# Building A Dictionary of Malay Language Part-of-Speech Tagged Words Using Bahasa WordNet and Bahasa Indonesia Resources

**Mohamed Lubani**
University of Malaya, Faculty of Computer Science and Information Technology,
Kuala Lumpur, 50603 Malaysia
mohamed.lubani@siswa.um.edu.my

**Rohana Mahmud**
Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603 Malaysia
rohanamahmud@um.edu.my

## ABSTRACT

Assigning grammatical categories to words in natural text is a vital step in processing natural language. Language resources and text processing tools such as part-of-speech (POS) can be used to assign each word the corresponding grammatical category based on its context. Such resources are available for the major languages such as English, Spanish and Japanese. However, the lack of resources for Malay language makes it very hard to develop new processing tools and contribute to the automation of the language processing. In this paper, a Malay POS dictionary is built using Bahasa wordnet and a POS tagged of Indonesian corpus, as well as a monolingual Malay dictionary. The output is a list of 25,778 Malay POS tagged words where each word is assigned all its possible grammatical categories. The proposed process can also be used as a guideline for future improvements.

**Key Words:** Malay Language, Natural Language Processing, Part-of-Speech Tagging.

## INTRODUCTION

Part-of-speech (POS) tagging is the process of categorizing words into their grammatical classes based on their contexts. This process is considered as one of the major Natural Language Processing (NLP) tasks and is used as input to many advanced NLP tasks including information retrieval and text understanding.

Malay language is classified as an underrepresented language in terms of the available text processing tools and language resources. Unlike resource-rich languages, there is a lack of a Malay POS tagged corpus as well as a lack of gold-standard data for Malay POS tagging.

Unlike conventional POS tagging which is considered as a static tagging with some ambiguities, the grammatical class of Malay words is considered dynamic (Knowles & Don, 2003). There are Malay words that correspond to verbs and are likely to appear as nouns. Similarly some Malay adjectives can occur as verbs or adverbs, even some Malay nouns can appear as prepositions. As mentioned by Knowles and Don (2003), linguists can argue the class of words even in the simplest text of Malay due to the lack of unified standards.

This paper aims to provide a POS tagged Malay dictionary using available language resources for Malay namely Bahasa Wordnet and a POS tagged Indonesian corpus. The paper is structured as follows: Section 2 describes the related work on existing Malay POS tagging; Section 3 highlights the used language resources and the process of building the POS dictionary. A detailed discussion is presented in Section 4 regarding the results. Finally, Section 6 concludes the paper with a discussion on the overall outcome achieved and future research directions.

## RELATED WORK

In (Baldwin & Awab, 2006) an open source toolkit contains text processing tools for Malay language is developed based on pre-existing language resources. This toolkit includes a Malay word and sentence tokeniser developed based on an English tokeniser, a Malay lemmatiser and a partial POS tagger. KAMI Malay English lexicon (Quah, Bond, & Yamazaki, 2001) is used in the process of building the lemmatiser and the partial POS tagger. KAMI lexicon contains 68k word-forms many of which are augmented with their possible POS tags and their corresponding lemma form. Words are extracted from KAMI lexicon and two lists are created: the first contains all words that have a lemma and POS category with a total of around 14k words, and the second list contains all words that have only a POS category with a total of around 31k words. After building the two lists, additional corrections are made using a set of 40 overlapping deaffixation rules to detect and remove specific affix types and phonologically normalize the result. Each deaffixation rule is linked with a specific POS category for its input and output which helps to filter the results. The final results are validated using KAMI lexicon and a list of heuristics to choose the appropriate lemma and POS tags for each word where only three possible POS categories are used: verb (V), noun (N) and adjective (J).[105] The accuracy of the assigned POS tags is not evaluated in (Baldwin & Awab, 2006) due to a lack of gold-standard data for Malay language.

