyanto, M.A.
NIP. 196408141990011001

VALIDATION

Approved by
Strata-1 Final Project Examination Committee
Faculty of Humanities, Diponegoro University
On September 27th, 2017

Chair Person	First Member
Dr. Agus Subiyanto, M.A.	Drs. Mualimin, M.Hum.
NIP. 196408141990011001	NIP. 196111101987101001
Second Member	Third Member
Dwi Wulandari, M.Hum. M.A.	Drs. Oktiva H. Chandra, M.Hum
NIP. 197610042001122001	NIP. 196710041993031003

ACKNOWLEDGMENT

Praise to Allah SWT, who has given strength and spirit, so this final project entitled "Errors by Auto-Morphological Analysis in a Children Story Corpus: An Evaluation of Morphind Program" could be accomplished as one of the requirements to achieve bachelor degree at English Department, Faculty of Humanities, Diponegoro University.

The deepest gratitude is given to Prihantoro, S.S., M.A as the writer's advisor, for his guidance, helpful correction, time, patience, advice, and kindness until this final project is completed.

The writer would also like to extend special thanks to:

1. Dr. Redyanto Noor, M.Hum., as the Dean of Faculty of Humanities, Diponegoro University Semarang;
2. Dr. Agus Subiyanto, M.A., as the Head of the English Department, Faculty of Humanities, Diponegoro University and the writer's academic advisor for his support and guidance from the first semester until now;
3. All of the lectures in the English Department, Faculty of Humanities Diponegoro University for their knowledge shared to me;
4. My beloved father Jupri, mother Nurhadiyah (almh), my aunty, my uncle, my nephew, and all my families who always support and motivate the writer;

5. All friends in English Department 2013 especially in class A and Linguistic class, thank you for the memorable years in campus;
6. My beloved friends “Rempongers” Nindi, Davin, Rista, Icha, Elfida, Vita, My beloved roomate Nike, Indah and, My beloved friends “Nongsky” thank you for you for the support and kindness;
7. My beloved partner Ghaniy, Denissa, Ersa, Nurul, Fina, Lala, and all my KKN squad Rowosari thank you for the support and kindness;
8. Everyone who had helped and supported the writer in finishing this final project;

The writer realizes that this final project is still far from being perfect. Therefore, any criticism and suggestion are expected to make this final project better. Finally, the writer expects that this final project will be useful for all the readers.

Semarang, 24th August 2017

Noveka Erviana Nur Alfiani

TABLE AND CONTENTS

LIST OF ABBREVIATION

1. Lemma Tag

n	Noun
p	Personal Pronoun
v	Verb
c	Numeral
q	Adjective
h	Coordinate Conjunction
s	Subordinating Conjunction
f	Foreign Word
r	Preposition
m	Modal
b	Determiner
d	Adverb
t	Particle
g	Negation
i	Interjectio
o	Copula
w	Question
x	Unknown
z	Punctuation

2. Morphological Tagset

1st Position		2nd Position		3rd Position	
N	Noun	P	Plural	F	Feminine
		S	Singular		
				M	Masculine
				D	Non-Specified
—		—		—	
P	Personal Pronoun	P	Plural	1	First Person
		S	Singular	2	Second Person
				3	Third Person
—		—		—	
V	Verb	P	Plural	A	Active Voice
		S	Singular	P	Passive Voice
—		—		—	
C	Numeral	C	Cardinal Numeral		
		O	Ordinal Numeral		
		D	Collective Numeral		
—		—		—	
A	Adjective	P	Plural	P	Positive
		S	Singular	S	Superlative
—		—		—	

Coordinating Conjunction

S Subordinating Conjunction

F Foreign Word

R Preposition

M Modal

B Determiner

D Adverb

T Particle

G Negation

I Interjection

O Copula

W Question

X Unknown

Z Punctuation

ABSTRACT

Indonesian Morphological Tool, Morphind, is meant to make a proper morphological analysis before doing further automatic language processing. Morphind is applied to enrich raw Indonesian text with morphological information, the preprocessing stage of an Indonesian corpus. In this study, the data is obtained from children's stories in the website ceritaanak.org by taking 500 types of total 2101 types. The purpose of this study is to identify and classify the types of errors present in data processing using morphind program. In the analalysis I uses the method Introspective and Dictionary Indonesian (KBBI) to validate the analysis. The findings of this research suggest that there are still many aspects that can be improved about morphind. Recommendations are fixing the data base especially for OOV (out of vocabulary) and dictionary accuracy, improving the display for the Allomorph, and improving the algorithm for morpheme segmentation.

