NANYANG TECHNOLOGICAL UNIVERSITY

SCHOOL OF HUMANITIES AND SOCIAL SCIENCES

Creating derivational morphology links in Wordnet Bahasa

Muhammad Zulhelmy bin Mohd Rosman

Matriculation No.: U1030148D

Supervisor: Associate Professor Francis Bond

Date:18 November 2013

A Final Year Project submitted to the School of Humanities and Social Sciences, Nanyang Technological University in partial fulfilment of the requirements of the Degree of Bachelor of Arts in Linguistics and Multilingual Studies.

Declaration of Authorship

I declare that this assignment is my own original work, unless otherwise referenced, as defined by the NTU policy on plagiarism. I have read the NTU Honour Code and Pledge.

No part of this Final Year Project has been or is being concurrently submitted for any other qualification at any other university.

I certify that the data collected for this project is authentic. I fully understand that falsification of data will result in the failure of the project and/or failure of the course.

_ _ _ _ _ _ _ _ _ _ _ _ _ _	_ _ _ _ _ _ _ _ _ _ _ _ _ _	_ _ _ _ _ _ _ _ _ _ _ _ _
Name	Signature	Date

Acknowledgements

I would like to express my gratitude and appreciation to my supervisor Associate Professor Francis Bond for his guidance and support throughout my Final Year Project and my University life as a whole. It is through him that I was introduced to the wonderful and intriguing world of Computational Linguistics. It is truly impressive that a non-native speaker would be interested in creating and developing the Wordnet Bahasa which in the long run will help the preservation of the Malay and Indonesian languages. From the start of this project, his inputs allowed me to develop this project well and enabled me to understand the subjects of morphology and word sense disambiguation better.

I would like to thank my friends and fellow course mates for their support throughout this period. Even though their input were not required for this project, their encouragement allowed me to push through the tough times.

Contents

Abstract

Derivational morphology links are created for the Wordnet Bahasa, a combined Indonesian and Malay online lexical dictionary (Nurril Hirfana, Suerya, & Bond, 2011). The focus was to link root words to affixed words as affixation is one of the more apparent word formation processes in *Bahasa Melayu*. MorphInd, an Indonesian morphological analyser (Larasati, Kubon, & Zeman, 2011), is used to breakdown affixed words into their root form and affixes. Using Python 2.7 with NLTK, a raw mapping is done by matching the analysed words to the root forms. The derivational links in the Princeton Wordnet (PWN) are used to verify if the same links exist in Wordnet Bahasa. Redundant links are removed by the Part-of-Speech (POS) filter and Semantic Super Type filter. The links are then disambiguated using the Lesk algorithm, where the definitions and other components of the sense (e.g. hypernyms, hyponyms and examples) are compared for their similarity. However, the disambiguation process is rendered ineffective because of the high amount of errors still existing in Wordnet Bahasa. The derivational links are released as a separate file and only those with similar derivational links to PWN are added into Wordnet Bahasa. Erroneous entries that were identified using MorphInd are removed from Wordnet Bahasa.

1 Introduction

Word Sense Disambiguation (WSD) is a process of identifying the most appropriate sense for words with multiple meanings. Words are usually ambiguous in nature (Lesk, 1986; Baldwin et al., 2010). For example, the English word ***bank*** has multiple meanings. It can either be “a sloping land” or “a financial institution”, depending on the context that the word is in. With a more robust WSD process, the most appropriate meaning can be tagged to the correct lexeme. However, having a standard framework for this process is difficult as different languages have different morphological and syntactical structure (Bosch, Fellbaum, & Pala, 2008). Thus, different methods are usually applied to comply with the uniqueness of each language. Adding morpho-semantic relations into a computer database such as an online dictionary is not an easy task. As some links are ambiguous, creating a fully automated program would not accurately reflect the links as in natural language (Bilgin, Cetinoglu, & Oflazer, 2004; Ranaivo-Malancon, 2004). These linkages between words would be easily understandable to humans through usage and experience. However, the computer does not have this ability and therefore, a WSD process needs to be done to ensure that a precise representation of the natural language will be shown (Ranaivo-Malancon, 2004).

This study aims to create one of the morpho-semantic relations, the derivational morphology links, in Wordnet Bahasa - a Wordnet created based on the Princeton Wordnet for the Malay and Indonesian languages (Nurril Hirfana, Suerya, & Bond, 2011). As the derivational links will encompass the whole Wordnet Bahasa, WSD process is needed to identify the validity of the links. In general, wordnets do not include derivational links as part of it basic set up as morphologically derived words are assumed to be regular. Therefore, affixes are also assumed to hold distinct information about the derived word’s meaning and part of speech (Fellbaum, Osherson, & Clark, 2009). However, as affixation in *Bahasa Melayu* is not regular, the derivational links must be created to show the construction of the derived words and their affixes. With the derivational morphology links added into Wordnet Bahasa, a more representative of the *Bahasa Melayu* structure will be shown. This study will only focus on the single word sense that is available in Wordnet Bahasa. As this is the first instance of creating derivational link in Wordnet Bahasa, it would be best to investigate the accuracy of the single words links before embarking into multi words.

The following section will describe the types of derivational morphology that exist in *Bahasa Melayu* (**S** 2). The current state of Wordnet Bahasa will be discussed followed by the description of MorphInd and how can it be useful in the creation of the derivational link (**S** 3). The methodology of the creation and disambiguation of the derivational links is thoroughly explained (**S** 4). The results of the mapping are presented (**S** 5) and discussed (**S** 6).

2 Derivational Morphology in *Bahasa Melayu*

The Malay Language (*Bahasa Melayu*) is a part of the Austronesian language family and is mostly spoken in the Malay Archipelago region. This consists of Indonesia, Malaysia, Brunei and Singapore where it is an official language in all of these countries. *Bahasa Melayu* is also an unofficial language in Thailand and Christmas Island. Currently, the total number of speakers stands at 163 million people comprising of 23 million native speakers and 140 million L2 speakers.[1] However, Kozok (2012) stated that because Malay and Indonesian are institutionalised as national languages after independence, the total number of speakers should be closer to 215 million. *Bahasa Melayu* has two writing systems, the Latin script and Arabic (known as *Jawi*). From the 17th century onwards, the Latin script is much more prominently used through the influence of the Dutch and British.[2]

Bahasa Melayu is an agglutinative language where new words are created by the manipulation of root words (Sneddon, 2010). This is done by three main methods: affixation, compounding and reduplication. Affixation is done by attaching affixes onto a root word. As compared to root words, affixes do not have standalone meaning. For example, the derived form ***berbatas*** ”having a boundary” is created by adding the affix *ber-* to the root form ***batas*** “limit". Compounding is achieved by joining two root forms to create a new word with a new meaning. For example, the word ***kakaktua*** "parrot" is from two root forms ***kakak*** "sister" and ***tua*** "old". Reduplication on the other hand is a repetition of the root word. The root word ***tiba*** "arrive" can be reduplicated to form ***tiba-tiba*** "suddenly". In this study, the focus of the derived morphology linkage is affixation as it is one of the more apparent word formation in *Bahasa Melayu* (Macdonald et al., 1976; Sneddon, 2010).

2.1 Affixes in *Bahasa Melayu*

Bahasa Melayu has all the four forms of affixes; prefixes, suffixes, infixes and circumfixes. This section will elaborate on the four different forms and how they are used to produce a new word. Most affixations in *Bahasa Melayu* are productive and create either derivational or inflectional forms (Macdonald et al., 1976; Sneddon, 2010).

[1] `http://archive.ethnologue.com/16/show_language.asp?code=ind`

[2] `http://www.omniglot.com/writing/malay.htm`

Prefix	POS→DPOS	Root word	Meaning	Derived Form	Derived Meaning
ber-	N→V	*anak*	child	*ber-anak*	giving birth
per-	N→N	*tiga*	three	*per-tiga*	third
peN-	V→A	*malas*	lazy	*pe-malas*	lazy bones
ter-	V→A	*tarik*	pull	*ter-tarik*	attracted
meN-	N→V	*tari*	a dance	*me-nari*	to dance
di-	N→V	*kapur*	chalk	*di-kapur*	to whitewash
ke-	V→N	*hendak*	want	*ke-hendak*	desire
se-	V→R	*lalu*	pass	*se-lalu*	always

Table 1: Examples of Prefixes in *Bahasa Melayu* with Derived Forms

2.1.1 Prefixes

Examples of the prefixes in Malay are shown in Table 1. Different prefixes in *Bahasa Melayu* have different functions. For example, the prefix *ber-* derives verbs from noun roots while the prefix *ke-* derived nouns from verb roots (Macdonald et al., 1976; Sneddon, 2010). Table 1 shows examples of how each prefix creates a new derived form of the root word. the prefixes *ber-*, *per-* and *ter-* only create derivational forms while the rest of the prefixes create both derivational and inflectional forms (Macdonald et al., 1976). There are also borrowed prefixes from other languages; such as *anti-*, *mikro-*, *ultra-* and *super-* from English (Warrillow-Williams, 2002), and *pra-* from Sanskrit (Sneddon, 2010). Affixes can also be derived from another affixed form such as the prefix *peN-* where the noun form is derived from the root word with prefix *meN-* (Macdonald et al., 1976) as seen in Table 2.

