# Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
ard51@ui.ac.id, fam.rashel@ui.ac.id, andry.luthfi@ui.ac.id, maruli@cs.ui.ac.id

***Abstract*—We describe our work on designing a linguistically principled part of speech (POS) tagset for the Indonesian language. The process involves a detailed study and analysis of existing tagsets and the manual tagging of an Indonesian corpus. The results of this work are an Indonesian POS tagset consisting of 23 tags and an Indonesian corpus of over 250.000 lexical tokens that have been manually tagged using this tagset.**

***Part of speech tagset; POS; Indonesian***

## I. INTRODUCTION

The part of speech (POS) of a word, also known as its grammatical category, is an indicator of the syntactic and morphological behaviour of a particular lexical item. Knowing whether a word is a noun, verb, adverb, or adjective can provide valuable linguistic insight, and thus POS tagging, the process of assigning a POS tag to a word, is a fundamental process that supports almost all NLP applications.

There has been some previous work on Indonesian POS tagsets and automatic POS taggers [1][2][3], but to our knowledge the tagsets have only been tested on very small manually tagged corpora of around 15.000 words.

In this paper we present our efforts in designing a linguistically principled POS tagset for the Indonesian language, which involves a detailed study and analysis of existing tagsets and manual tagging of an Indonesian corpus of over a quarter of a million words. The results of this work are a POS tagset for Indonesian and a manually tagged Indonesian corpus using this tagset. These results can be used for further work on Indonesian NLP, such as developing statistical and rule-based automatic POS taggers.

The process of designing POS tagset is divided into two phases. The first phase is defining initial POS tagset. The second phase is testing and completing POS tagset which involves manually tagging Indonesian corpus.

## II. ANALYSIS AND DESIGN OF INITIAL PART OF SPEECH TAGSET

We first analyzed and compared POS tagsets from various previous works and consulted authoritative Indonesian grammar references [4][5]. In particular, two POS tagsets that we base our design on are from Adriani et al. and Larasati et al. [2][6]. These tagsets can be seen in Table I. Adriani et al. proposed two variants of tagsets, one with 37 tags and a reduced one with 25 tags, both of which are originally based on the Penn Treebank tagset [7], hence the use of the same abbreviations for the tag labels. Larasati et al. do not define a POS tagset *per se*, but instead develop an Indonesian morphological analyzer that uses 19 lemma tags.

Our guiding principle in designing a tagset was that we wanted to maintain useful linguistic distinctions whilst reducing the manual effort that would be required by the annotators.

From this principle we derived a desiderata for our tagset as follows:

1. **Linguistically valuable**. The tagset should be able to characterize morphological and syntactic behaviour of lexical tokens appearing in the context of a text that would be useful for various linguistic analyses and natural language processing tools.
2. **Simplicity**. Given that we aim to manually tag a substantial corpus of roughly 250.000 words, we would like to maintain a small number of tags in the tagset, so as to reduce the cognitive load on the annotators. Tagsets vary greatly in size, e.g. for English the Penn Treebank tagset has 36 tags, whereas the Claws C7 tagset has a size of 146 tags.
3. **Automatically refinable**. Our goal is to encapsulate the level of linguistic decision that requires human judgment, where any further refinement could plausibly be done by an automatic procedure sometime in the future. For example, one could imagine a very broad single category of *Noun*, where any nominal word was assigned that same tag. However, to distinguish between, say proper nouns and regular nouns, would require human judgment. Thus, we maintained separate tags for these two categories. On the other hand, one could imagine specifying separate tags for singular vs. plural nouns. Given that plurality is clearly marked through the morphological process of reduplication, this distinction can be further refined in the future through an automatic analysis, e.g. using Indonesian morphological analyzers [8], and thus we do not maintain separate tags for this. As another example, with respect to the various forms of verbs, we chose to simply collapse them into a single tag, given that automatic morphological analysis can reveal much more details about verbs if needed.