Keywords : *Morphology, Morphind, Automatic morphological analysis, error analysis*

1. INTRODUCTION

1.1. Background of the Study

The Indonesian language, is the official language of Indonesia. Language technology research in this language is quite encouraging lately but without a well-developed long-term plan. There are many language tools such as parsers, semantic analyzers and speech recognition tools. The Indonesian Morphological Tool, Morphind, is meant to make a proper morphological analysis before doing further language processing. Morphind is applied to enrich raw Indonesian text with morphological information, the preprocessing stage of an Indonesian corpus.

> "Morphind introduces a more finegrated tagset compared to IndMA its predecessor and provides output in the form of segmented morphemes as

> an added value. In addition, lemmata is also independently tagged for the purpose of lemmatization" (Larasati, 2011).

In the research I used Morphind program to analysis the data. Morphind produces analysis that only covers morphology phenomena; it does not handle syntax, but its output can be used as input to many other Natural Language Processing (NLP) tasks.

> "Morphind analyzes tokens as unigrams and does not take into account any neighbouring tokens. MorphInd does not return any syntactical functions on the analyses, although some functions are easily recognized by the word order or the clitics" (Larasati, 2011).

In writing this research I have two purpose of research: to use morphind program created by computer scientists to analyze words in the data and To identify and classify the types of errors contained in data processing using the morphind program.

In the research I collected data taken from children's stories as much as 7806 tokens. The authors limit the data used in project work as much as 500 types of total 2101 types, where the data has been processed by Unitex applications (Paumier, 2003).

The result of this analysis will help reader to know more about morphind program created by computer scientists to analyze words and identify the errors. I hope there will be other research an evaluation morphind program that can be used as the object of other research.

I found there are three similar research discuss Morphind program. First (Larasati, 2011) , Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. The result of this research MorphInd produces robust morphological information in the output format i.e. morphemic segmentation, lemma morpheme position, lexical category, and morphological feature. The new robust tagset with broader categorization that it uses is also suitable for a further language processing such as parsing. Second, (Rashel, Luthfi, Dinakaramani, & Manurung, 2014), Building an Indonesian Rule-Based Part-of-Speech Tagger. The result of this result is still many opportunity to improve the system, such as foreign language detector, expanding the language resources, and improving the tokenizer. Third (Afini, 2016), *Penerapan Analisis Morfologi untuk Penanganan kata berimbuhan pada Pos Tagger Bahasa Indonesia bebasis Statistik.* The result of this research is for Morphological Analysis by applying MorphInd can be used for the process of preprocessing the cutting of clitik on the shape of the form phrase. and to indicate MorphInd can be used to label the OOV word on IPOSTagger.

The differences of my research with the previous studies is the object of my research and the focus of this research. The data in this research will process with corpus linguistic method.

2. THEORETICAL FRAMEWORK

2.1. Morphology

Morphology is a study about word structure (Fromkin, Rodman, & Hyams, 2009). Morphology is a study that systematically learns about the internal structure of words (Haspelmath & Sims, 2010). Words can still be broken down into several more complex parts such as roots, affixes, stems and bases. Morphemes, In the science of morphology are used to identify smaller part of words. In this study I analyze the word using morphind program (Larasati, 2011) and identify the types of morphologycal errors. The word categories form-class words and structure-class words. In general, the form classes provide the primary lexical content; the structure classes explain the grammatical or structural relationship. Class classification of open class there are noun, verb, adjective and adverb, while class classification is closed there are determiner, pronoun, auxiliary, conjunction (or conjunct), interrogative, preposition, and particle.