Inspired by the morphological analysis of Malay words proposed in (Baldwin & Awab, 2006), Hamzah and Kamaruddin (2014) manually developed a knowledge base containing of 7,545 words classified based on their morphological properties into six categories: root words, compound words, reduplicate words, “kata luar biasa”, “kata pandu” and affix. All words in each category are tagged with one of four possible POS tags: noun, verb, adjective and adverb. In total, there are 2758 nouns, 1459 verbs, 1102 adjectives and 226 adverbs in the knowledge base. Based on the developed knowledge base, the authors also introduced a set of ordered rules to tag new Malay words based on the morphological properties of the words by looking for the corresponding POS tags in the knowledge base. In addition to the four possible POS tags, the tagger can also detect and label proper nouns and numbers making the list of POS tags contains six possible tags.

Alfred, Mujat and Obit (2013) introduced a set of rules to determine the POS tags for Malay words based on their contexts. To find the correct POS tag, a Malay POS tag dictionary is used at the beginning to assign words all possible POS tags. Then the system applies affixing and word relation rules to narrow down the list of POS tags to only a single tag. The POS tag dictionary used is a list of more than 8,700 POS Malay tagged words extracted manually from Thesaurus Bahasa Melayu (DBP, 2008). The word relation rules are used to assign the correct POS to the word out of 14 possible

[105] The results of the Malay tokeniser and lemmatiser are available on: https://code.google.com/p/malay-toklem/.

tags: noun (NN), verb (VB), adjective (JJ), adverb (RB), direction (DR), preposition (IN), auxiliary (AUX), cardinal number (CD), penekan (PEN), pembenda (BND), conjunction (CC), penguat (GUT), interrogative (WP) and pangkal ayat (PNG). The affixing rules are used if the word is not found in the POS tag dictionary and the word is given the POS tag based on the applied affixing rule. The proposed POS tagger is able to detect the POS tags for Malay words based on their contexts and a set of affixing rules. However, it uses a limited set of rules that doesn't cover all important POS tags such as proper nouns. In addition the affixing rules involve only the prefix, suffix and circumfix without considering the infix which can be found in Malay words.

Another Malay POS tagging method that uses pre-existing language resources is proposed in (Zamin, Oxley, Bakar, & Farhan, 2012) where an unsupervised Malay tagger is introduced. The method translates the Malay sentences into English and uses an English POS tagger called Brill's tagger (Brill, 1992) to tag the translated English version. Then the method uses a Malay English lexicon to map Malay words to their English translations and uses similarity measures to find the closest translation in the translated English version. The Malay words are then assigned the POS tags of their corresponding English words which can be given one of 38 different POS tags. This system can be used for POS tagging because of the limited available language resources for Malay language. However, Malay has a different structure than English and this can be a major error source for the system which gives no considerations to the specific nature of Malay language.

As concluded in (Alfred, Mujat, & Obit, 2013), the availability of a Malay POS tag dictionary is important to most Malay POS tagging systems that apply rules to perform the POS tagging. The following section describes the process of building such a dictionary for Malay language using available language resources.

## METHODOLOGY

In this section, the process of building the POS tagged Malay dictionary is introduced. The main aim is to utilize the available Malay language resources in a novel way in order to build the POS Malay dictionary. First we describe the language resources used in the process, and then the process of building the Malay POS dictionary is described.

## LANGUAGE RESOURCES

The following are the language resources used in the process:

## WORDNET BAHASA

This is a free large scale semantic dictionary for the two Malay languages Malaysian and Indonesian (Bond, Lim, Tang, & Riza, 2014). It contains a large number of words from Malaysian and Indonesian languages categorized in four different grammatical categories: noun (n), verb (v), adjective (a) and adverb (r). Wordnet Bahasa is inspired by Princeton Wordnet and is built collaboratively with the help of a large online community.[106] It also contains contributions from previous projects to build Malaysian and Indonesian Wordnets.

## WORDNET BAHASA STRUCTURE

[106] Princeton WordNet: http://wordnet.princeton.edu/.