The allomorphic forms prefix *meN-* and *peN-* correspond to the phonetic environment of the stem. For example, if the stem begins with a bilabial stop (eg /b/, /p/) then the allomorphic form *mem-* will be used (Macdonald et al., 1976; Sneddon, 2010). These allomorphs appears in complementary distribution (Sneddon, 2010). For *meN-* and *peN-*, the allomorphic forms does assimilate to the root word. A single affix *meN-* has various forms; *men-*, *mem-*, *meng-*, *me-*. While the prefix *peN-* has the allomorphic forms; *pen-*, *pem-*, *peng-* and *pe-*. The allomorphs of the prefix *meN-* are shown in Table 3.

A combination of two prefixes can occur in *Bahasa Melayu* (Ranaivo-Malancon, 2004; Nakov & Ng, 2011). The prefix *meN-* and *per-* can be combine with the root word ***buat*** “create” to produce the word ***memperbuat*** “commit”. The prefix *per-* derives an instrument or agent noun while the prefix *meN-*

Prefix	POS→DPOS	Root word	Meaning	Derived Form	Derived Meaning
peN-	V→N	*me-nulis*	write	*pe-nulis*	writer
peN-	V→N	*meng-gosok*	rub	*peng-gosok*	polisher
peN- -an	V→N	*me-luas-kan*	to widen	*pe-luas-an*	broadening
peN- -an	V→N	*meng-irim-kan*	send	*peng-irim-an*	delivery

Table 2: *peN-* Derived from Root Words with *meN-*

Allomorphic form	POS→DPOS	Root word	Meaning	Derived Form	Derived Meaning
me-	V→V	*lihat*	see	*me-lihat*	seeing
men-	V→V	*cari*	find	*men-cari*	search
mem-	V→V	*buka*	open	*mem-buka*	to open
meng-	V→N	*gambar*	picture	*meng-gambar*	to describe

Table 3: Allomorphs of the Prefix *meN-*

indicates an active voice, However, when they are combined, the meaning of the combined prefix is not equal to the sum of the two previous prefixes. Unlike their singular forms, the prefix *memper-* derives a transitive verb which is also causative (Macdonald et al., 1976). As the pattern in *Bahasa Melayu* is regular, the double prefix can be analysed as one unit (Ranaivo-Malancon, 2004; Sneddon, 2010) . This is to ensure that they remain separate from the individual prefix form. Examples of the double prefixes are in Table 4.

Double Prefix	POS→DPOS	Root word	Meaning	Derived Form	Derived Meaning
mem-per-	N→V	*kaya*	rich	*meN-per-kaya*	to enrich
di-per-	N→V	*sembah*	homage	*di-per-sembah*	to pay homage
ber-per-	N→V	*lembaga*	establishment	*ber-per-lembaga-an*	having constitution
ber-ke-	N→V	*liar*	wild	*ber-ke-liar-an*	to stray

Table 4: Double Prefixes in *Bahasa Melayu*

Suffix	Root word	POS→DPOS	Meaning	Derived Form	Derived Meaning
-an	*kotor*	A→N	dirty	*kotor-an*	dirt
-kan	*hamil*	N→V	pregnant	*hamil-kan*	to impregnate
-i	*luka*	N→V	bruise	*luka-i*	to hurt
-nya	*akibat*	N→R	result	*akibat-nya*	consequently
-anda	*ibu*	N→N	mother	*ibu-anda*	dear mother
-man	*seni*	N→N	art	*seni-man*	artist
-wan	*usaha*	N→N	effort	*usaha-wan*	entrepreneur
-at	*akhir*	V→N	end	*akhir-at*	the afterlife

Table 5: Examples of Suffixes in *Bahasa Melayu* with Derived Forms

2.1.2 Suffixes

There are eight forms of suffixes in *Bahasa Melayu* and examples are shown in Table 5. Similar to the prefixes, each suffix has its own function. For example, the suffix *-an* derive a noun ***kotoran*** "dirt" from an adjectival root ***kotor*** "dirty". It can also derive nouns from a verb root (e.g. ***makan*** "to eat" to ***makanan*** "food") or even from a noun root (e.g. ***ruang*** "space" to ***ruangan*** "a generic room"). Unlike the prefixes, suffixes are more regular in their orthography. There are no occurrences of double suffixes and allomorphs in any of the suffixes. However as shown in Table 6, clitics and particles can be added to the end of words as case markers (Ranaivo-Malancon, 2004). Clitics and particles differ from affixes in their functions as new words are not derived from them. They also vary in terms of position where words with prefixes can have clitics or particles added to it (Ranaivo-Malancon, 2004). For example, the derived word ***makanan*** can have the particle *-ku* attached to it to form ***makananku*** "my food". However, prefixes cannot be attached to words that contain particles or enclitics (Ranaivo-Malancon, 2004). As clitics and particles do not create new words, they must not be included as part of affixation in *Bahasa Melayu*.

2.1.3 Circumfixes

The scope of circumfix in *Bahasa Melayu* can be broad to include all the derived words that are formed with both a prefix and an affix in them (See, 1980; Verhaar, 1984; Nakov & Ng, 2011). The definition of circumfix can also be narrow, where a circumfix is only considered when the derived word

Enclitic/Particles	Marker			
	Possessive	Objective	Subjectve	Definite
–kau	x	x		
–ku	x	x		
–mu	x	x		
–nya	x	x	x	x
–kah	x			

Table 6: Examples of Enclitic and Particles in *Bahasa Melayu* with Derived Forms

Root form	With Suffix *-an*	With Prefix *ke-*	With Circumfix *ke- -an*
boleh	**bolehan*	**keboleh*	*kebolehan*
lapang	*lapangan*	**kelapang*	*kelapangan*

Table 7: Circumfix in *Bahasa Melayu*

is invalid if either of the prefix or suffix was removed (Macdonald et al., 1976; Ranaivo-Malancon, 2004; Sneddon, 2010). As shown in Table 7, both ***boleh*** and ***lapang*** do have a derived form with the circumfix *ke- -an*. However, only the word ***lapang*** has the derived form with the suffix *-an* while ****bolehan*** is considered to be invalid. As such, the word ***kebolehan*** is derived from the root word ***boleh*** with the circumfix *ke- -an* while the word ***kelapangan*** is derived from the prefix *ke-* and the word ***lapangan***; which in itself comprises of the root word ***lapang*** and the suffix *-an*.

To create a standardised categorisation, this paper will use the definition by See (1980), Verhaar (1984) and Nakov & Ng (2011) and consider all words containing a combination prefix and suffix as a circumfix. Looking at the examples in Table 7, one could argue that the meaning of ***kelapangan*** “spaciousness” is much closer to ***lapang*** “spacious” than ***lapangan*** “field”. Therefore, the root word for ***kelapangan*** should be ***lapang***. This also enforced the fact that most forms of circumfix are different from the stand-alone prefix and suffix as stated in Macdonald et al. (1976). For example, the prefix *per-* derive agent or instrument nouns from verbs while the *per-* with the suffix *-an* derived nouns showing the process of the action from the root verb.

2.1.4 Infixes

Infixation in *Bahasa Melayu* occurs after the first consonant sound of the base word (Macdonald et al., 1976). The most common infix is *-le-* with words such as ***lelelaki*** “men” and ***kelelawar*** “bat” formed by adding that infix to the words ***lelaki*** “man” and ***kelawar*** “bat” respectively. However, studies have shown that infixation is not productive in both Malay and Indonesian as compared to other related language like Tagalog or Javanese (Macdonald et al., 1976; See, 1980). Studies also agreed that infixation is not part of the modern Indonesian language and new examples are hard to find (Macdonald et al., 1976; See, 1980; Sneddon, 2010; Nakov & Ng, 2011) . As infixation is not productive, native speakers would usually recognise it as part of the root word and not as a separate affix (Macdonald et al., 1976). Infixation can also be argued as partial reduplication like in the example of ***lelelaki*** “men” where one can argue that it is a partial reduplication of *le-* (Ranaivo-Malancon, 2004; Sneddon, 2010) . Also, the meaning of the root and derived form can be similar as in the case of ***kelelawar*** and ***kelawar*** where both words refer to the meaning “bat” (See, 1980). Therefore, as there are compelling evidence against the existence of infixation in *Bahasa Melayu* , and that they are now discussed under root formation (Macdonald et al., 1976), it would be best to consider that infixation does not exist in modern Indonesian and Malay.

2.1.5 Transparency of Root Word to Derived Forms

In general, the derived forms can be easily linked to the root word via its definition or meaning. As stated previously, the prefix *peN-* derives the person or instrument that does the action as described by the root verb with the prefix *meN-*. For example, ***penulis*** “writer” is derived from the word ***menulis*** “to write”. With the suffix *-an*, it shows the activity that arises from the verb, as in the word ***penulisan*** which means “writing” (Macdonald et al., 1976). While the prefix *ter-* shows an accidental or uncontrollable action of the root verb (Macdonald et al., 1976). For example, the derived word ***tertutup*** “unintentionally closed” comes from the root word ***tutup*** “closed”. As seen from the examples, the meaning of the derived form would be close to the root form. Thus, as the definitions in dictionaries are comprehensive enough to describe each word, comparing the definition of the root word to the derived forms should suffice to determine the validity of the derived forms.