2.2. Affix

Morphemes are the smallest grammatical units that have meaning. Free morpheme is a morpheme that can stand alone as a word, such as *tour* and *walk* (Fromkin, Rodman, & Hyams, 2009). Bound morpheme is a morpheme that can not stand alone. Suffixes and Prefixes are examples of bound morpheme. (Plag, 2002). Affix is a bound morpheme that attaches to bases (Plag, 2002, page. 100). Lexeme is a word in the abstract sense, lexemes are abstract entities that do not

have their own phonological form. While, the word form is a word in a concrete sense (Haspelmath & Sims, 2010, page. 15).

2.3. Allomorph

Allomorph is variant form of a morpheme but it does not change the meaning. Allomorph has different in pronunciation and spelling according to their condition (Plag, 2002, page. 124). It means that allomorph will have different sound, pronunciation or spelling in different condition. The condition depends on the element that it attaches to.

2.4. Token and Type

Token is simply defined as running word, while type is the distinct tokens. In type, same token are only counted one. However, when counting tokens they are counted depending on the occurences not the variation.

> "A token is any instance of a particular wordform in a text; while type is a particular, unique wordform – can tell us how large a range of vocabulary is used in the text" (McEnery & Hardie, 2012).

The exmple of token and type, Mary goes to Edinburgh next week and she intends going to Washington next month. The same word of the sntence are distinct tokens of a single types. The term word would be ambiguous between a 'type' interpretation and a 'token' interpretation, the ambiguity would be just the same as is exhibited by many other terms not specifically related to language. for example: tune a tune you heard this morning may be "the same" as one you heard yesterday.

3. RESEARCH METHODS

3.1. Morphind Survey

In this study I use morphind program to analyze the data. Morphind program is a program used to analyze linguistic data in the form of words. Morphind programs can only be used on Linux and Ubuntu computers. The program can not be used on windows, because before using morphind the program must install other programs such as, Pearl, Java, Foma and Subversion. This is quite challenging installation procedure.

3.2. Collecting and Category Data

In the collecting data, I extracted the types through the unitex program. Of the total data, the authors took as much as 20% from 2101 types, which is 500 types. The data is then analyzed using morphind program. The research data is categorized thematically based on the type of error.

3.3. Classifying Errors

The author used the Introspective method and Dictionary Indonesian (KBBI) reference to validate the analysis of morphind program. Errors are categorise thematically. Introspective method is a method of providing data by utilizing the language intuition of researchers who examine the language dikuasainya (mother tongue) to provide the necessary data for the analysis in accordance with the purpose of this research. this method is intended as an attempt to reveal the identity of the formation of language form that can allow

people to carefully determine certain lingual units whose unclear-lingual status is unclear.

4. RESULTS AND DISCUSSION

In this section, the writer will show the data that has been collected and processed by Morphind program. In morphology word can be divided into several smaller components (Chaer, 2008, p. 13).

The author will also categorize Morphind error analysis into several categories. A word, can be analyzed manually or by a morphological analyzer program. Consider manual analysis of example (1):

(1) *Membeli*

In example number (1), the word *membeli* can be analyzed manually and divided into smaller elements, that are two morphemes {*meN-}* and {*beli}*. See example analysis (2) :

(2) *{meN-} + {beli}*

Each morpheme can be described by its form and meaning. *{meN-}* is a morpheme, which is grammatically known as prefix. *{beli}* is a lexical morpheme. Example (3) is the analysis of the word *membeli* by morphind program :

(3) meN + beli<v>_VSA

In example (3), Morphind program analyzed word *membeli* into 2 morphemes and provided tags. The first morpheme is prefix *meN-*; the prefix

describes the type of allomorph prefix by writing the letter 'N' using capital letter, where the letter can turn into one of these nasals *(ng, m, n, ny).* (Chaer, 2008, p. 16).