We eventually chose the 25-tag tagset in [2] as the template for our tagset, and made the following modifications:

978-1-4799-5330-1/14/$31.00 ©2014 IEEE

TABLE I. POS Tagsets from previous works

| Adriani et al. [2] | Larasati et al. [6] |
|---|---|
| CC (coordinate conjunction) | H (coordinating conjunction) |
| CD (cardinal numerals) | C (numeral) |
| | B (determiner) |
| FW (foreign words) | F (foreign word) |
| IN (prepositions) | R (preposition) |
| JJ (adjectives) | A (adjective) |
| MD (modal or auxiliaries verbs) | M (modal) |
| NEG (negations) | G (negation) |
| NN (common nouns) | N (noun) |
| NNP (proper nouns) | |
| PR (common pronouns) | |
| PRP (personal pronouns) | P (personal pronoun) |
| RB (adverbs) | D (adverb) |
| | T (particle) |
| SC (subordinate conjunction) | S (subordinating conjunction) |
| SYM (symbols) | |
| | I (interjection) |
| VB (verbs) | V (verb) |
| WDT (*wh*-determiners) | |
| WH (WH) | W (question) |
| . (sentence terminator) | |
| , (comma) | |
| : (colon or ellipsis) | |
| ( (opening parenthesis) | |
| ) (closing parenthesis) | |
| " (opening quotation mark) | |
| " (closing quotation mark) | |
| . -- (dash) | |
| | O (copula) |
| | X (unknown) |
| | Z (punctuation) |

1. Added tags for determiners, particles, and interjections, which are missing (but are present in [6]).
2. Following [6], we also added an "unknown" tag for tokens that cannot be resolved to any other tag, e.g. typos and erroneous entries.
3. Removed WDT (*wh*-determiners), which we consider to be a tag that was erroneously copied over from the Penn Treebank tagset in [2], as it is quite particular to English.
4. Following [6], we also conflated the 8 punctuation tags into one punctuation tag, as the refinement can be automatically performed in the future if deemed necessary.
5. We did not include a separate copula tag as in [6] as we believe that for Indonesian, its syntactic behaviour is consistent with verbs in general and thus does not warrant a separate tag.

As a result, we defineed 22 POS tags as our initial POS tagset.

## III. DATA DESCRIPTION

In developing our tagset and manually tagged corpus, we used the Indonesian sentences from the IDENTIC parallel corpus [9]. This corpus consists of 27.325 pairs of Indonesian and English sentences originating from the Penn Treebank corpus that were translated into Indonesian, newspaper articles in economy, international news, science, and sports from the PAN Localization project output[1], and movie subtitles. We chose this resource as it is readily available under a Creative Commons licence, and has already been enriched with morphological tags using the MorphInd morphological analyzer, although it appears that no contextual disambiguation has been performed. Nevertheless, these morphological tags provide a good starting point for analysis.

## IV. TESTING AND REVISIONS OF TAGSET

The process of testing and improving the POS tagset consists of several steps. In this section we present our methodology. The process involves manually tagging the Indonesian corpus by two human annotators. The POS tagset is tested by assigning the POS tags to words in the corpus. During this manual tagging process, issues and conditions that arise about POS tags for particular words are noted. These issues are then discussed to find the solutions and to refine the tagset.

*1) The first step:* Manually tag the first 100 sentences of the IDENTIC corpus using the intial POS tagset. Each word in is tagged mainly by seeing its syntactic context, i.e. grammatical environment of the word. Several issues arose during this manual tagging, such as the appropriate POS tags for currency symbols, abbreviations, television shows, movie titles, and book titles, whether in Indonesian or in a foreign language. The solutions of issues found during this initial tagging were agreed upon by the annotators and the POS tagset was revised accordingly.

*2) The second step:* The revised tagset was used to retag the first 100 sentences that was initially tagged. During this step, besides manually retagging the data, possible subcategories of each part of speech were also noted, as we considered the possibility of decomposing each part of speech into subcategories and defining different tags for each subcategory.

*3) The third step:* Define subcategories of parts of speech. The possible subcategories that were noted in the previous step are completed by consulting authoritative grammar references [4][5]. The output of this step is a revised tagset complete with subcategories of each part of speech.