3 Resources

3.1 Current state of Wordnet Bahasa

Wordnet Bahasa was created as an online lexical dictionary based on the Princeton WordNet (PWN) created by Fellbaum (1998). The database contains the complete wordnet structure; with information such as lexical relations (e.g. hypernym-hyponym) and meanings of particular synsets (mainly in English). As the name suggests, the Wordnet Bahasa combines both Bahasa Indonesia and Malay (Nurril Hirfana, Suerya, & Bond, 2011). The two languages are joined under the generic Malay language code (**msa**) as they are mutually intelligible.[3]

Currently the Wordnet Bahasa contains 142,488 and 119,152 senses for Bahasa Indonesia and Malay respectively. The senses were taken from dictionaries such as the French-English-Malay dictionary **FEM**, Kamus Melayu-Inggeris **KAMI**, and aligned with wordnets from other languages (Nurril Hirfana, Suerya, & Bond, 2011). The amount of senses that is available in Wordnet Bahasa would be sufficient for formal links to be added as the extend of derivational morphology in the senses, will be explicit. Thus, it would be vital to create the link between the root words and the derived forms to show a better representation of how the words in *Bahasa Melayu* are linked through their morphology (Nurril Hirfana, Suerya, & Bond, 2011). .

3.1.1 Derivational Morphology in Other Wordnets

In the Princeton Wordnet, the derivation link was first created for noun-verb pairs (Fellbaum, Osherson, & Clark, 2009). For example, the noun-verb pair of ***act-actor*** in the Princeton WordNet was connected via the derivational related forms link. The first focus was to look at the nouns derived from verbs via the suffix *-er*/*-or* and then proceeded to other derivational nouns which ends in suffix such as *-al*, *-ment* and *-ion*. As shown in Table 8, the links were then categorised into eleven semantic categories (Fellbaum, Osherson, & Clark, 2009). However, in the current version of the Princeton Wordnet, the derivational links are only categorized under the part of speech and not its semantic categories. A comparative study of the English, Czech and Zulu wordnets was done in Bosch et al. (2008) where the categorization and spread of derivation morphology was found to be more apparent in the Czech and Zulu as the languages themselves contains more derived words.

[3] A relationship between languages in which speakers of related language varieties can easily understand each other without the need to learn each other's language (Sugiharto, 2008).

Semantic Categories
Agent
Instrument
Inanimate agent/Cause
Event
Result
Undergoer
Body part
Purpose
Vehicle
Location

Table 8: List of Semantic Categories used in Princeton Wordnet (Fellbaum, Osherson, & Clark, 2009)

The derivation link was also proposed in the wordnets pertaining to Slavic languages such as Bulgarian (Koeva, Krstev, & Vitas, 2008) and Serbian (Koeva, 2008). As derivational relations in the Slavic languages are different than in English, creation of a semi-automatic derivational link generator was proposed in this paper. In Bulgarian, there is a pattern of nouns derived from adjective via suffix *-oct* or *-ost*. As such, this link can be made by matching the root form of the noun to the adjective . The process was not fully automated as some complex issues arose from it. For example, some of the derived words losses their gender and thus, the word category changes (Koeva, 2008).

Bilgin et al. (2004) and Mititelu (2012) stated that using the semantic relations in another language are an effective way of creating derivational links for the target language. Importing derivational links from other languages will allow for a quicker set up (Bilgin, Cetinoglu, & Oflazer, 2004). In the final stage, Bilgin et al. (2004) suggested for the derived morphology links to be categorized into the semantic categorises or the result of the derived words.

During the creation of the derivational link in Romanian wordnet, the links were validated via an 'automatic heuristics' process (Mititelu, 2012). As most of the root and derived forms in Romanian must be in the same parts-of-speech (POS), those links which does not comply were considered as invalid. Exceptions to the invalid data were picked out manually. A sample of 1000 links were chosen and manually validated. The precision and recall of the links were tabulated and used to validate all the links that were created in the Romanian wordnet. Precision is the percentage agreement of relevant links

Synset	MorphInd Analysis	Meaning
mengawal	*meN+**kawal**<v>*	to control
bercepat	*ber+**cepat**<v>*	travel rapidly
urusan	***urus**<v>+an*	appointment
sebabnya	***sebab**<v>+dia*	originate
kejujuran	*ke+**jujur**<a>+an*	honesty
memperhalus	*meN+per+**halus**<a>*	refine

Table 9: Examples of MorphInd Process

while recall is the percentage of valid links that were retrieved from the sample data (Mititelu, 2012). Mititelu (2012) argued that this method is the most effective way in creating the basic derivation link in a language.

Therefore, adding derivational links in Wordnet Bahasa will help to increase its usefulness as the morpho-semantic relations in *Bahasa Melayu* would be highlighted more clearly. The creation of derivational links in Wordnet Bahasa will be a modification of the methods created in the other wordnets. The derivational links from another language (for example English) will be used in building the links for *Bahasa Melayu*. As this is the first attempt in creating the derivational link in Wordnet Bahasa, the validation of the link will be done semi-automatically similar to the process done in Mititelu (2012).

3.1.2 MorphInd

MorphInd is an open-source, Linux-based morphology tool which analysed the morphological structure of Indonesian words (Larasati, Kubon, & Zeman, 2011). This included derived words such as affixation, reduplication and compounding. MorphInd was an improvement over a previous morphological analyser for Indonesian (IndMA) with a wider coverage of derivational and inflectional morphology and also a much bigger database for its dictionary (Larasati, Kubon, & Zeman, 2011). Table 9 shows examples of how the MorphInd tool analysed a given word to its root word and affixes.

The Part of speech categorisation for Wordnet and MorphInd differs slightly, with nouns in MorphInd separated in different categories as shown in Table 10. However, as MorphInd will only be used in this study to breakdown the morphology of the senses in Wordnet Bahasa, the difference in categorisation will not have any effect in this process. Ultimately, the Wordnet Bahasa categorisation will be used during the linking process.

Wordnet	MorphInd
Noun (n)	Noun (NN)
	Proper noun (NNP)
	Personal pronoun (PRP)
	Foreign word (FW)
Verb (v)	Verb (VB)
Adjective (a)	Adjective (JJ)
Adverb(r)	Adverb (RB)

Table 10: Comparison of Wordnet Part of Speech Categories with MorphInd

Due to MorphInd's high coverage of analysed words on an Indonesian corpus (at 84.5% out of 10,000 sentences) (Larasati, Kubon, & Zeman, 2011), it would be appropriate to include it as the first step of analysing the derivational morphology in Wordnet Bahasa. MorphInd will be used to break down the derived words into its root word and the affixes. The linking of the derived words to its root form will be more efficient with this process. As the dictionary used in MoprhInd is different from Wordnet Bahasa, not all of the senses in Wordnet Bahasa is expected to be be processed by MorphInd.

4 Methodology

The derivational links in Wordnet Bahasa are created by mapping the derived words to its root words. Both the Indonesian and Malay data under Wordnet Bahasa were used. The derived forms must be identified and processed to its root forms and affixes before the links can be created. This is done by using MorphInd. The link creation process and other secondary processes (for example; cleaning of data) is done using Python 2.7 with NLTK (Bird et al., 2009).

4.1 Analysis of Wordnet Bahasa using MorphInd

Even thought MorphInd was used only on the Indonesian Corpus and was created to analyse the Indonesian language (Larasati, Kubon, & Zeman, 2011), the mutual intelligibility of Malay and Indonesian would allow for the program to be used for Malay. As shown in table 9, the output given by MorphInd splits the derived word into the root form and affixes. For example, the word ***urusan*** "dealings" will be separate into the root form ***urus*** “deals” and the verbal suffix *-an*. The only exception for MorphInd is for the suffix/enclitic *-nya*, where the tag *+dia* is used.

4.2 Mapping

The root words and the derived forms were matched by stripping off the affixes from the derived words. For example, the sense ***nyata*** was mapped to the derived form ***ternyata***. As MorphInd has already analysed ***ternyata*** to *ter-**nyata**<a>*, the two words are mapped by stripping off the prefix *ter-* and the POS tag <*a*>. However, as the links created only use the root form to match the two words, some of them might not be valid in *Bahasa Melayu*. This may be due to the root and derived forms being too far apart to be considered as a derivational link. Also, each word can contain multiple meanings which may not match the derived forms. Over-generation of the derivational links will occur at this stage. Therefore, there was a need to filter out the invalid links to ensure that only the correct senses are matched.

4.2.1 Using Wordnet Relations

The Princeton Wordnet has its own semantic relations under `derivational_related_link` (Fellbaum, Osherson, & Clark, 2009). One can test if the same relation occurs for *Bahasa Melayu* by using this link that exist in Princeton WordNet. Another Wordnet relation that was applicable is `pertainym` (Fellbaum, Osherson, & Clark, 2009). Pertainym is a word (usually an adjective) which is related to another word. For example, in the Princeton Wordnet, the adjective ***daily*** is a pertainym of the noun ***day***.

The derivational related links will be expected to have a small impact on Wordnet Bahasa as English is not as rich in derivational morphology as compared to *Bahasa Melayu*, As the derivational link in one language can be used to prove the existence of derivational link in another language (Bilgin et al., 2004; Mititelu, 2012), these two links will be used to prove that the derivational link does exist both in English and *Bahasa Melayu*. Therefore, if a `derivational_related_link` or `pertainym` link from PWN can be found in Wordnet Bahasa, this proves that the derivational link in *Bahasa Melayu* is valid.

