Word classes in Indonesian: A linguistic reality or a convenient fallacy in natural language processing?

Meladel Mistica[a], Timothy Baldwin[b], and I Wayan Arka[a]

[a] CHL, The Australian National University,
Canberra, Australia
meladel.mistica@gmail.com
wayan.arka@anu.edu.au
[b] CSSE, The University of Melbourne,
Parkville, Australia
tb@ldwin.net

Abstract. This paper looks at Indonesian (Bahasa Indonesia), and the claim that there is no noun-verb distinction within the language as it is spoken in regions such as Riau and Jakarta. We test this claim for the language as it is written by a variety of Indonesian speakers using empirical methods traditionally used in part-of-speech induction.

In this study we use only morphological patterns that we generate from a pre-existing morphological analyser. We find that once the distribution of the data points in our experiments match the distribution of the text from which we gather our data, we obtain significant results that show a distinction between the class of nouns and the class of verbs in Indonesian. Furthermore it shows promise that the labelling of word classes may be achieved only with morphological features, which could be applied to out-of-vocabulary items.

Keywords: Indonesian, Word class, Word class induction, Morphology

1 Introduction

The notion of word classes, such as the nouns, verbs and adjectives, is fundamental in both linguistics and computational linguistics. Word classes are the basis for the labels in part-of-speech tagging, and also the building blocks for parsing. In grammar engineering, they are the primitives upon which context-free grammar rules are written. In linguistics, they are considered the categories that shape the organisation of the language, and the way the world is conceived through language. These categories may not align across languages: what is expressed as a verb in one language may be expressed as an adjective or noun in another. But one linguistic universality hypothesis that remains despite these variations is that the categories *noun* and *verb* exist in all languages (Croft, 2003).

This paper examines the noun–verb distinction specifically for Indonesian (*Bahasa Indonesia*) as, at least in spoken Indonesian in particular regions such as Riau and Jakarta, it has been claimed that open class categories are indistinguishable (Gil, 2001; Gil, 2010). If correct, this would refute the universality of the noun–verb distinction in Linguistics, with significant implications for Linguistic Typology and Theoretical Syntax.

The primary aim in this paper is to apply unsupervised data-driven analysis of Indonesian text to determine whether we can automatically learn the noun–verb distinction, and in doing so, shed light on whether Indonesian conforms to Croft's noun–verb universality hypothesis, or is indeed a counter-example to the hypothesis as claimed by Gil. In addition, these experiments are a litmus test to see if "morphological signatures", as used by Goldsmith (2001), are sufficient in

determining parts-of-speech, and can be used to determine the part-of-speech of out-of-vocabulary items (OOV) in Indonesian text processing. One assumption we make throughout this work is that the same basic word class distinctions are invariant across different styles, genres and registers (such as spoken vs. written) of a given language. That is, a conclusion on word class distinctions drawn based on written data should apply equally to spoken data, for example.

This paper is laid out as follows. In Section 2 we look at how word classes are determined, both qualitatively in Linguistics and statistically in the field of part-of-speech induction. Next, we look at the formal properties of Indonesian, and we give examples to show how the distinction between nouns and verbs can be illusive. In the next two sections (Sections 4 and 5) we describe the data, tools and method for our experiments, followed by the results in Section 6. Finally, we discuss our findings and the impact and contribution to both Linguistics and Computational Linguistics, particularly in natural language processing for Indonesian (Sections 7 and 8).

2 Determining Word Classes

In the following sections, we outline the linguistic methodology for determining word classes, and how we emulate this computationally using methods borrowed from part-of-speech induction.

2.1 Linguistic Categories

Within a descriptive framework, the current methodology for determining word classes is rather formal, relying on the combinatorics of the form. At the clausal or phrasal level, we look at the syntagmatic possibilities of the units within the phrase or clause. When looking at the word level, we look at how each of the morphological components combine. The heuristics for this *combinatorics* approach are outlined by Evans (2000), which takes into consideration semantic properties as a way of labelling these classes rather than determining them.