The second morpheme *beli* is the lemma. In example number (3), *beli* is analyzed as a verb with tag symbol <v>. In the analysis, Morphind separates each morpheme with a'+' sign, and then _VSA tag, which is written in upper case to describe the full from *membeli*.

There are 3 different categories of errors from Morphind. They are tagset, allomorph, and morpheme break. These three errors will be described in more detail, in the next chapter.

The first error is the Tagset; this error will be divided into four parts such as Clitic, word or surfaceform tagset, entry code tagsets, and word that are not in the data base.

4.1. Tagset

Tagset shows part of speech tags or can be called as POS tag. It is a set of writing symbols used to show POS (Larasati, 2011). In syntax word is the smallest unit of analysis and can further be labelled as subject, object, predicate. In the morphological process, a word can be formed by affixation, composition, acronym and conversion (Chaer, 2008, p. 5).

Tagset errors in this corpus are divided into 4, i.e. Clitic tagset, Word or surfaceform tagset, entry tagset, and tagset that is not in the data base (OOV).

4.1.1. Clitic Tagset

Clitic is a free morpheme, where the underlying form is orthographically attached to another component or a word (Chaer, 2008, p. 5). Clitics in Indonesian are divided into 2 :

- Proclitic is a clitic that can be attached in front of the word base / base word.
- Enclitic is a clitic that can be attached to the back of the base word / base word.

ku- is one proclitic embedded in front of the base. And then, the example of an enclitic attached to the back of the word are *-nya, -mu,* Here is an example of clitic analysis that is not correct :

(4) ^lilin<n>_NSD+dia<p>_PS3$

(5) *Ucapkan keinginan Siti dulu sebelum meniup lilinnya ya*

(6) ^letak<n>_NSD+dia<p>_PS3$

(7) *Letaknya hanya berselisih lima rumah dari rumah Ranu*

(8) ^jendela<n>_NSD+dia<p>_PS3$

(9) *Bu kucing tak bisa masuk karena tak cukup jendelanya*

Examples (4), (6), (8) show *-nya* clitic analysis by Morphind program, while examples (5), (7), (9) shows their sentences contexts. Morphind analyzed *-nya* as a third person pronoun. In example (4) the word *lilinnya, -nya* is not a pronoun. *-nya* refers to a particular candle rather than in general. In example (6) the word *letaknya*, *-nya* is not a pronoun. *–nya* refers to the description of a

particular place rather than in general. And in example (8) the word *jendelanya*, -*nya* is not a pronoun. –*nya* refers to a particular window rather than in general. In short, -*nya* in those sentences are used to mark definiteness.

The following examples are word analysis by using a new tag for the definite article, because in the analysis of morphind program there is no tag of definite article. The author uses tag <e> for definite article. They are correct analysis of clitic :

(10) ^lilin<n>_NSD+nya<e>_ES3$

(11) ^letak<n>_NSD+nya<e>_ES3$

(12) ^jendela<n>_NSD+nya<e>_ES3$

Examples number (10), (11), (12) are correct analysis of –*nya* as definite articles. The author analyzes -*nya* not as a pronoun *dia*. In example (11), -*nya* in *letaknya* refer the word *lapangan basket* in the sentence.

4.1.2. Entry Tagset

The basic form is a base that can take morphological process to be a word (Chaer, 2008, p. 21). Some examples of base are *deras* (adjective), *cemas* (adjective), *bunyi* (noun). When affixiation or other word formation processes take place to the base we call the word full form. Morphind analyzis the full form and predicts in it is entry.