4.2.2 POS Based Filter

As show in Table 1 to Table 5, all the affixes in *Bahasa Melayu* can only occur in certain Part-of-Speech (POS) (Macdonald et al., 1976; Sneddon, 2010). However, unlike Romanian where the root and derived forms must be in the same POS (Mititelu, 2012), the POS for the root word and derived forms can differ in *Bahasa Melayu*. Table 11 shows the breakdown of all available POS as described by Macdonald et al. (1976), Verhaar (1984) and Sneddon (2010). For example, the prefix *meN-* can only form derived verbs from root nouns, verbs or adjectives. A POS based filter is the most appropriate first step as this will ensure the right derived form is linked to the right root word. For example the word ***tegak*** can be used as a noun "upright" or an adjective "erect" or even as an adverb "straight". However, the derived word ***menegak*** "to erect" cannot be derived from ***tegak*** as an adverb. Furthermore, derived forms are also subjected to the POS constraint. The derived form with the prefix *meN-* can only exist as a verb in *Bahasa Melayu* (Macdonald et al., 1976; Sneddon, 2010). Without the POS filter, any derived words with the prefix *meN-* that exist in the wrong POS (such as noun or adjective) will still be linked to the root word. Therefore the POS filter will remove the redundant links and should increase the accuracy of the mapping.

4.2.3 Semantic Super Type Filter

Wordnet synsets were categorised into forty five different files super types on the syntactic category and logical grouping of the synsets. As seen in Table 12, the spread of the semantic super types (also known as Lexicographer Files) are more towards nouns and verbs. Using the semantic super types, semantic categories for the root words and the derived forms are shown. As the prefix *peN-* will derive an agent or instrument noun out of a verb (Macdonald et al., 1976; Sneddon, 2010), the derived form should be under the super type of either $\texttt{person}_n$ or $\texttt{artifact}_n$. As semantic features are more apparent in nouns (Macdonald et al., 1976), the super type filter will be used on the derivation links with derived noun forms. The affixes with concrete semantic categories and have suitable super types in Wordnet are shown in Table 13. Using the semantic super type filter, the semantic type of the derived forms should agree to the semantic features that are available in *Bahasa Melayu* as studied in Macdonald et al. (1976) and Sneddon (2010).

4.2.4 Using Wordnet Definitions

As stated in the previous section, a word can contain many different meaning and also exist in different parts of speech. For example, the word ***hidup*** can exist as a noun, verb or adjective. As a noun, *hidup* can have a number of definitions. Referring to Wordnet Bahasa, ***hidup*** can either be "a characteristic state or mode of living" (**WN** ID:13963192-n) or "a state of surviving" (**WN** ID:13962166-n). While for the derived form ***menghidupkan*** the definition can either be "cause to perform" (**WN** ID:00517529-v) or "be brought back to life" (**WN** ID:00169298-v). Therefore, based on the definitions for ***hidup*** and

POS	Affixes
Noun	
N→N	*per-*, *se-*, *pra-*, *anti-*,
	-an, *-wan*, *-wati*, *-i*,
	per- -an
N→V	*di-*, *meN-*, *ber-*,
	-kan,
	meN- -kan, *meN- -i*, *di-per-*, *di- -kan*, *di- -i*, *meN-per- -i*, *meN-per- -kan*,
	di-per- -i, *di-per- -kan*, *ber-ke- -an*, *ber-per- -an*
N→A	*se-*, *-wan*, *-i*,
N→R	*se-*, *se- -nya*
verb	
V→N	*peN-*, *ke-*,
	peN- -an, *ke- -an*,
V→V	*di-*, *meN-*, *ter-*, *per-*, *ber-*,
	-kan,
	per- -an, *meN- -kan*, *ke- -an*, *ber-ke- -an*
V→A	*ter-*
V→R	*-nya*, *se- -nya*
adjective	
A→N	*peN-*, *per-*, *ke-*,
	-an,
	per- -an, *ke- -an*
A→V	*ke- -an*
A→A	*ter-*, *se-*
A→R	*se-*,
	se- -nya
adverb	
R→R	*se- -nya*

Table 11: Distribution of Affixes based on POS

$\texttt{Tops}_n$	$\texttt{act}_n$	$\texttt{animal}_n$	$\texttt{artifact}_n$
$\texttt{attribute}_n$	$\texttt{body}_n$	$\texttt{cognition}_n$	$\texttt{communication}_n$
$\texttt{event}_n$	$\texttt{feeling}_n$	$\texttt{food}_n$	$\texttt{group}_n$
$\texttt{location}_n$	$\texttt{motive}_n$	$\texttt{object}_n$	$\texttt{person}_n$
$\texttt{phenomenon}_n$	$\texttt{plant}_n$	$\texttt{possession}_n$	$\texttt{process}_n$
$\texttt{quantity}_n$	$\texttt{relation}_n$	$\texttt{shape}_n$	$\texttt{state}_n$
$\texttt{substance}_n$	$\texttt{time}_n$		
$\texttt{body}_v$	$\texttt{change}_v$	$\texttt{cognition}_v$	$\texttt{communication}_v$
$\texttt{competition}_v$	$\texttt{consumption}_v$	$\texttt{contact}_v$	$\texttt{creation}_v$
$\texttt{emotion}_v$	$\texttt{motion}_v$	$\texttt{perception}_v$	$\texttt{possession}_v$
$\texttt{social}_v$	$\texttt{stative}_v$	$\texttt{weather}_v$	
$\texttt{all}_a$	$\texttt{pertainyms}_a$	$\texttt{participial}_a$	$\texttt{all}_r$

Table 12: WordNet Lexicographer Files.

menghidupkan, the derivational link should be made for “a characteristic state or mode of living” and “be brought back to life” as they are connected by the word “life/living”. As such, to do this process automatically, the definition of the words must be compared for its similarity. In this example, the similarity of “a characteristic state or mode of living” and “be brought back to life” should be higher than “a state of surviving” and “cause to perform” as there is a similar word in the first pair. Therefore, a string similarity measure can be used to disambiguate the definition.[4] Other components of the sense such as its synonyms and examples can also be added as items of comparison.

The Lesk algorithm is a disambiguation method where the definition of each sense in a word is compared to the definition of every word in a phrase (Lesk, 1986). A word is assigned to the most appropriate sense which definition is similar to the definition of the other words surrounding it. For example, using the phrase ***time flies like an arrow***, the Lesk algorithm compares all the definitions of ***time*** to the definitions of ***fly*** and ***arrow*** and assigns the most appropriate definition according their similarity. As this algorithm can be used to compare the root words and derived forms, the Lesk algorithm will be used for this study to evaluate the validity of the derivational links that are created from the raw mapping.

The Lesk Algorithm used for this study was adapted from Banerjee & Pedersen (2002) and Baldwin et al. (2010) where an extended version was used. In the extended Lesk, the definition of the semantic

[4] metric that measures the distance of one string to another.

POS→DPOS	Affixes	Semantic Features	Derived Super Type
N→N	*per-*	Agent or Instrument noun	$\texttt{person}_n$, $\texttt{artifact}_n$
	se-	similar, singular, collection	$\texttt{group}_n$ or similar to root
	pra-	previous state	similar to root
	-an	variety, collective	similar to root
	-wan	Agent Noun (male)	$\texttt{person}_n$
	-wati	Agent noun (female)	$\texttt{person}_n$
	per- -an	Process or Activity	$\texttt{act}_n$
V→N	*peN-*	Agent or Instrument noun	$\texttt{person}_n$, $\texttt{artifact}_n$
	per-	Agent or Instrument noun	$\texttt{person}_n$, $\texttt{artifact}_n$
	-an	resultative, undergoer, or Instrument noun	$\texttt{act}_n$, $\texttt{communication}_n$, $\texttt{person}_n$, $\texttt{artifact}_n$, $\texttt{food}_n$
	per- -an	Process or Activity	$\texttt{act}_n$
	peN- -an	Process or Activity	$\texttt{act}_n$
A→N	*peN-*	Agent or Instrument noun	$\texttt{person}_n$, $\texttt{artifact}_n$
	per-	Agent or Instrument noun	$\texttt{person}_n$, $\texttt{artifact}_n$
	peN- -an	Process or Activity	$\texttt{act}_n$

Table 13: Suitable Semantic Features using Wordnet Super Types

relations (e.g. hypernym-hyponym relations) in the word was added to word definition during the comparison process (Banerjee & Pedersen, 2002). Baldwin et al. (2010) refined the algorithm even further by looking at the definition of each word that exists in the main word's definition. For example, the definition of the word ***goal*** in PWN is “a successful attempt at scoring” (**WN** ID:00187337-n). The extended Lesk algorithm by Baldwin et al. (2010) will take the definition of ***successful***, ***attempt*** and ***scoring*** on top of the main definition of the word. These extended definitions can be taken from the Princeton WordNet Gloss Corpus.[5] The Gloss Corpus used the definition that exist in the Wordnet synsets and manually link each word in the definition to the most appropriate sense in Wordnet (Szymański & Duch, 2012). Therefore in the Gloss Corpus, the Princeton Wordnet acts as the dictionary in which the definition from the same Wordnet is tagged. This was similar to the spreading activation process where during sentence comprehension, words in the sentence are automatically linked to the related concept (Szymański & Duch,

[5] `http://wordnet.princeton.edu/glosstag.shtml`

2012). This could serve as a rough semantic representation of the target sense as the tagging process in the Gloss Corpus is constrained by the other senses surrounding it. Thus, adding the Gloss definition on top of the sense definition, hypernyms and hyponyms will increase the accuracy of representation for the particular sense.