There are of course problems in taking an overly formal approach, without any reference to the semantics of the word or stem, or its function, as outlined by Croft (2000), using the English adjective class as a case in point: the adjective *big* in its superlative form is *biggest*, whereas the adjective *beautiful* cannot combine with the *-est* suffix to form a superlative. Instead, it must be formed analytically with the adverb *most*, as in *the **most beautiful** sunset*. Even given this fact about these two stems/words, it may be difficult to argue that these are not of the same word class because in this instance they do not combine in the same way morphologically. One could concede that these form sub-classes of adjectives.

Our assumption is that the two aforementioned levels, phrasal/clausal and word level, to some extent have their own localised characteristics and constraints on combination. However the predictions we make at the word level, by and large, have consequences at the phrasal/clausal level, and vice versa. For example, if we decide that, at the word level, the stem *stool* is a noun, we then assume that within a phrase it will behave like other nouns and therefore form a group with other terms that are like it; i.e., it will combine with other terms in the same way that other nouns do.

In an article that led to many responses, Evans and Osada (2005) argue against the claim that Mundari (an Austroasiatic language from India) has no noun–verb distinction. They outline three criteria for establishing the lack of word classes within a language, as summarised in Figure 2.1. Additionally, Evans and Osada state that these criteria must apply *exhaustively* throughout the language. We address each of the issues in Figure 2.1 for Indonesian in Section 3 to show how one could possibly analyse Indonesian as being a language that had no noun–verb distinction.

2.2 Part-of-speech Induction

According to Biemann (2009), part-of-speech induction aims to approximate syntactic labels based on unlabelled token sequences, usually resulting in the discovery of categories that are not considered linguistically motivated. It is traditionally approached using unsupervised methods,

I. EQUIVALENT COMBINATORICS
"Members of what are claimed to be merged classes should have identical distributions in terms of both morphological and syntactic categories."
II. COMPOSITIONALITY
"Any semantic differences between the uses of a putative 'fluid' lexeme in two syntactic positions (say argument and predicate) must be attributable to the function of that position."
III. BIDIRECTIONALITY
"[T]o establish that there is just a single word class, it is not enough for Xs to be usable as Ys without modification: it must also be the case that Ys are usable as Xs."

Figure 1: Criteria for determining word classes by Evans and Osada (2005)

otherwise known as clustering, and then assigning labels after the fact. However in this study we use linguistic features in order to induce linguistically motivated clusters.

A recent survey of systems developed for part-of-speech (POS) induction by Christodoulopoulos *et al.* (2010) shows that these systems are notoriously difficult to evaluate due to the nature of the task. Rather than focusing on the individual learners developed in these systems, for our purposes, we observe that two main feature types were utilised in all the systems that were surveyed: morphological, and collocational (or pseudo-syntactic). Of the seven systems surveyed, five developed systems based on solely on collocational or distributional properties of tokens in a text, while two systems (Clark, 2003; Biemann, 2006) additionally included morphological features.

Relevant to the morphological features we develop in this study is the research of Goldsmith (2001), and his use of the term **signatures** to mean a collection of morphological patterns relevant to a stem. He employs the MDL (minimum description length) algorithm as a way of discovering and grouping verbs that inflect in the same way. Although this field focuses on collocational or distributional features that emulate syntactic position, Goldsmith's work focuses solely on morphological patterns as a way of grouping *inflectional categories*. Rather than inducing the signatures in the way the Goldsmith does, we derive them from a corpus using a non-class biased morphological analyser (See Section 5 for more details).

3 Formal Properties of Indonesian

In this section, we discuss the properties of Indonesian in terms of the criteria set up by Evans and Osada (2005), as summarised in Section 2.1. We show how the combinatorics of lexical items could lead one to analyse the language as having one catch-all open class lexical category, i.e. have no distinction between nouns and verbs or other open word classes such as adjectives or adverbs.

First we address the issue of *equivalent combinatorics*. The following examples focus on the word *lari* "run". As is evident in Examples (2) and Example (1) , *lari* can occur in both the *subject* and the *predicate* position.

(1) *Ia* *lari.*
(S)he run
"(S)he runs."

(2) *Lari* *menyehatkan.*
run cause.to.be.healthy
"Running is healthy."