Currently available Bahasa Wordnet contains more than 600,000 entries. Each entry has four properties:

1. Synset: which has the form “offset-pos” where “offset” is the index of the corresponding entry from Princeton wordnet 3.0 and “pos” is the grammatical category for this entry.
2. Language: This indicates the language of the entry. Since Bahasa Wordnet merges different previously built Wordnets, the language also refers to the source Wordnet of the entry. This property can be assigned one of the following three possible tags: “B” stands for Bahasa Wordnet and it means that this entry is either taken from Malaysian language or Indonesian language. “I” stands for Indonesian and “M” stands for Malaysian.
3. Goodness: It is estimated that 5-10% of Bahasa Wordnet entries are incorrect.[107] Therefore, entries are given one of five different goodness measures: “Y” means that this entry has been hand checked and good. “O” means that this entry is automatically identified high quality. “M” means that this entry is automatically identified medium quality. “L” means that this entry is automatically identified as bad or poor quality. “X” means that this entry has been hand checked and identified as bad or poor quality.
4. The word: The forth property is the actual word and not necessarily in the lemma form.

Figure 1 shows an actual sample of the Bahasa Wordnet dictionary.

**Figure 1: Sample of the Bahasa Wordnet dictionary where each line represents an entry with four different properties in the following order: synset, language, goodness and the word.**

| | | | |
|---|---|---|---|
| 00035189-n | B | X | perolehi |
| 00035189-n | B | X | selesaikan |
| 00035189-n | B | X | siap |
| 00035189-n | B | Y | kejayaan |
| 00035189-n | B | Y | prestasi |
| 00035189-n | I | O | pemenuhan |
| 00035189-n | I | O | penyempurnaan |
| 00035189-n | M | X | penyempurnaan |
| 00035189-n | M | X | perlaksanaan |
| 00035254-a | B | L | laut bebas |
| 00035254-a | B | L | tertutup |
| 00035254-a | B | L | tutup |
| 00035259-v | B | L | bilik mandi dgn pancuran mandi hujan |
| 00035259-v | B | L | cucur |
| 00035259-v | B | L | hujan |
| 00035259-v | B | L | hujan lebat |

## PROCESSING WORDNET BAHASA

The complete Wordnet Bahasa dictionary is downloaded and saved in an excel file.[108] A special script is written to scan the excel file to first locate unique words with

[107] As mentioned in the disclaimer: http://wn-msa.sourceforge.net/eng/index.html.

[108] The complete dictionary can be downloaded from http://sourceforge.net/p/wn-msa/tab/HEAD/tree/trunk/.

goodness measures not equal to “L” or “X”. Then these unique words are aggregated with their possible POS tags found in the dictionary. The output is a list of 81,362 unique words with at least medium quality and their corresponding POS tags.

## INDONESIAN PART-OF-SPEECH TAGGED CORPUS

The second resource is an Indonesian POS tagged corpus developed as a part of the Pan Localization project which aims to develop language standards and technology across multiple South and South-East Asian countries.[109] Currently there are 11 countries involved in the project.

## INDONESIAN CORPUS STRUCTURE

The Indonesian POS tagged corpus contains more than one million POS tagged tokens. Tokens can be actual words, numbers, symbols or punctuations. These tokens are tagged with one of 37 different POS tags for Bahasa Indonesia introduced in (Pisceldo, Adriani, & Manurung, 2009) and shown in Table 4. In the corpus, tokens are separated by a space and each token is followed by “/” then the corresponding POS tag.

**Table 4: POS tags used in the Indonesian POS tagged corpus.**

| No. | Tag(s) | Description |
|---|---|---|
| 1-8 | ) ( , . : -- “ ” | punctuations |
| 9,10 | $, Rp | Currency (dollar, Rupiah) |
| 11 | SYM | Symbols |
| 12 | NNC | Countable common nouns |
| 13 | NNU | Uncountable common nouns |
| 14 | NNG | Genitive Common nouns |
| 15 | NNP | Proper nouns |
| 16 | PRP | Personal pronouns |
| 17 | PRN | Number pronouns |
| 18 | PRL | Locative pronouns |
| 19 | WP | WH-pronouns |
| 20 | VBT | Transitive Verbs |
| 21 | VBI | Intransitive Verbs |
| 22 | MD | Modal or auxiliaries verbs |
| 23 | JJ | Adjectives |
| 24 | CDP | Primary cardinal numerals |
| 25 | CDO | Ordinal cardinal numerals |
| 26 | CDI | Irregular cardinal numerals |
| 27 | CDC | Collective cardinal numerals |
| 28 | NEG | Negations |
| 29 | IN | Prepositions |
| 30 | CC | Coordinate conjunction |