Three different measurements are created for this study to ensure that the different components of a Wordnet sense are covered and to get the most accurate measurement for the similarity of the root and derived form. Lesk 1 (Algorithm 1) measures the similarity of the lemmas, examples and definition. Lesk 2 (Algorithm 2) uses the sense's definition, hypernyms and hyponyms while Lesk 3 (Algorithm 3) uses the definitions, hypernyms, hyponyms and Gloss definition. Even though the three string measurements uses different items, there should be similarity in them regarding the appropriateness of the root and derived forms. If the root and derived forms are very far apart (where the meaning is not transparent), the scores for all the three measurements should be lower than when the two forms are related.

```
for each synset in wordnet do
    for each sense and examples and definition in synset do
        string = lemmas+examples+definition
    end
    score(main,derived) = similarity($string^{\text{main}}$,$string^{\text{derived}}$)
end
```

Algorithm 1: Lesk 1=synonyms+examples+definition

```
for each synset in wordnet do
    for each definition and hypernyms and hyponyms in synset do
        string = definition+hypernym+hyponym
    end
    score(main,derived) = similarity($string^{\text{main}}$,$string^{\text{derived}}$)
end
```

Algorithm 2: Lesk 2=definition+hypernyms+hyponyms

```
for each synset in wordnet do
    for each definition and hypernyms and hyponyms in synset do
        for definition in Gloss_Corpus do
            if definition in synset == definition in Gloss_Corpus then
                Gloss_definition = definition of synset in Gloss_Corpus
            end
        end
        string = definition+hypernyms+hyponyms+Gloss_definition
    end
    score(main,derived) = similarity( $string^{\text{main}}$, $string^{\text{derived}}$)
end
```

Algorithm 3: Lesk 3=definition+hypernyms+hyponyms+Gloss_definition

5 Results

5.1 MorphInd Process

MorphInd was able to analyse the senses in Wordnet Bahasa into the root words and affixes. Table 14 shows the breakdown of the MorphInd process on Wordnet Bahasa's senses. As MorphInd uses its own dictionary for analysis, only 70% of the total number single word senses is processed by the tool as shown in Table 14. The borrowed prefixes found in Wordnet Bahasa are *anti-* and *pra-* . MorphInd also processed words with the particle *-lah* and the clitic *-nya* and futher breakdown of the MorphInd process into the different affixes are shown in Appendix C . For the suffix *-nya*, derived forms will only occur from a verb root or as a circumfix with a prefix *se-* (Macdonald et al., 1976; Verhaar, 1984). While those sense with the particle *-lah* and the clitic *-nya* will not be linked to the root word as they are part of inflectional morphology Ranaivo-Malancon (2004).

5.1.1 Derivational Link From Princeton Wordnet

The `derivation_related_link` and `pertainym` link from the Princeton Wordnet did apply to Wordnet Bahasa as shown in Table 15 . The number is quite minute as compared to the links present in the whole data. However, the presence of the Princeton Wordnet derivational links as shown in Table 15 affirms the derivational links in Wordnet Bahasa especially for nouns and verbs (Bilgin et al., 2004). Those Wordnet Bahasa sense with either the derivational or pertainym link from the Princeton

	Indonesian		Malay	
Wordnet senses	142,488		119,152	
	Processed	Not Processed	Processed	Not Processed
Multi-word	23,261	13,271	20,036	11,146
Single-word	71,042	34,914	63,386	24,584
root word	25,645		23,634	
Prefix	18,405		15,934	
Suffix	5,127		4,878	
Cirumfix	21,837		18,913	

Table 14: Result of MorphInd Process

Wordnet as shown in Table 16 will be considered as a concrete derivational link.

5.2 POS and Semantic Super Type Filter

A raw mapping is created between the root and analysed words. This is done for all POS in the Wordnet Bahasa and the number of links is shown in Table 17. Overall, root verbs have the highest number of derived forms followed by nouns, adjective and adverbs for both databases.

As seen from the Table 17, the POS filter helped in reducing the redundancy of the derivational links. The number of links decreases for all the POS for both Indonesian and Malay data. As stated previously, derivational forms would only arise from certain POS (Macdonald et al., 1976; Sneddon, 2010). Examples of the validated links are shown in Table 18. The link betweeen ***nafas*** and ***pernafasan*** was valid as the POS filter will allow a noun to be derived from a verb via the circumfix *per- -an*. While for ***terbelakang*** and ***bertahan***, the links are invalid as the prefix *ter-* and *ber-* cannot derive nouns from verbs. For the derived verb ***mencuba***, the link to ***cuba*** was invalid as verbs with prefix *meN-* cannot be derived from adverbs. The semantic super type filter also eliminates those links with the derived form in the incorrect super type category as shown in Table 19. The super type filter can only increase the accuracy of derived nouns, as concrete semantic categories are only found in this category (Macdonald et al., 1976; Sneddon, 2010). However, the super type filter was still effective as it was able to reduce the number of invalid links for the derived noun categories as shown in Table 17.

	Indonesian		Malay	
POS→DPOS	Derivational	Pertainym	Derivational	Pertainym
noun				
N→N	547	0	501	0
N→V	2,398	0	2,156	0
N→A	10	0	9	0
N→R	1	0	1	0
verb				
V→N	1,262	0	1,200	0
V→V	1,677	0	1,403	0
V→A	647	0	695	0
V→R	0	0	0	0
adjective				
A→N	573	18	552	19
A→V	2	0	2	0
A→A	0	0	0	0
A→R	0	0	0	0
adverb				
R→R	0	0	0	0

Table 15: Princeton Wordnet Derivational and Pertainym Links in Wordnet Bahasa

5.2.1 Using Wordnet Definitions

The different components of the synset (e.g. definition. hypernym, hyponym, examples) have to be concatenated into a string before the string similarity measurement can be done as shown in Table 20. For example in Lesk 1, the first string will consist of synonyms, examples and definition of the root word while the second string will consist of synonyms, examples and definition of the derived form. The strings are cleaned up of all punctuation such as commas and full stop. These punctuations usually exist in the definitions and examples, and may affect the string measurement if they are not removed.

A similarity string measurement will use the two strings that were generated and return a score of their similarity. As with the extended Lesk Algorithm that was used by Baldwin et al. (2010), this paper also

POS→DPOS	**WN** ID	Sense	From Eng Wordnet
	02460964-a	*jati*	Root: genuine
A→N	13955341-n	*ke-jati-an*	DRV: genuineness
	02024367-v	*meN-tekan*	Root: press
V→N	03999992-n	*peN-tekan*	DRV: press
	02743261-a	*manusia*	Root: human
A→N	04829182-n	*ke-manusia-an*	PRT: humanity
	02767378-a	*lahir*	Root: birth
A→N	15142167-n	*ke-lahir-an*	PRT: birth

Table 16: Wordnet Bahasa Derivational Links with Princeton Wordnet Equivalent

$$d = \frac{2(|X \cap Y|)}{|X| + |Y|}$$

Figure 1: Dice Coefficient

uses the Dice coefficient as a similarity measure. The cosine string measure was first tested in this study but the score generate was insignificant and irregular. Dice coefficient is a metric used to compare two strings in terms of common bigrams (pair of adjacent letters) in a string. The formula for Dice coefficient is shown in Figure 1. In this formula, ***X*** is the common bigrams of the root word while ***Y*** is the common bigrams of the derived form. ***d*** is the quotient of similarity, with the score of 1 begin the most similar and 0 for no similarity between the 2 strings.

To achieve the best accuracy during the WSD process , a “Gold standard” sample data is created. A “Gold standard” is a standard accepted as the most valid for the particular study and in this case, the most valid derivational links that were created in Wordnet Bahasa. Boyd-Graber et al. (2006) stated that researchers agree more with a manual tagged “Gold standard” than a fully automated version. A sample size of 966 sense link was manually tagged. The Lesk 1 string of the root word and the derived forms were used to determine the validity of the links. The objective of the threshold level is to find the best score with the highest correct pairing (also known as precision). Even with the use of the POS and super types filters, errors exist in the links. This is mainly caused by the mismatch of the definition to the root and derived forms. These errors are tagged as "D" and "R" to indicate error in derived definition and root definition respectively and are excluded. Those which are considered for analysis, are either tagged as

	Indonesian			Malay		
Derived form	Raw	Pos	Super Type	Raw	POS	Super Type
noun						
N→N	28,079	13,874	3,802	26,333	13,051	3,603
N→V	67,984	65,446	65,446	55,443	53,331	53,331
N→A	8,183	565	565	8062	498	498
N→R	898	229	229	845	225	225
verb						
V→N	76,870	59,720	46,185	73,435	56,758	44,356
V→V	97,214	85,162	85,438	75,635	75,635	75,635
V→A	5,070	647	647	5174	695	695
V→R	514	93	484	484	84	84
adjective						
A→N	33,283	29,605	27,568	31,889	28,531	26,643
A→V	52,252	369	369	45,817	361	361
A→A	6,067	242	242	6,348	223	223
A→R	1,810	240	240	1,753	255	255
adverb						
R→N	3,108	0	0	3,058	0	0
R→V	8,612	0	0	7,796	0	0
R→A	992	0	0	1,017	0	0
R→R	595	69	69	585	72	72

Table 17: Breakdown of Mapped Derivational Links before and after Filters

"Y" for valid or "N" of invalid link. In total, 693 links were considered valid out of 966 sample senses link and the threshold level (precision score) is determined using the recall of “Y” and “N” . The precision score and sample recall are shown in Table 21 to Table 24.