The examples below employ the stems *bunyi* "sound" and *bangun* "to wake up", and show how words that can be traditionally thought of as noun and verb, respectively, can occur in the same morphological environment, by combining with the same morphological affixes:

(3) *membunyikan*
mem-bunyi-kan
AV-sound-CAUS
"make X make a sound (instrument)"

(4) *membangunkan*
mem-bangun-kan
AV-wake.up-CAUS
"make X wake up (someone)"

Take first the issue of *Compositionality* from Section 2.1, which states that given a word in a position, we should be able to predict its semantics. We see that in Examples (1) and (2), the predicative or referential function of *lari* "run" is simply determined by its position in the clause, unlike English that requires us to employ derivational morphology.[1]

In Examples (3) and (4), the semantic and morpho-syntactic effects of the affixes are predictable: the suffix *-kan* CAUS is a **causative**, meaning that it will have causative semantics associated with it. The prefix *mem-* AV simply tells us that the agent is the subject, which would be true for both examples above. In other words these examples satisfy the semantic predictability criterion stipulated by the *Compositionality* requirement.

Testing the criterion of *Bidirectionality* is not a trivial task simply because we would have to rely on grammaticality judgements, which can be subjective and continuous rather than binary in nature (Keller, 2001). Assuming that there is no distinction between open word classes in Indonesian, this criterion states that those classes traditionally labelled as nouns must be able to behave as all other parts of speech in the open class category, and vice versa. We have an example of what would be traditionally analysed as a verb in a syntactic slot normally reserved for nouns in Example (2) . However, we would also have to find possibilities of traditional nouns in verbal positions, without the need for morphological affixation to license its usage in that position.

Rather than rely on grammaticality judgements, we analyse how the language is used by the Indonesian speaking population in publicly available web data.

4 Data and Tools

This section outlines how we gathered our data and developed our data set, obtained our evaluation data, and modified existing tools for the task in order to create our morphological features.

4.1 Getting the Word-forms for Morphological Analysis

The text we use for this study is the Indonesian Wikipedia[2] because not only is it a large source of naturally-occurring text, but also because the data is produced and curated by many authors; it is representative of the way the language is used throughout the Internet-connected areas of Indonesia, and Indonesian speakers throughout the world.

We gathered approximately 2 million Indonesian tokens from Wikipedia articles. After removing markup using WikiPrep[3], we sorted the data and counted token frequencies, leaving capitalisation untouched and preserving any errors in the removal of mark-up. We then ran the morphological analyser over all tokens that occurred 5 or more times, mainly to eliminate typographic or other errors. In actuality this is only around 17% of all word forms found, with a long tail of singleton occurrences.

4.2 Stem lexicon

The method we employ in discovering the word class clusters is fully unsupervised, and does not incorporate any possibly biasing information on word class distinctions. In order to ascertain whether the induced word clusters reflect the noun–verb distinction, however, we require evaluation data that we consider to encode the conventional definition of noun and verb. For this, we

[1] In English we can say: *I run.* but not: **Run is healthy.*

[2] We used the dump produced on October 15, 2009, downloaded on October 19, 2009 from `http://dumps.wikimedia.org/idwiki/`

[3] `http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep`

PART-OF-SPEECH	COUNT
Noun	8,096
Verb	821
Other (Prep, Pron, Num, etc.)	1,770
TOTAL	10,687

Table 1: Part-of-speech distribution in the stem lexicon for the morphological analyser.

PREFIX	Active Voice; Passive Voice; Causative; Passive *Ter/ter; ber; OrdKe/ke; pe/peN; se*
CIRCUMFIX	*ke_an; per_an; pe/peN_an*
SUFFIX	*an; i; kan; wi*
CLITICS	*ku; mu; nya*
OTHER	Reduplication

Figure 2: Types of affixes from the morphological analyser.

extract the lexicon from a morphological analyser which encodes a hard distinction between the two word classes, in a manner which is compatible with the criteria of Evans and Osada (2005) and word class definitions of Croft (2003). As can be seen in Table 1, the lexicon is biased towards nouns, having almost ten times as many nouns as there are verbs. The class labelled "Other" consists of all stems not marked as a noun or verb, and includes adjectives, pronouns, prepositions, numbers, determiners, etc.