[109] Pan Localization project: http://www.panl10n.net/ (on going).

| | | |
|---|---|---|
| 31 | SC | Subordinate conjunction |
| 32 | RB | Adverbs |
| 33 | UH | Interjections |
| 34 | DT | Determiners |
| 35 | WDT | WH-determiners |
| 36 | RP | Particles |
| 37 | FW | Foreign words |

## PROCESSING INDONESIAN CORPUS

The tagged corpus is freely available on the Pan Localization project website.[110] The corpus is downloaded and processed using a special script written to read tokens from the corpus. Numbers and non-alphabetical tokens are ignored in this process resulting of around 721,154 possible words. Each of these words is associated with all its POS tags found in the corpus. The output is a list of 35,473 unique words with their corresponding POS tags.

## MALAY LANGUAGE DICTIONARY

The third and final resource used is a Malay Language dictionary to be used to verify and distinguish Malay words. In the proposed process, the widely used Malay dictionary “Kamus Dewan” published by Dewan Bahasa dan Pustaka is used in its third edition (DBP, 1996).

## KAMUS DEWAN STRUCTURE

This dictionary contains more than 36,000 entries. Each entry is provided with a description that explains the meaning, as well as possible synonyms and abbreviations. Figure 2 shows the structure of this dictionary.

**Figure 2: The structure of Kamus Dewan Malay dictionary.**

[110] One Million POS Tagged Corpus of Bahasa Indonesia: http://www.panl10n.net/english/OutputsIndonesia2.htm

**efek II** (éfék) IB surat yg boleh dicagarkan (spt bon, surat saham, dll).

**efektif** (éféktif) ada kesannya (pengaruhnya), berkesan, mujarab (bkn ubat): *satu sistem perkhidmatan awam yg ~ tidak akan wujud semata-mata kerana pelaksanaan disiplin yg kuat sahaja.*

**efemeral** (éfemeral) (Bio) bkn tumbuhan yg menyelesaikan kitar hidupnya dlm jangka masa yg pendek, terutama tumbuhan di padang pasir.

**efendi** (éféndi) (Turki) tuan (sebutan bagi golongan bangsawan).

**efisien** (éfisien) bkn individu, organisasi, mesin, dll yg boleh menjalankan tugas dgn cekap (tanpa pembaziran waktu dan tenaga);

## PROCESSING KAMUS DEWAN

This dictionary is downloaded as multiple Microsoft word files corresponding to different alphabetical letters.[111] These files are processed and only words in bold are extracted and saved in a text file. Numbers in bold such as “I”, “II”, “1”, “2” will be excluded from the detected words. The output is a list of 54,152 extractions including individual letters and synonyms and abbreviations of all entries.

## BUILDING MALAY PART-OF-SPEECH DICTIONARY

The list of words extracted from Bahasa Wordnet along with their possible POS tags contains 81,362 unique words. Also the list of words extracted from the POS tagged Indonesian corpus contains 35,473 unique words with their corresponding POS tags. Both lists contain large number of Malay words which need to be extracted to build the Malay POS dictionary.

To extract only Malay words from these lists, two conditions are used. The first is to check the language property of the word in the Bahasa Wordnet looking for the type “M”. The second condition is to look for the word in the 54,152 extractions from the Malay dictionary. If the word satisfies either the first or the second condition, it will be extracted as a Malay word along with its corresponding POS tags from both lists. The output is a list of 25,778 Malay words with their possible POS tags. Table 5 shows a sample of the built Malay POS dictionary.

[111] Kamus Dewan (.doc): https://lamanbahasa.wordpress.com/kamus/.