Out of the three Lesk measurements, Lesk 3 has the best sample recall of 481 Y-Y pairing and 1 N-N pairing. As the three Lesk algorithms measure different components of a sense, there was no improvement in the precision or recall when they are combined as shown in Table 24. As such, Lesk 3 is used as the most appropriate similarity measurement with the threshold value of 0.149. After the

POS→DPOS	**WN** ID	Sense	Definition	Validity
	14841770-n	*nafas*	the air that is inhaled and exhaled in respiration	
N→N	00831191-n	*per-nafas-an*	the bodily process of inhalation and exhalation	Valid
	00061203-r	*belakang*	happening at a time subsequent to a reference time	
N→N	08502507-n	*ter-belakang*	a place or condition in which no development or progress is occurring	Invalid
	00459776-v	*tahan*	cause to be slowed down or delayed	
V→N	10303654-n	*ber-tahan*	someone who exhibits great independence in thought and action	Invalid
	00004722-r	*cuba*	and nothing more	
R→V	02373336-v	*men-cuba*	proceed somewhere despite the risk of possible dangers	Invalid

Table 18: Validation of Derivational Links via POS Filter

baseline threshold level was set, an automated word sense disambiguation (WSD) process was done to see the recall of the Lesk measurements and the validity of the derivational links across both the Malay and Indonesian data. The results of the WSD process are shown in Table 25.

For comparison purposes, each of the Lesk measurements' threshold level is parsed through both the Indonesian and Malay data of Wordnet Bahasa. As seen from table 25, the recall level for all the Lesk measurements are really high with most of them giving a score of 0.90 or above. This means that 90% of the links are considered as valid. This was due to the very low threshold levels that were set for all the three Lesk. Nevertheless, when the Lesk measurements were compared with each other, Lesk 3 has

POS→DPOS	**WN** ID	Sense	Super Type	Validity
	06723908-n	*kata*	$\texttt{communication}_n$	
N→N	07009421-n	*pra-kata*	$\texttt{communication}_n$	Valid
	00259755-v	*meN-usik*	$\texttt{change}_v$	
V→N	10305192-n	*peN-usik*	$\texttt{person}_n$	Valid
	01217499-n	*wakil*	$\texttt{act}_n$	
N→N	01140839-n	*per-wakil-an*	$\texttt{act}_n$	Valid
	13828075-n	*arah*	$\texttt{relation}_n$	
N→N	06786629-n	*arah-an*	$\texttt{communication}_n$	Invalid
	04998530-n	*rupa*	$\texttt{attribute}_n$	
N→N	04683814-n	*rupa-wan*	$\texttt{attribute}_n$	Invalid

Table 19: Validation of Derivational Links via Super Types Filter

the lowest recall level for most of the POS and derived form. On the other hand, Lesk 1 has the highest recall with a score of 1.00 for adverbs derived from verbs for both Malay and Indonesian. This confirmed Lesk 3 as the most appropriate string measurement of the derivational links. However, with the high level of errors as stated above, they render the Lesk algorithm ineffective in disambiguation the derivational links.

Example	04617562-n	*ke-peribadi-an*
Lesk 1	synonym	personality
	Definition	the complex of all the attributes behavioral temperamental emotional and mental that characterize a unique individual
	Examples	their different reactions reflected their very different personalities it is his nature to help others
Lesk 2	Hypernym	attribute
	Hyponyms	narcissistic personality identity oral personality genital personality anal personality personableness obsessive compulsive personality
	Definition	the complex of all the attributes behavioral temperamental emotional and mental that characterize a unique individual
Lesk 3	Hypernym	attribute
	Hyponyms	narcissistic personality identity oral personality genital personality anal personality personableness obsessive compulsive personality
	Definition	the complex of all the attributes behavioral temperamental emotional and mental that characterize a unique individual
	Gloss Definition	highly unusual or rare but not the single instance relating to or caused by temperament of or relating to behavior a human being involving the mind or an intellectual process the complex of all the attributes behavioral temperamental emotional and mental that characterize a unique individual of more than usual emotion

Table 20: Breakdown of the Different Lesk Strings

		Computer	
Threshold	0.144	Y	N
Human	Y	481	1
	N	15	1
	D	22	
	R	153	

Table 21: Result of Sample using Lesk 1

Computer			
Threshold	0.044	Y	N
Human	Y	479	3
	N	153	3
	D	22	
	R	153	

Table 22: Result of Sample using Lesk 2

Computer			
Threshold	0.149	Y	N
Human	Y	482	0
	N	155	1
	D	22	
	R	153	

Table 23: Result of Sample using Lesk 3

Computer			
Threshold	0.044	Y	N
Human	Y	481	1
	N	155	1
	D	22	
	R	153	

Table 24: Result of Sample by Combining all Lesk Results

	Indonesian			Malay		
Measure	Lesk 1	Lesk 2	Lesk 3	Lesk 1	Lesk 2	Lesk 3
Threshold	0.144	0.044	0.149	0.144	0.044	0.149
Noun						
N→N	0.995	0.999	0.999	0.995	1.0	0.999
N→V	0.990	0.993	0.991	0.989	0.994	0.991
N→A	0.992	0.948	0.961	0.992	0.975	0.958
N→R	0.986	0.973	0.995	0.986	0.977	0.968
Verb						
V→N	0.989	0.994	0.988	0.989	0.995	0.988
V→V	0.991	0.986	0.976	0.991	0.987	0.975
V→A	0.991	0.979	0.984	0.997	0.978	0.985
V→R	1.00	0.989	0.989	1.00	0.988	0.988
Adjective						
A→N	0.987	0.986	0.986	0.992	0.987	0.986
A→V	0.99	0.975	0.959	0.994	0.974	0.96
A→A	0.995	0.983	0.971	0.995	0.982	0.972
A→R	0.988	0.954	0.975	0.988	0.951	0.972
Adverb						
R→R	1.00	0.898	0.913	1.0	0.892	0.909

Table 25: Recall Value using Lesk Measurements

6 Discussion

Errors in the MorphInd process were found in this study. For example, words like ***sebarang*** “anything” and ***pemarah*** “angry person” were separated into ***se+barang*** and ***peN+parah***. For the word ***sebarang*** it is considered a word unit and not a prefix *se-* with the root word ***barang***. Native speakers would agree that ***sebarang*** “anything” is not derived from ***barang*** “item” as the definition for the two words are quite far apart. Regarding the word ***pemarah***, the analysed form is wrong, as it should be ***peN+marah*** instead of ***peN+parah***. This causes the word to be linked to ***parah*** instead of ***marah*** and therefore, causing the link to be invalid. These errors are highlighted to the creator of MorphInd so that they can be rectified in the future release of the software. Further improvements to MorphInd are explained in Appendix B.

6.1 Evaluation of POS and Super Type Filter

The POS filter is much more comprehensive as compared to the Super Type filter. The POS filter encompasses the whole data and was able to remove derivational links with the root and derived forms in the wrong POS. For Wordnet Bahasa, the Super Type filter can only filter out those affixes with concrete semantic features in their derived form. Thus, the Super Type filter is only able to filter out only the derived nouns because of this limitation. Also, as the distributions of the super types are not even for all POS in Wordnet, it is difficult to use the filter for other part of speeches. This is especially true in the case of adjectives and adverbs where most of them are grouped in one major type - $\texttt{all}_a$ and $\texttt{all}_r$. Thus, it would be difficult to differentiate the types of adjective and adverbs using the super types. However, as both filters were able to increase the accuracy of the derivational link by reducing the redundant and unwanted entries through POS and semantic features, the filters are useful to the creation of the derivational link in Wordnet Bahasa. As the redundant and error links are removed, the links will become more representative of the derivational morphology in *Bahasa Melayu*.

Further filters can be added to enhance the accuracy of derivational links in Wordnet Bahasa. One area that can be studied is transitivity, where the number of objects a particular verb can hold is analysed. As change in transitivity was involved when deriving a verb from a root verb using the prefix *meN-* and *di-* (Macdonald et al., 1976), the wordnet Verb Frames can be used to analyse the transitive link. The Verb Frames that represents transitive verbs are shown in Table 26. For the prefix *di-*, it cannot be derived from transitive root verbs (Macdonald et al., 1976). Therefore using Verb Frames, those derived forms with prefix *di-* and are link to a transitive root verb will be removed from the database. However, as the

Frame No.	Verb Frame
8	Somebody —-s something
9	Somebody —-s somebody
10	Something —-s somebody
11	Something —-s something
12	Something —-s to somebody
13	Somebody —-s on something
14	Somebody —-s somebody something
15	Somebody —-s something to somebody
16	Somebody —-s something from somebody
17	Somebody —-s somebody with something
18	Somebody —-s somebody of something
19	Somebody —-s something on somebody
20	Somebody —-s somebody PP
21	Somebody —-s something PP

Table 26: Verb Frames indicating Transitive Verbs

transitivity filter will only be used for the prefix *di-* and that the number of derived words with the prefix *di-* in the data is small, using this filter now will not have a big impact in disambiguating the links.

6.2 Evaluation of WSD using Lesk algorithm

Good examples in the derivational links can be seen in Table 27. The word ***relatif*** has a derived form which is above the Lesk 3 threshold of 0.149. Furthermore, the derived form ***kerelatifan*** has a similar derived form ***relativity*** in the Princeton Wordnet. While the word ***asuh*** has a derived form ***asuhan*** with a pertainym link ***education*** in English and a Lesk 3 value of 0.521. As such, this links are considered concrete examples of good derivational link in Wordnet Bahasa. However, as stated before, the threshold level set by the “Gold standard” is very low for all the Lesk algorithms. This caused a substantial number of erroneous links to be passed as valid links, even for the best Lesk algorithm - Lesk 3. Three main reasons can be attributed to the low threshold levels.