4.3 Morphological Analyser

Due to Indonesian being a morphologically rich language, we need some way of mapping word forms onto lexemes, to capture as complete a signature as possible for each lexeme. For this, we used the same morphological analyser as provided the source of the stem lexicon, but altered the analyser to remove all distinctions between word classes, to avoid biasing the clustering results. For example, we defined the (verbal) circumfix *peN+..+an* (e.g. such as *pem+bakar+an* peN+burn+an, meaning "the process of incinerating") to be able to flank any stem. This leads to uninhibited over-generation, but it also enables the possibility of analysing all classes of stems in a uniform manner.

We obtained the original morphological analyser by Mistica *et al.* (2009), which was built using XFST (Beesley and Karttunen, 2003) and includes the 10,687 lexicalised stems from Table 1. We then merged all stems in the lexicon into a single undifferentiated word class. The set of affixes the morphological analyser can process is outlined in Figure 2.

5 Method

5.1 Feature Engineering

We perform two kinds of experiments, based on two distinct feature representations: multinomial and binomial. Multinomial features take into account the number of occurrences of a morphological pattern, and the binary features record the existence of a particular morphological pattern (irrespective of how often it occurs). Linguists conventionally manually classify lexemes into morpho-syntactic categories using a binomial representation, i.e. testing (often using elicited data or intuition) whether a given lexeme is compatible with each of a range of morpho-syntactic configurations. Computational linguistic methods, on the other hand, tend to use a multinomial lexeme representation. Part of our interest in comparing these two feature representations is to empirically test the relative expressiveness of the two approaches.

		STEM	STEM+S	un+STEM+ed	STEM+ed
bird	a.	10	20	0	0
	b.	1	1	0	0
plug	a.	1	3	5	15
	b.	1	1	1	1

Figure 3: An example of features extracted from *bird*, *birds*, *plug*, *plugs*, *unplugged* and *plugged*.

DATA	# Features	N	V	O	TOTAL
All POS (All parts-of-speech)	259	8,096	821	1,770	10,687
Noun & Verbs	250	8,096	821	-	8,917
All POS, NoRR (No Reduplication)	208	8,096	821	1,770	10,687
Nouns and Verbs, NoRR (No Reduplication)	200	8,096	821	-	8,917

Table 2: Part-of-speech distribution for the 'Clustering' experiments, with number of features.

The morphological patterns are collated from the output of the morphological analyser. That is, we do not just note which suffixes are appended to each stem, but we look at each word form in its entirety, i.e. what combinations of morphs occur with stem. The combination of these morphs make up each feature.

If, for example, we were looking at an English corpus and found the tokens *bird*, *birds*, *plug*, *plugs*, *unplugged* and *plugged* in the corpus 10, 20, 1, 3, 5, 15 times respectively, then the features for this set would be as shown in Figure 3, wherein the **a.** rows indicate the multinomial values and the **b.** rows are the binary values.

5.2 Clustering

For our clustering experiments, we use two basic datasets: (1) lexemes from all word classes (noun, verb and other) in the stem lexicon, and (2) lexemes from just the noun and verb classes in the stem lexicon. We additionally experiment with reduplication either enabled or disabled in the morphological analyser, as it has been claimed that reduplication is the only instance of non-derivational morphology in Indonesian, and leads to marked differences in semantic effect between predicates and nominals (Musgrave, 2001). We hypothesise that this may have an influence on the assignment of instances to clusters, and test this by optionally disabling analysis of reduplication.

Table 2 shows the breakdown of data across the different logic possibilities of word class composition in the stem lexicon, and reduplication in the morphological analyser (with "NoRR" indicating that reduplication is disabled). In each case, "N", "V" and "O" indicate the breakdown of nouns, verbs and other word classes in the dataset in question. We also observe that, of the possible combinations of morphs (from Figure 2) for a given stem, a relatively small number is attested in the actual data (as indicated in the column "# Features").