In this case the error is in giving tag to the base of a full form, which that call entry tag. Entry tag is written in lowercase, such as <n>, <v>, <p>. Examples of incorrect tags for the analysis are shown in examples (13) to (17) :

(13) ^dahulu<d>_D--$

(14) *Rara dan Riri sama sekali tidak menghiraukan rumah yang dahulu menjadi tempat tinggal mereka*

(15) ^bisa<n>_NSD$

(16) ^alang<n>_NSD+kah<t>_TSA$

(17) *Menurut Siti, seandainya apa yang selalu ia inginkan dapat segera terwujud dengan cara itu, alangkah bahagianya dia*

The word *dahulu* in the morphind program is wrongly analyzed. *dahulu* according to KBBI (Indonesian dictionary) is a noun, which should be symbolized by the tag <n>, while in the analysis of Morphind, *dahulu* is an adverb denoted by <d>. While in Morphind's analysis, *alangkah* a descriptive word denoted by <n>. In writing morphind writing is written with two morphemes that is alang<n> + kah<t>_TSA$.

(18) *Sang Paman hanya <u>bisa</u> memandangi keponakannya dengan geleng-geleng kepala*

Some words are ambiguous, For example *bisa* can be a noun or a verb. In example (15) the word *bisa* is analyzed as <n> by Morphind program. However

the correct tag is <v>, because in (18) *bisa* is a modal verb that describes its subject to be able to perform an activity.

In the next example (16), the word *alangkah* in the morphind program is wrongly analyzed. *alangkah* according to KBBI (Indonesian dictionary) is a adverb, which should be symbolized by the tag <d>, while in the analysis of Morphind, *damai* is an adjective denoted by <a>.

(19) ^dahulu<n>_N--$

(20) ^bisa<v>_VSD$

(21) ^alangkah<d>_DSA$

4.1.3. Word of Surface form Tagset

Word or surface form may take the same form like the base, or different form (undergo morphological process). In Morphind word tag is indicated by uppercase letters on the right side, which is different from entry tag, which is indicated by lowercase letters. Examples (22) to (27) show words that are given wrong tags :

(22) ^bahu <v> _VSA$

(23) *Shasa mencolek bahu Nia yang duduk di depannya*

(24) ^bungsu<a>_ASP$

(25) *Koko si bungsu juga berlatih berburu, ia mendapatkan tikus yang lebih besar dari ukuran tubuhnya*

(26) ^ramai<n>_NSD$

(27) *Teman-teman baru Ranu ramai berceloteh tentang aneka les yang harus mereka ikuti, les pelajaran sekolah sampai les piano atau les lukis*

The surface form tagset in example (22), (24), (26) are wrong. The word *bahu* is a noun. But on the results of the analysis by Morphind program shows the word *bahu* is a type of verb (VSA). The tag of the word *bahu* in (22) is ideally (NPS) not (VSA). In example (24) the wrong tag is for the word *bungsu*. The word *bungsu* is noun. But on the results of the morphind analysis program the word *bungsu* is a type of adjective (ASP). The correct tag in example (24) should be (NPS). And then the last example (26) the wrong tag is for the word *ramai*. The word *ramai* is adjective. But on the results of the morphind analysis program the word *ramai* is a type of noun (NSD). The correct tag in example (26) should be (ASD).Consider the correct tags for the errors in examples (22) – (27) in example (28) – (30) :

(28) ^bahu<n>_N--$

(29) ^bungsu<n>_NSP$

(30) ^ramai<a>_ASD$

4.1.4. Tagset that are not in the Data Base

In the analysis, this type of error analysis occurs both in entry or surface form tags. The symbol of the error is with <x>, which means unknown or undefined. Consider examples (31) - (36) :

(31) ^betah<x>_X--$

(32) *Tidak betah karena merasa selalu diperintah*

(33) ^becak<x>_X--$

(34) *Pak Somad mengayuh becak tuanya yang tak berpenumpang menuju rumah*

(35) ^capek<x>_X--$

(36) *Kok diam aja San, capek ya*

From example (31), (33), (35), the tag <x> and X are given because Morphind cannot detect the word in the data base. *betah* is supposed to be an adjective. *becak* here belongs to a noun in (33). And in (35) the word *capek* here belongs to an adjective. *betah, becak and capek* are not in the data base. Therefore, they are given with the <x> symbol. Example (37) – (39) show the correct analysis of *betah, becak and capek* respectively :