Firstly, there is a mismatch in Wordnet Bahasa with regards to the sense and its definition. This error

POS→DPOS	**WN** ID	Sense	PWN	Lesk 3
	00006032-a	*relatif*		
A→N	06106502-n	*ke-relatif-an*	DRV: relativity	0.351
	02946221-a	*asuh*		
A→N	04921900-n	*asuh-an*	PRT: education	0.521
			DRV: education	
	00004475-n	*manusia*		
N→N	04726938-n	*ke-manusia-an*	None	0.533
N→N	04829182-n	*ke-manusia-an*	None	0.404
	00002684-n	*barang*		
N→R	00024509-r	**se-barang*	None	0.352
N→A	02267686-a	**se-barang*	None	0.332
	00004227-v	**mati*		
V→R	00024509-r	**ke-mati-an*	None	0.22
V→A	02267686-a	**ke-mati-an*	None	0.26
	00010435-v	*tindak*		
V→N	06532095-n	**tindak-an*	None	0.509

Table 27: Evaluation of Derivational Links after Filters and Lesk

is still quite apparent in Wordnet Bahasa where some words are misrepresented in the wordnet due to the ambiguity of the English translation. thus causing wrong links to be created. For example, in Table 27, the definition of the root word ***mati*** (**WN** ID:00004227-v) is “to expel air”. As ***mati*** can only relate to “endings” or “death”, it was wrongly added to the particular definition. This caused the link to the any of derived word of ***mati*** to be invalid (tagged as “R” in the sample data). Another example would be the word ***tindak*** (**WN** ID:00010435-v) and its derived word ***tindakan*** (**WN** ID:06532095-n). The derived definition is "a legal document codifying the result of deliberations of a committee or society or legislative body". Even though morphologically the link is correct, the discrepancy between the derived word and definition causes this link between the two words to be invalid (tagged as “D” in the sample data) . As the definition is one of the main items tested for all three Lesk algorithms, this causes a huge number of noise to be present in the data consisting of either wrong root definition or wrong derived definition. Even as the error links were removed during the “Gold Standard” test, there was no significant improvement to threshold level. As seen from Table 21 to Table 24, the errors in root words (tagged as "R") and derived forms (tagged as "D") make up almost 20% of the links in the sample data. As shown in Table 27, the Dice coefficient score for the invalid links are still higher than the threshold level and this is even after they were removed from the "Gold Standard" test. Therefore, it is very likely that the error entries are extensive enough to render the Lesk method to be ineffective.

Another reason for the WSD process to be ineffective is that the ambiguity between the *Bahasa Melayu* word and English definition. Translated text might not be true representation of the intended meaning, so much so that the sense in one language can differ in meaning in another (Uchida & Zhu, 2001). For the word ***memotong*** in Wordnet Bahasa, one of the definitions is "end or extinguish by forceful means". This was incorrect as ***memotong*** actually means "to cut" or "to slice". Therefore, the English definition for the word sense does not suit that particular word. However, when the Lesk algorithm was used to measure the string similarity for Lesk 3 of the root word ***pemotong*** to the derived form ***memotong***, the Lesk 3 score was 0.546. This is way above the threshold level of 0.149 for Lesk 3. Therefore, adding and using Malay definitions may solve this issue as the ambiguity element will be removed.

Finally, errors in the MorphInd analysis affect the Lesk measurement. Even though the words ***barang*** and ***sebarang*** passed the POS filter, they are invalid in actual fact as their definition was very far apart. Interestingly, the Lesk 3 score for this link (0.352) is not much different from the Lesk score for the valid link ***relatif*** to ***kerelatifan*** (0.351). Thus, this caused the Dice coefficient measure and Lesk algorithm as a whole to be ineffective as it was unable to make the distinction between the valid and invalid links.

7 Further Work

As seen in the results section, the amount of noise in Wordnet Bahasa did affected the WSD process. This causes the WSD process to be unable to disambiguate the wrong links from the correct ones. As the Lesk score for the wrong link is higher than or similar to the threshold level they are deemed as valid links. Steps were taken to try and improve the score such as eliminating those with wrong definition for the root or derived forms. However, even with these steps, very little improvements can be seen in the precision and recall values.

Therefore, the Wordnet Bahasa has to be cleaned further before the derivational link can be added. Tools such as a morphological analyser can be used to identify words in the wrong POS. As shown by the

POS filter, the part of speech of a derived word is determined by the affixes. In this study, the MorphInd tool was able to extract a total of 6,500 wrong entries of affixed single words for both the Indonesian and Malaysian data. However, more can be done to improve the Wordnet Bahasa. As MorphInd targets Indonesia words, the Malay data can be further improved by using morphological analyser that is specific for Malay. With more morphological analysers being used, more wrong entries can be identified and thus increasing the accuracy of the Wordnet Bahasa.

Currently, Wordnet Bahasa uses the English definitions as found in the Princeton Wordnet. This was adequate for activities such as corpus tagging, as the definitions are use only to identify the most apportioned sense. However, in this study, the definitions were used for comparison during the WSD process. As information is always lost in translation (Uchida & Zhu, 2001), the English definition will not fit the Malay word fully. As stated above, misrepresentation of the Malay word in Wordnet Bahasa was caused by this ambiguity. For example, the current definition for ***tembak*** (**WN** ID:02123175-v) is "cause a sharp and sudden pain in". Even though ***tembak*** is a direct translation of the lemma "shoot", the more appropriate sense in Bahasa Melayu would be ***cucuk***. Thus, with the addition of the Malay definitions, the differences in meaning would be clearer and it will be more representative of the senses using the same language. Furthermore, the misalignment of the Malay sense to the English definition will be more explicit and this will assist in removal of the invalid senses.

Another method of disambiguating is via corpus tagging process. Similar to the “Gold Standard” test, the targeted affixed word would be tagged with the most appropriated root word and derived form. In this way, the link between the root and derived word can be firmly established manually. Even though tagging is not efficient and time consuming, it would be one of the most effective way of establishing the derivational links. As a native speaker would be using his/her native knowledge of Bahasa Melayu to do the tagging process, the validity of the derivational links through this process will be much more concrete.

On top of affixes, *Bahasa Melayu* has other derivational word formations such as compounding and reduplication. These two word formations should be processed similarly to the affixed words. As MorphInd is able to handle these two word formation process, the software can still be used to analyse the derived words. However, these two processes should be looked at when the links for the affixed words are improved to an acceptable level. The semantic categories should also be added to better represent the function of the derived words (Fellbaum, Osherson, & Clark, 2009).

8 Conclusion

This study was able to create a basic derivational link for the root and affixed words in Wordnet Bahasa for both the Malaysian and Indonesian data. Using the MorphInd, derived words were broken down into its root word and affixes. This process helped in the raw linkage of the root words to the analysed derived word. These links were then refined using the POS and semantic Super Type filter. However, when the automated WSD process was executed via the Lesk algorithm, it was proven to be ineffective. One of the main factors is the high amount of noise that is still in Wordnet Bahasa. Therefore, because of the high level of inaccuracy and noise, it would be inappropriate to include all the derivational links into the Wordnet Bahasa for now. As stated above, only those links with a derivational or pertainym link in English will be included into the Wordnet Bahasa. The wrong derived words that were discovered using MorphInd are removed from the Wordnet Bahasa. The rest of the links are released as an external file and will be included in the Wordnet Bahasa website. The "Gold standard" test that was done prior to the automated process is also release with the definition and tags clearly shown.

In conclusion, by attempting to create the derivational link in Wordnet Bahasa, errors were discovered in the Wordnet Bahasa. Even though the study did not achieve the full automated WSD process for the derivational link in Wordnet Bahasa, basic links for affixes words were created. Some of the links are also supported by the existence of the same derivational link in the Princeton Wordnet. As, it would be inappropriate to add all the derivational links to the current Wordnet Bahasa, they are released as a separate file as shown in Appendix A. With the MorphInd tool, most of the derived words were broken down it its affix and root word. Additionally, MorphInd is able to extract the error words from Wordnet Bahasa.and thus, assisting in the clean-up of the dictionary. The accuracy of Wordnet Bahasa will increase progressively with the database cleaned up and addition of Malay definitions. Thus, after improvements made to Wordnet Bahasa, the Lesk algorithm (or its equivalent) can be used effectively in disambiguating the derivational links and a more representative data of the derivational link can then be added into Wordnet Bahasa.

References

Baldwin, T., Kim, S., Bond, F., Fujita, S., Martinez, D., & Tanaka, T. (2010). A reexamination of MRD-based word sense disambiguation. *ACM Transactions on Asian Language Information Processing (TALIP)*, *9*(1), 4.

Banerjee, S., & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the third international conference on computational linguistics and intelligent text processing* (pp. 136–145). London, UK, UK: Springer-Verlag. Retrieved from `http://dl.acm.org/citation.cfm?id=647344.724142`

Bilgin, O., Cetinoglu, O., & Oflazer, K. (2004). Morphosemantic relations in and across wordnets. In *Proceedings of the global wordnet conference* (pp. 60–66).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly. ((`www.nltk.org/book`))

Bosch, S., Fellbaum, C., & Pala, K. (2008). Derivational relations in English, Czech and Zulu wordnets. *Literator*, *29*(1), 139–162.

Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the third international wordnet conference* (pp. 29–36).

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.

Fellbaum, C., Osherson, A., & Clark, P. (2009). Putting semantics into WordNet's "Morphosemantic" Links. In *Human language technology. challenges of the information society* (Vol. 5603, p. 350-358). Springer Berlin Heidelberg.

Koeva, S. (2008). Derivational and morphosemantic relations in Bulgarian wordnet. *Intelligent Information Systems*, *16*, 359–389.