*4) The fourth step:* Evaluate subcategories of parts of speech. Subcategories of parts of speech defined during the previous step are evaluated. There are three evaluation results:

*a) No different tag for each subcategory:* We do not define different tags for each subcategory based on two assumptions: (1) defining different tags for each subcategory will result in an excessive number of POS tags that will complicate the manual tagging process and (2) some subcategories have too few members that render it unnecessary to define different tags for those subcategories.

*b) Keep the definition and characteristics of all subcategories in part of speech tagset:* We keep the information of all subcategories in the tagset so we can determine the subcategory membership of a tagged word by looking at the POS tag of that word.

[1] http://www.panl10n.net/indonesia/

*c) Add new part of speech tag:* Based on our analysis, new POS tag, NND, is added to the tagset. It marks classifiers, partitives, and measurement nouns. Classifiers classify nouns into particular noun classes [4][5]. Partitives indicate a particular amount of something based on the way it is measured, assembled, or processed [5]. Measurement nouns refer to size, distance, volume, speed, weight, or temperature [5].

*5) The fifth step:* Manually tag the first 5.000 Indonesian sentences of the IDENTIC corpus. All issues that are found during this step are noted and discussed to find the solutions. During this step, the methodology for manually tagging the corpus done by following the procedure described below.

*a) Assign a part of speech tag to a word without seeing the syntactic context of that word:* This is the first process that is carried out. This is done for (1) words which have only one possible POS tag based on [4][5][10] and (2) words which are identified as a verb, noun, or foreign word by human annotators.

*b) Assign a part of speech tag to a word by seeing the syntactic context of that word:* This is the second thing process that is carried out. This is done for (1) words which have more than one possible POS tags based on [4][5][10], (2) words which are identified as proper nouns by human annotators, and (3) ambiguous words. Human annotators may look up the part of speech of those words in an Indonesian dictionary [11] to help determine the POS tag of the word. If the POS tag is still undetermined, the word in question will be tagged with X, i.e. an unknown tag.

*c) Evaluate and correct the tagged data by studying the context:* This is the last process that is carried out, and entails another pass of the entire data.

*6) The sixth step:* During this step the next 5.000 Indonesian sentences from the IDENTIC corpus are manually tagged, following the same procedure as described during the previous step.

*7) The seventh step:* Finally, we evaluate the manually tagged data and complete the POS tagset. Human annotators evaluate and correct all manually tagged data, which currently consists of the first 10.000 Indonesian sentences of the IDENTIC corpus, containing 262.330 lexical tokens. After being evaluated and corrected, all tagged data is sorted by POS tag. All words with the same POS tag are then grouped and listed in the definition of the POS tag as the members of that part of speech. The definition and characteristics of each part of speech are evaluated and detailed. The output of this step is considered to be the final resulting tagset.

## V. RESULT

We have produced two main outputs as the result of this work: (1) a linguistically motivated POS tagset for Indonesian that has been thoroughly verified against a substantial corpus and (2) a manually tagged Indonesian corpus consisting of over a quarter of a million lexical tokens.

### A. Part of speech Tagset

The final version of the POS tagset we design in this work consists of 23 POS tags. A summary of this tagset can be seen in Table II. The detailed subcategorization of these 23 tags is not presented due to space constraints, but detailed subcategorization analysis is planned for future work involving the construction of an automatic POS tagger.