Rather than apply a hard clustering algorithm such a k-means, we decided to employ a soft probabilistic clustering algorithm, namely the EM algorithm, as a way of modelling the fact that these categorial distinctions are not necessarily absolute. The first component of the EM algorithm (the *Expectation Step*) is the same as the process by which k-means assigns a data point to a cluster, however in the second step of EM (the *Maximisation Step*), the cluster centroids are recomputed such that the likelihood of the parameters of the distributions are maximised. The expectation and maximisation steps are repeated iteratively until the parameters do not change or reach a specified threshold (Tan *et al.*, 2006).

We used the EM implementation in the Weka package,[4] (Witten and Frank, 2005) maintaining the default parameters for the maximum number of iterations ($I = 100$), the minimum allowable standard deviation in log-likelihood ($1e - 6$), and the number of random seeds initially selected ($S = 100$). We ran the EM cluster either with no pre-defined cluster number ($N = -1$, where the number of clusters is dynamically optimised by maximising the log-likelihood) or exactly 2 clusters ($N = 2$).

5.3 Experimental Set-up

We perform 2 groups of experiments, which we term "All Clustering" and "Subsampling". For "All Clustering", we use all available data in a given dataset, across each of the full datasets described in Table 2. The stem lexicon is biased towards nouns, with 75.8% of the stems being nouns in the All POS case, and 90.8% being nouns in the Nouns & Verbs case, as shown in Table 2.

For the "Subsampling" experiments, we aimed for a noun–verb split which is more representative of the split seen in actual text, based on hand-analysing an Indonesian document external to our document collection.[5] For each unique word token in the article (except English words, words in the footer, the menu, and tab items not relevant to the document), we looked up that word form in the Indonesian government's official dictionary *Kamus Besar Bahasa Indonesia* "The Big Indonesian Language Dictionary" (Sugiono, 2008). If the dictionary lists it and gives a POS or multiple POSs, we count oncee for each POS listing. We found that the proportion of nouns to verbs was 65:35 (with 30 noun and 16 verb types). Equipped with this knowledge, we created a sub-sampled dataset with these proportions of nouns and verbs, in two datasets: (1) a 650–350 mix of nouns to verbs, and (2) a 1300–700 mix of nouns and verbs.

As briefly discussed in Section 4.2, we evaluate the discovered clusters against the classes in the original stem lexicon. We do this in accordance with the evaluation methodology adopted in word class induction (Clark, 2003; Biemann, 2006), by first mapping each lexeme to the cluster with the highest probability for that instance. In the 2-class ($N = 2$) clustering case, we assign one cluster noun and one cluster verb, according to the assignment which maximises the F-score relative to the labels in the original stem lexicon; in the $N = -1$ case where the number of clusters is automatically determined, we determine the assignment of labels to classes which maximises F-score (accepting that the same class can be used to label multiple classes) subject to the constraint that each class is used to label at least one cluster.

In all cases, we use computationally-intensive *randomised shuffling* to test statistical significance (Yeh, 2000).

6 Results

6.1 All Clustering

As can be seen in Table 3, the experiments over all parts of speech, fared rather poorly, with F-scores (F) falling below the (supervised) majority class baseline, as can be seen in Table 3. Bear in mind, however, that our primary question is whether nouns and verbs can be distinguished in Indonesian, where these results are over the much harder task of differentiating nouns, verbs and a heterogeneous assortment of closed- and open-class words. We thus move directly on to consider the results for the binary "Noun & Verb" case, and observe that our precision (P), recall (R), and F-score (F) are all above the majority class baseline, for both the multinomial and binomial feature representations.

[4] `http://www.cs.waikato.ac.nz/ml/weka`

[5] The Indonesian *Linguistik komputasional* "Computational Linguistics" stub article, as accessed in May, 2011 at the URL: `http://id.wikipedia.org/wiki/Linguistik_komputasional`; this article was not contained in the 2009 dump of Indonesian Wikipedia.