Koeva, S., Krstev, C., & Vitas, D. (2008). Morpho-semantic relations in wordnet–a case study for two Slavic languages. In *Proceedings of the fourth global wordnet conference* (pp. 239–254).

Kozok, U. (2012). *How many people speak Indonesian?* `http://ipll.manoa.hawaii.edu/indonesian/2012/03/10/how-many-people-speak-indonesian`. (Retrieved October 28, 2013)

Larasati, S., Kubon, V., & Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Systems and frameworks for computational morphology* (Vol. 100, p. 119-129). Springer Berlin Heidelberg.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on systems* *documentation* (pp. 24–26). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/318723.318728 doi: 10.1145/318723.318728

Macdonald, R. R., et al. (1976). *Indonesian reference grammar*. Georgetown University Press Washington, DC.

Mititelu, V. B. (2012). Adding morpho-semantic relations to the Romanian wordnet. In *Lrec*.

Mohamed Noor, N., Sapuan, S., & Bond, F. (2011). Creating the open Wordnet Bahasa. In *Proceedings of the 25th pacific asia conference on language, information and computation (paclic 25)* (pp. 258–267). Singapore.

Nakov, P., & Ng, H. T. (2011, June). Translating from morphologically complex languages: A paraphrase-based approach. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1298–1307). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1130

Ranaivo-Malancon, B. (2004). Computational analysis of affixed words in Malay language. In *Proceedings of the 8th international symposium on Malay/Indonesian linguistics, Penang, Malaysia.*

See, C. M. (1980). The morphological analysis of Bahasa Malaysia. In *Proceedings of the 8th conference on computational linguistics* (pp. 578–585). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/990174.990281 doi: 10.3115/990174.990281

Sneddon, J. N. (2010). *Indonesian: A comprehensive grammar* (2nd ed.). London: Routledge.

Sugiharto, S. (2008). *Indonesian-Malay mutual intelligibility?* http://www.thejakartapost.com/news/2008/10/25/indonesianmalay-mutual-intelligibility.html. (Retrieved November 8, 2013)

Szymański, J., & Duch, W. (2012). Annotating words using WordNet semantic glosses. In T. Huang, Z. Zeng, C. Li, & C. Leung (Eds.), *Neural information processing* (Vol. 7666, p. 180-187). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-34478-7_23 doi: 10.1007/978-3-642-34478-7_23

Uchida, H., & Zhu, M. (2001). The Universal Networking Language beyond machine translation. In *International symposium on language in cyberspace, Seoul* (pp. 26–27).

Verhaar, J. W. (1984). Affixation in contemporary Indonesian. *Towards a Description of Contemporary Indonesian: Preliminary Studies Part, 1*.

Warrillow-Williams, D. (2002). Evidence for the emergence of new bound morphemes in Indonesian. *Te Reo: Journal of the Linguistic Society of New Zealand*, *45*, 91 - 109. Retrieved from `http://search.ebscohost.com.ezlibproxy1.ntu.edu.sg/login.aspx?direct=true&db=mzh&AN=2003902505&site=ehost-live`

Appendix A Release

In total, using the MorphInd tool, there are 3,255 and 3,299 wrong entries in Wordnet for the Indonesia and Malay data respectively. They will be removed from Wordnet Bahasa. As compare to the number of synsets available in Wordnet Bahasa, this accounts for 2% of the whole database. Through the MorphInd analysis, the derived words contained in this list are incorrect because they exist in the wrong POS. Taking an example from Table 28, ***mendatangkan*** cannot be a noun as the circumfix *me- -kan* can only derive verbs.

WN ID	Sense
00001740-v	pernafasan
00003829-s	melahirkan
00003846-r	menindih
00004032-v	keluhan
00007347-n	mendatangkan
00007347-n	menetaskan
00007347-n	mengakibatkan
00007347-n	menimbulkan
00021766-a	ketepatan
00024720-n	menyatakan

Table 28: List of Error Words

POS→DPOS	Affixes	root **WN** ID	derived **WN** ID	Lesk1	Lesk2	lesk3	Tag	PWN
A→N	ke++an	00631391-a	04899201-n	0.463	0.496	0.434	Y	DRV: correctness
A→N	ke++an	00631391-a	04803209-n	0.369	0.252	0.305	Y	None
V→N	peN+	02302220-v	10180178-n	0.367	0.272	0.256	Y	None
V→N	peN+	02302220-v	03485997-n	0.5	0.271	0.247	Y	None
V→V	ter+	01621555-v	01494310-v	0.414	0.591	0.699	Y	None
A→N	ke++an	02386612-a	14450339-n	0.452	0.207	0.228	Y	DRV: shortness
A→N	per++an	02748635-a	00923444-n	0.448	0.37	0.622	Y	PRT: industry
A→N	ke++an	00964470-a	06212422-n	0.343	0.16	0.521	Y	None

Table 29: Example of Automated Derivation Link in Wordnet Bahasa

As the level of inaccurate derivational links was substantially high, it is inappropriate to include all the links into the current Wordnet Bahasa. Only those which are verified by the Princeton Wordnet through the derivational and pertaiym link will be added into Wordnet Bahasa as these are considered to be good and concrete links. For the rest of the automatic links, they will be release as an external file showing the root ID, derived ID and the affix that links these two IDs as shown in Table 29.

The evaluation data set or "Gold standard" will also be released. This will consist of the tags, **WN** ID and sense and the string used for comparison (Lesk 1) as shown in Table 30. This data will show how the links were adjudicated by the experimenter. Further studies can look at this data set and evaluated if the tagging process was done appropriately.

Similar to Wordnet Bahasa, these data sets will be posted onto the Wordnet Bahasa website.[6] They are also released under the MIT license.[7] This allows of usage, modification, publish and even selling of the data with the proper acknowledgement given.

[6] http://wn-msa.sourceforge.net/

[7] http://www.opensource.org/licenses/mit-license.php

Tag	**WN** ID	Sense	Lesk 1 String
	00956131-a	jujur<a>	fair just free from favoritism or self interest or bias or deception conforming with established standards or rules a fair referee fair deal on a fair footing a fair fight by fair means or foul
Y	04867130-n	ke+jujur<a>+an	sincerity the quality of being open and truthful not deceitful or hypocritical his sincerity inspired belief they demanded some proof of my sincerity
N	04650731-n	ke+jujur<a>+an	frankness outspokenness the trait of being blunt and outspoken
N	04871720-n	ke+jujur<a>+an	candor candour candidness frankness directness forthrightness the quality of being honest and straightforward in attitude and speech
Y	04871374-n	ke+jujur<a>+an	honesty honestness the quality of being honest
D	04701943-n	ke+jujur<a>+an	pellucidness pellucidity limpidity passing light without diffusion or distortion

Table 30: Example of "Gold Standard" data Release

Appendix B Evaluation of MorphInd

As mentioned before, the MorphInd tool was able to analyse all the affixes including clitics and particles that exist in Wordnet Bahasa. This allowed for the affixed words to be fully linked to all the root words. Those senses with clitics and particles, as detected by MorphInd, were removed from the linkage process. However, because of the difference in the dictionary used by MorphInd, not all the senses were processed. Attempts were made to add Wordnet Bahasa senses to the MorphInd's dictionary but they were unsuccessful as only the binary version of the dictionary was available in the public release version of MorphInd. The creator of MorphInd was contacted regarding this issue via email and the developer's version of MorphInd was given. With the developer's version, the dictionary can be edited under the file `lemma.lexin` with the format as shown in Table 31. Thus, by improving and increasing the size of the dictionary as more words will be analysed and linked. This will create more derivational links in Wordnet Bahasa during the raw mapping process. However, this process should be done after the Wordnet Bahasa has been cleaned as there are a number of *Bahasa Melayu* senses in the wrong POS in the current version. A such, adding the senses from the current version will also increase the number of wrong entries in MorphInd.

```
adjective||adekuat
adjective||adem
foreign||dubinsky
foreign||dublin
foreign||ducati
noun||dwitunggal
noun||ebek
noun||ebi
verb||junjung
verb||kabruk
 verb||kabung
verb||kabur
```

Table 31: Format of MorphInd Dictionary

Appendix C MorphInd Process by Affixes

Affix	Indonesian	Malay
meN-per-	254	127
ber-peN-	23	10
ter-peN-	10	10
di-per-	1	1

Table 32: Result of MorphInd Process for Double Prefixes

Affix	Indonesian	Malay
meN-	11,121	9,364
ber-	4,046	3,629
peN-	1,549	1,508
ter-	963	849
se-	166	167
di-	162	159
ke-	30	31
pra-	18	14
anti-	8	5
-an	3,770	3,565
-kan	680	673
-i	265	249
*-nya**	225	228
-wan/-wati	61	52
*-lah**	17	23

Table 33: Result of MorphInd Process for Single Affixes

Circumfix	Indonesian	Malay
meN- -kan	9,832	8,364
ke- -an	3,889	3,586
meN- -i	2,598	1,968
per- -an	1,348	1,271
meN-per- -kan	556	392
ber- -an	471	358
di- -kan	202	105
meN-per- -i	109	78
di- -i	66	73
ber- -kan	64	58
ber-ke- -an	42	31
se- -nya	35	49
peN- -an	29	5
ber-peN- -an	23	10
ber- -i	11	11
ter- -kan	9	12
*ter- -nya**	7	7
ber-per- -an	7	7
*ber- -nya**	2	2
ke- -an	2	2
di-per- -kan	2	2
di-per- -i	1	1
*ke- -nya**	1	1
ter- -i	1	1
*meN- -nya**	1	1

Table 34: Result of MorphInd Process for Circumfixes