TABLE II. Final proposed Indonesian tagset

| Tag | Description | Example |
|---|---|---|
| CC | Coordinating conjunction | dan, tetapi, atau |
| CD | Cardinal number | dua, juta, enam, 7916, sepertiga, 0,025, 0,525, banyak, kedua, ribuan, 2007, 25 |
| OD | Ordinal number | ketiga, ke-4, pertama |
| DT | Determiner / article | Para, Sang, Si |
| FW | Foreign word | climate change, terms and conditions |
| IN | Preposition | dalam, dengan, di, ke, oleh, pada, untuk |
| JJ | Adjective | bersih, panjang, hitam, lama, jauh, marah, suram, nasional, bulat |
| MD | Modal and auxiliary verb | boleh, harus, sudah, mesti, perlu |
| NEG | Negation | tidak, belum, jangan |
| NN | Noun | monyet, bawah, sekarang, rupiah |
| NNP | Proper noun | Boediono, Laut Jawa, Indonesia, India, Malaysia, Bank Mandiri, BBKP, Januari, Senin, Idul Fitri, Piala Dunia, Liga Primer, Lord of the Rings: The Return of the King |
| NND | Classifier, partitive, and measurement noun | orang, ton, helai, lembar |
| PR | Demonstrative pronoun | ini, itu, sini, situ |
| PRP | Personal pronoun | saya, kami, kita, kamu, kalian, dia, mereka |
| RB | Adverb | sangat, hanya, justru, niscaya, segera |
| RP | Particle | pun, -lah, -kah |
| SC | Subordinating conjunction | sejak, jika, seandainya, supaya, meski, seolah-olah, sebab, maka, tanpa, dengan, bahwa, yang, lebih ... daripada ..., semoga |
| SYM | Symbol | IDR, +, %, @ |
| UH | Interjection | brengsek, oh, ooh, aduh, ayo, mari, hai |
| VB | Verb | merancang, mengatur, pergi, bekerja, tertidur |
| WH | Question | siapa, apa, mana, kenapa, kapan, di mana, bagaimana, berapa |
| X | Unknown | statemen |
| Z | Punctuation | "...", ?, . |

### B. *Manually Tagged Indonesian Corpus*

The first 10.000 Indonesian sentences of the IDENTIC corpus, containing 262.330 lexical tokens, have been manually tagged in this work to test and complete our POS tagset. This corpus will shortly be made freely available online and under a Creative Commons licence.

## VI. CONCLUSION

We have designed a linguistically motivated POS tagset for the Indonesian language. The design process is divided into two phases: (1) define initial POS tagset and (2) test and revise POS tagset which involve manually tagging the Indonesian sentences in the IDENTIC corpus. Our manually tagged Indonesian corpus consists of 10.000 sentences.

The results of our work can be used for further work on Indonesian NLP, such as developing rule-based or statistical-based POS taggers for the Indonesian language.

## REFERENCES


[1] S. Sari, H. Hayurani, M. Adriani, and S. Bressan, "Developing Part-of-Speech Tagging for Bahasa Indonesia", In Proceedings of the 2nd International MALINDO Workshop. Cyberjaya, Malaysia, 12-13 June 2008.

[2] M. Adriani, R. Manurung, and F. Pisceldo, "Statistical Based Part Of Speech Tagger for Bahasa Indonesia", in Proceedings of the 3rd International MALINDO Workshop, Co-located Event ACL-IJCNLP 2009. Singapore, August 1, 2009.

[3] A.F. Wicaksono and A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia", In Proceedings of the 4th International MALINDO Workshop, Jakarta, August, 2010.

[4] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. M. Moeliono, Tata Bahasa Baku Bahasa Indonesia, 3rd ed. Jakarta: Balai Pustaka, 2003.

[5] J. N. Sneddon, A. Adelaar, D. N. Djenar, and M. C. Ewing, Indonesian Reference Grammar, 2nd ed. Crows Nest: Allen & Unwin, 2010.

[6] S. D. Larasati, V. Kuboň, and D. Zeman, "Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus", in Proceedings of the Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011). August 2011. Zurich, Switzerland

[7] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics, Vol. 19, No. 2, pp. 313—330, 1993.

[8] F. Pisceldo, R. Mahendra, R. Manurung, and I W. Arka, "A Two-Level Morphological Analyser for the Indonesian Language", in Proceedings of the Australian Language Technology Association (ALTA) Workshop. Tasmania: CSIRO ICT Center, December 8-10 2008.

[9] S. D. Larasati, "IDENTIC Corpus: Morphologically Enriched Indonesian-English Parallel Corpus", in Proceedings of LREC 2012. May 2012. Istanbul, Turkey

[10] H. Kridalaksana, "Kelas Kata dalam Bahasa Indonesia", 2nd ed. Jakarta: PT Gramedia Pustaka Utama, 2007.

[11] Pusat Bahasa, Kamus Besar Bahasa Indonesia Pusat Bahasa, 4th ed. Jakarta: PT Gramedia Pustaka Utama, 2008.