(37) ^betah<a>_ASP$

(38) ^becak<n>_NSD$

(39) ^capek<a>_ASP$

4.2. Allomorph

Allomorph is a term used in the field of linguistics for the variation of a form of morpheme (Chaer, 2008, p. 15). Therefore, allomorph is a realization of a real or existing morpheme. In data analysis, I found some examples of words that have an allomorph error in writing, among others:

(40) ^ber+gegas<v>_VSA$

(41) *Shasa bergegas membuka pintu pagar*

(42) ^ber+semangat<n>_VSA$

(43) *Ketika mobil berhenti di depan rumah, Shasa membuka pintu mobil dengan bersemangat*

(44) ^ber+doa<n>_VSA$

(45) *Biar ia bisa berdoa sambil meniup lilin di atas kue ulangtahunnya*

On analyzed number (40), (42), (44) with the Morphind program the word *bergegas, bersemangat, berdoa* is wrong in writing the morpheme prefix. The morpheme is an allomorph where the writing of the R must be capitalized. The correct samples of the analysis are :

(46) ^beR+gegas<v>_VSA$

(47) ^beR+semangat<n>_VSA$

(48) ^beR+doa<n>_VSA$

The above example is an example that prefix writing at the beginning of sentence is written correctly. The prefix is an allomorphic form so that in its writing, the allomorph must be written using capital letters i.e. beR-.

4.3. Morpheme Break

Morpheme break is a kind of mistake in delimiting words and the overall error of the word where the word should be written separately. This morpheme break has several categorizations or combinations of affix, suffix, and confix.

4.3.1. Morpheme Break Surface form or Boundaries

This section will analyze the error of a basic word that includes verbs, properties, objects and others as a whole where the word has been affixed either affix or suffix.

(49) ^berkacamata<x>_X--$

(50) *Hai! siapa seorang anak kurus berkacamata*

(51) ^di+perbaiki<v>_VSP$

(52) *Sambil menunggu mobil diperbaiki, Santi berjalan-jalan di sekitar penginapan bersama Tante Lusi*

(53) ^bercat<x>_X--$

(54) *Ring basketnya juga masih utuh dengan tiang bercat merah terang*

In example (49), (51), (53) these words are not delimited correctly. Each words in the examples are affixed. There Fore they are supposed to be delimited properly. The analysis by Morphind program is still wrong. Therefore, the author wrote re-analysis of the three words such as shown in examples (55) – (57):

(55) ^ber+kacamata<n>_V--$

(56) ^di+per+baik<a>+i_NSP$

(57) ^ber+cat<n>_V--$

kacamata that originally standalone without a prefix called a noun, in which the word is attached with prefix beR- that turn the word into *berkacamata* into a kind of verb, which *beR*- means to use. *berkacamata* becomes a verb that has the

meaning of using glasses. In example (51) *baik* is prefixed by prefix *di-* and *per-* and suffixed suffix *-i*. Like (49) in (53), bercat is analysed as a standalone word. However, it should beR- has to be delimited from cat, because it is a prefix. The table 1.1 in the apendix summarizes the finaling in this research.

5. CONCLUSION

The results of research on data processing using morphind program shows that the error rate is 39% of the total data in the analysis of 500 words. Larasati, Kuboˇn, & Daniel (2011) evaluated Morphind with different variables, which are: Test Set and Metric. This two variables are commenly used in computer scince field. Therefore they are not dirrectly comparable here. However they believe that Morphind is better than it is predecessor, IndMA.

There are 3 types of errors from our analysis, which can proportionally described as: 64% Tagset, 19% allomorph, and 17% morpheme break errors. Tagset error can be subdivided 4 types: Clitic, Entry, POS, OOV. Among the errors, the largest error is tagset, while the smallest number of errors occur in morpheme break.

Based on the results of the research, the researcher suggests that there are still many aspects that can be improved about morphind. Recommendations are fixing the data base especially for OOV (out of vocabulary) and dictionary accuracy, improving the display for the Allomorph, and improving the algorithm for morpheme segmentation. Chart 1 and 2 in the appendix are diagrammatic views of the error rate by the Morphind program.