Data Type	ALL POS $N=-1$			NOUNS & VERBS $N=2$			NOUNS & VERBS $N=-1$		
	P	R	F	P	R	F	P	R	F
Multinomial	.985	.588	.737	.930	.931	.930	.990	.990	.990
Binary	.977	.579	.727	.914	.915	.915	.973	.973	.973
Random	.623	.341	.441	.512	.512	.512	.512	.512	.512
Majority Class	.756	.756	.756	.908	.908	.908	.908	.908	.908
Multinominal (NoRR)	.989	.591	.740	.942	.942	.942	.942	.942	.942
Binominal (NoRR)	.940	.556	.699	.918	.918	.918	.971	.971	.971
Random	.616	.338	.436	.501	.502	.502	.501	.502	.502
Majority Class	.757	.757	.757	.908	.908	.908	.908	.908	.908

Table 3: Results for the "All Clustering" experiments.

Data	$N=2$				$N=-1$					
	650–350	p	**1300–700**	p	**650–350**	N-V	p	**1300–700**	N-V	p
Multinomial	.783	.184	**.788**	**.043**	**.808**	3-1	**.000**	**.788**	1-1	**.044**
−NoRR	.778	.452	**.803**	**.000**	.780	2-1	.490	**.793**	3-1	**.003**
Binomial	**.812**	**.000**	**.825**	**.000**	**.822**	3-3	**.000**	**.848**	5-4	**.000**
−NoRR	**.812**	**.001**	**.828**	**.000**	**.824**	3-4	**.000**	**.849**	3-5	**.000**
Random	.538		.529		.538			.529		
Majority Class	.650		.650		.650			.650		

Table 4: F-score for "Subsampling" experiments

The results we found most surprising were the data that had no morphological reduplication (NoRR). These were consistently outperformed by the systems that had reduplicated stems. This is unexpected because reduplication has different interpretations across word classes in Indonesian, and as mentioned previously is the only morphologically inflectional process (Musgrave, 2001).

At this stage, none of the differences in results over the Majority Class baseline are statistically significant ($p > 0.05$), although all of the differences over the random baseline are significant ($p < 0.001$).

6.2 Subsampling

We next move on to experiments with the subsampling data, which is a better reflection of the relative occurrence of nouns and verbs in actual text. We run two types of experiments: one with a 650–350 mix of nouns to verbs, and another with a 1300–700 mix of nouns to verbs, as described in Section 5.3. This ratio is based on manual analysis of the rate of occurrence of nouns and verbs in a held-out document.

Firstly, all F-scores surpassed the 0.650 majority class baseline, as shown in Table 4. Secondly, almost all of these results were significant based on our random sampling method. As with the previous "All Cluster" experiments, all F-scores exceeded the random baseline.

The column "n-v" indicates how many clusters were found, and how they were merged for evaluation. For example 3-1 indicates that 4 (3+1) clusters were induced, with three of them combining to form the "N" class and the other forming the "V" class.

The boldfaced figures in Table 4 highlight the results that exceed the majority-class baseline at a level of statistical significance ($p < 0.05$). As can be seen, in most cases there was a significant improvement over the baseline for both the smaller and larger datasets. This provides strong weight to the claim that the noun–verb distinction is discernible for Indonesian, refuting the claims

Features with Greatest Probability Density as a Proportion per Class			
Noun		Verb	
STEM+NYA	SE+STEM	STEM+PEN_AN	TER+STEM
REDUP[STEM]	STEM+I+NYA	Passive+STEM+KAN	Active+STEM+KAN+NYA
STEM		Active+STEM+KAN	Passive+STEM+I
BER+STEM		STEM+AN+NYA	STEM+KAN
REDUP[STEM]+NYA		STEM+AN	Active+STEM+I

Figure 4: Morphological patterns associated with nouns and verbs.

of Gil (2001, 2010). With the exception of Multinomial $N - 1$, the results for the larger dataset were slightly better than the smaller dataset, and the binomial feature representation was superior to the multinomial feature representation. Again, the experiments containing no reduplicated stems (NoRR) either did not fare better than those without, or were only slightly better.

7 Discussion

One experimental question we had was whether using only morphological features would suffice in determining word classes, and whether syntactic features were a requirement for this kind of experiment. We were also curious to see (given that our experiments were successful) if there were morphological signatures or a group of morphological patterns that defined a class. For each of our features in the 1300–700 binomial experiment, we collected the accumulated probabilities for each class and compared them to each other. We found that there were about 10 morphological patterns that we could associate with verbs, and 7 with nouns. These are shown in Figure 4.