**Table 5: Sample of the built Malay POS dictionary.**

| Malay Word | Tags from Indonesian corpus | Tags from Wordnet |
|---|---|---|
| Aa | NN | - |
| Ab | NN | n |
| Aba | NN | n |
| Abad | NN | n |
| Abadi | VBI, NN | r, a, n |
| Abah | - | n |
| Abah-abah | - | n |
| Abai | - | n, a |
| Abalone | - | n |
| Abang | - | n |
| Abar-abar | - | n |
| Abc | NN | n |
| Abd | NN | - |
| Abdi | - | n |
| Abdomen | - | a, n |
| Abiad | - | a |
| Abiaz | - | a |

## DISCUSSION

As shown in Table 5, the built Malay POS dictionary consists of three columns: the word-form, the corresponding POS tags from the Indonesian tagged corpus and the corresponding tags from Wordnet. A more detailed look at the output shows that there are many words tagged with similar tags from both sources. It is also noticed that tags from Wordnet cover all possible grammatical slots that fit a specific word. Indonesian tagged corpus helps to discover the most frequently used words and their most used grammatical classes which may not have a complete coverage of all possible contexts like Wordnet. The use of Wordnet tags helps to assign POS tags to rare Malay words as well as complete the missing contexts in the Indonesian corpus. The results of the built POS dictionary are yet to be evaluated by a linguist. However, assuming that both the Wordnet and the Indonesian tagged corpus have been previously checked and evaluated, the results are considered to be of good accuracy.

Taking only entries with at least medium quality from Wordnet reduces the number of extracted words in the results. However, this considered to be better than considering all entries regardless their goodness which may lead to inaccurate extractions and bad POS assignments.

As previously mentioned, Malay language has its special nature which is different than other languages. A carefully built POS tag set maybe needed to tag Malay words rather than using existing grammatical classes from other languages. Using the tag set from Wordnet (which is very limited) or from the Indonesian corpus can be seen as an intermediate step towards mapping Malay words to their POS tags. However, as

mentioned in section 2, there is no standard Malay POS tag set available and Malay linguists are using different groups of POS tags.

## CONCLUSION

In this paper, a process of building a Malay POS dictionary based on available language resources is introduced. The results can be used as input to many rule-bases Malay POS taggers to narrow down the tags for a given word based on its context. The proposed process can be used as a guideline to build richer Malay POS dictionaries using other resources and a richer Malay monolingual dictionary.

In the future, a list of Malay root words with their POS tags can be built. Also the set of Malay grammatical classes which is considered as a property of the Malay language can be built and the results of the proposed process are to be mapped to the target Malay POS tag set.

## REFERENCES

Alfred, R., Mujat, A., & Obit, J. H. (2013). A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles. *Intelligent Information and Database Systems*, (pp. 50-59). Berlin, Heidelberg.

Baldwin, T., & Awab, S. (2006). Open Source Corpus Analysis Tools for Malay. *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, (pp. 2212-5). Genoa, Italy.

Bond, F., Lim, L. T., Tang, E. K., & Riza, H. (2014). The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia, 57*, 83 -100.

Brill, E. (1992). A simple rule-based part of speech tagger. *Applied natural language processing*, (pp. 152-155). Stroudsburg, PA, USA.

DBP, D. B. (1996). *Kamus Dewan, Third Edition.* Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

DBP, D. B. (2008). *Tesaurus bahasa Melayu Dewan.* Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Hamzah, M. P., & Kamaruddin, S. f. (2014). PART OF SPEECH TAGGER FOR MALAY LANGUAGE BASED ON WORDS MORPHOLOGY. *International Symposium on Research in Innovation and Sustainability (ISoRIS '14)*, (pp. 1499-1502). Malacca, Malaysia.

Knowles, G., & Don, Z. M. (2003). Tagging a corpus of Malay texts, and coping with 'syntactic drift'. *Proceedings of the corpus linguistics 2003 conference*, (pp. 422-428). Lancaster.

Pisceldo, F., Adriani, M., & Manurung, R. (2009). Probabilistic Part Of Speech Tagging for Bahasa Indonesia. *International MALINDO Workshop, Colocated Event ACL-IJCNLP.* Singapore.

Quah, C. K., Bond, F., & Yamazaki, T. (2001). Design and construction of a machine-tractable Malay-English lexicon. *In Asialex 2001 Proceedings*, (pp. 200–205). Seoul, Korea.

Zamin, N., Oxley, A., Bakar, Z. A., & Farhan, y. A. (2012). A Lazy Man's Way to Part-of-Speech Tagging. *Knowledge Management and Acquisition for Intelligent Systems*, (pp. 106-117). Berlin, Heidelberg.