REFERENCES

CeritaAnak.org. (2015, Desember 14). Retrieved Desember 14, 2015, from Ceritaanak.org.

Afini, U. (2016). Penerapan Analisis Morfologi Untuk Penanganan kata berimbuhan pada pos tagger Bahasa Indonesia Berbasis Statistik. Thesis, 72.

Chaer, A. (2008). Morfologi Bahasa Indonesia. Jakarta: PT Rineka Cipta.

Fromkin, V., Rodman, R., & Hyams, N. (2009). An Introduction to Language.

New York: Michael Rosenberg.

Haspelmath, M., & Sims, A. D. (2010). Understanding Morphology. London: An Hacette UK Company.

Larasati, S. D. (2011, August). Larasati, S. D., Indonesian morphology tool (morphind): Towards an indonesian corpus. International Workshop on Systems and Frameworks for Computational Morphology (pp. 119-129). Berlin: Springer Heidelberg.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics. New York: Cambridge University Press.

Paumier, S. (2003). Unitex 3.0 User Manual. Paris: Université Paris-Est Marne-la-Vallée.

Plag, I. (2002). Word-formation in English. Cambridge: Cambridge University Press.

Rashel, F., Luthfi, A., Dinakaramani, A., & Manurung, R. (2014). Building an Indonesian Rule-Based Part-of-Speech Tagger. 4.

APPENDIX

1. Table 1.1 Error Summary

2. Diagram 1.1 overall error rate

3. Diagram 1.2 error proportion

1. Appendix 1

no	Error				Error	Corrected
1	Tagset	Clitic	proclitic			^aku<p>_PS1+bilang<v>_VSA$
						^langkah<n>_NSD+dia<p>_PS3$
			enclitic	25	^lilin<n>_NSD+dia<p>_PS3$	^lilin<n>_NSD+nya<t>_PS3$
					^letak<n>_NSD+dia<p>_PS3$	^letak<n>_NSD+nya<t>_PS3$
					^jendela<n>_NSD+dia<p>_PS3$	^jendela<n>_NSD+nya<t>_PS3$
		Entry		66	^dahulu<d>_D--$	^dahulu<n>_N--$
					^bisa<n>_NSD$	^bisa<v>_VSD$
					^bisik<v>_VSA$	^bisik<n>_NSA$
		POS			^bahu<v>_VSA$	^bahu<n>_N--$
					^bungsu<a>_ASP$	^bungsu<n>_NSP$
					^harus<v>_VSA$	^harus<d>_DSA$
		OOV		40	^betah<x>_X--$	^betah<a>_A--$
					^becak<x>_X--$	^becak<n>_N--$
					^capek<x>_X--$	^capek<a>_A--$
2	Allomorph				^ber+gegas<v>_VSA$	^beR+gegas<v>_VSA$
				96	^ber+semangat<n>_VSA$	^beR+semangat<n>_VSA$
					^ber+doa<n>_VSA$	^beR+doa<n>_VSA$
3	morpheme break				^berkacamata<x>_X--$	^beR+kacamata<n>_V--$
				26	^di+perbaiki<v>_VSP$	^di+per+baik<a>+i_NSP$
					^bercat<x>_X--$	^ber+cat<n>_V--$
4	Overall token sample			500		