From a linguistic perspective, one result that we feared was that the multinomial experiments, which took into account the number of times a pattern occurred with a stem, would fare better than the binomial experiments. The implications of this would suggest that the means by which linguists determine word classes, by adding to their inventory of possible combinatorics, would not suffice. However, we see that the binomial features have the same score or do better than the multinomial features, and therefore seeing a token only once, or simply having the knowledge that particular forms are possible, is enough to assist in analysing and determining classes of stems.

These morphological signature experiments look promising in determining the class of unknown words or out of vocabulary items, as a means of extending the lexicons. For future experiments we would like to mix syntactic features with the morphological features used in this study. We would also like to extend the study to see how different the morphological signatures are from one source of text to another.

8 Conclusion

We had designed an experiment that applied the linguistic criteria based on Evans and Osada (2005) in determining word classes for Indonesian, focusing specifically on the question of whether the noun–verb distinction is discernible to an unsupervised word class induction system. The results have shown that there are certainly distinguishable properties between nouns and verbs in Indonesian, even when we restrict ourselves to only examining features at the morphological level. The experiments used solely morphological features based on Goldsmith's (2001) signatures, showing promise that the labelling of word classes may be achieved only with morphological features, with potential application to out-of-vocabulary items.

From a natural language processing perspective, the labels 'nouns' and 'verbs' may be used felicitiously, and not merely as 'convenient labels' in Indonesian text processing. On a broader,

more general note, this study has shown how issues in linguistics can be tackled using methods developed and used in the field of computational linguistics.

References

Biemann, Chris. 2006. Chinese Whisper - An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, New York, USA.

Biemann, Chris. 2009. Unsupervised part-of-speech tagging in the large. *Research on Language and Computation* 7.101–135.

Christodoulopoulos, Christos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 575–584, Massachusetts, USA.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th Annual Meeting of the European Association for Computational Linguistics (EACL)*, 59–66, Budapest, Hungary.

Croft, William. 2000. Parts of speech as language universals and as language-particular categories. In *Approaches to the typology of word classes*, ed. by P. Vogel and Bernard Comrie, 65–102, Berlin, Germany. Mouton de Gruyter.

Croft, William. 2003. *Typology and Universals*. Cambridge University Press.

Beesley, Kenneth and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Evans, Nicholas. 2000. Word classes in the world's languages. In *Morphology: a handbook on inflection and word formation*, ed. by Christian Lehmann and Geert Booij and Joachim Mugdan, 708–732, Berlin, Germany. Mouton de Gruyter.

Evans, Nicholas and Toshiki Osada. 2005. Mundari: The myth of a language without word classes. *Linguistic Typology* 9.351–390.

Gil, David. 2001. Escaping eurocentrism: fieldwork as a process of unlearning. In *Linguistic Fieldwork*, ed. by Paul Newman and Martha Ratliff, chapter 5. Cambridge, UK: Cambridge University Press.

Gil, David. 2010. The acquisition of syntactic categories in Jakarta Indonesian. In *Part of Speech: Empirical and theoretical advances*. John Benjamins Publishing Company.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Association for Computational Linguistics* 27.153–198.

Keller, Frank, 2001. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. The University of Edinburgh dissertation.

Mistica, Meladel, I Wayan Arka, Timothy Baldwin, and Avery Andrews 2009. Double Double, Morphology and Trouble: Looking into Reduplication in Indonesian. In *Proceedings of the 2009 Australasian Language Technology Workshop (ALTW 2009)*, 44–52.

Musgrave, Simon, 2001. *Non-subject Arguments in Indonesian*. Melbourne, Australia: The University of Melbourne dissertation.

Sugiono, Dendy (ed.) 2008. *Kamus Besar Bahasa Indonesia - Pusat Bahasa*. Departemen Pendidkan Nasional, Jakarta, Indonesia: PT Gramedia Pustaka Utama, edisi keempat edition.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Boston, USA: Addison Wesley.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Data Management Systems. Morgan Kaufmann Publishers Inc.

Yeh, Alexander. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th of Conference on Computational Linguistics COLING*, 947–953.