Figure 1. table 1.1 Error Summary

2. Appendix 2

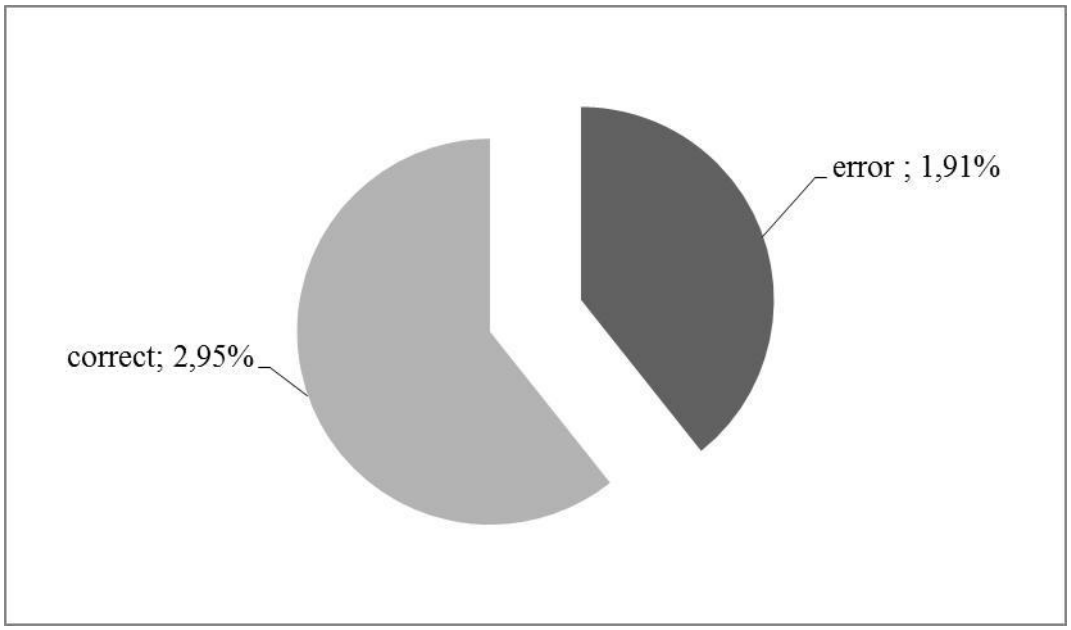

Figure 2. Diagram 1.1 overall error rate

3. Appendix 3

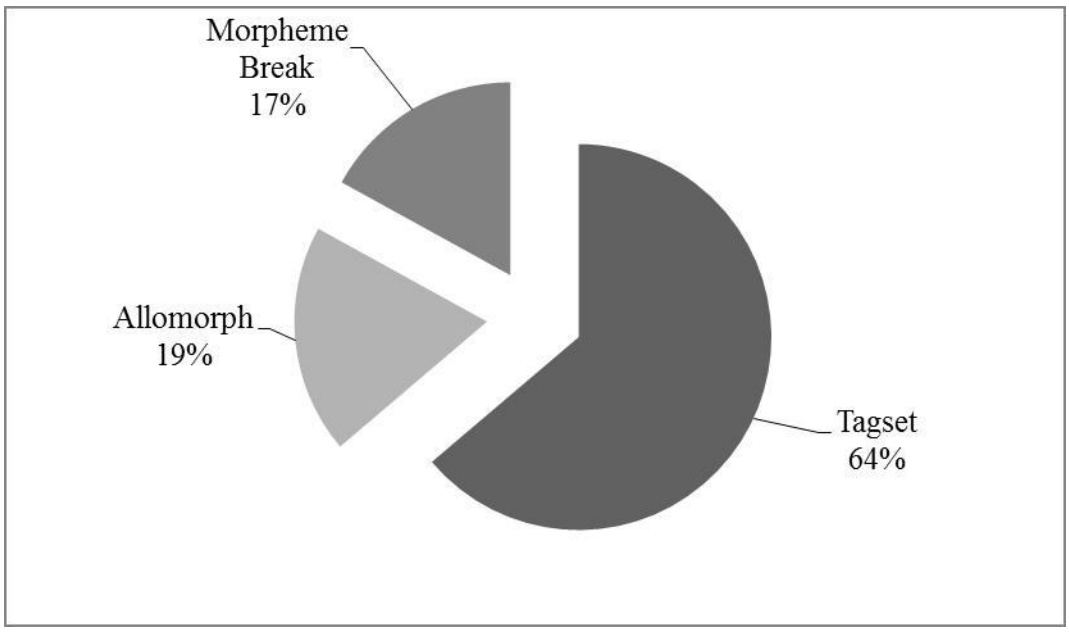

Figure 3. Diagram 1.2 Error proportion