Unsupervised Word Class Induction for Under-resourced Languages: A Case Study on Indonesian

Meladel Mistica
The Australian National University
meladel.mistica@gmail.com

Jey Han Lau and Timothy Baldwin
The University of Melbourne
jeyhan.lau@gmail.com
tb@ldwin.net

Abstract

In this study we investigate how we can learn both: (a) syntactic classes that capture the range of predicate argument structures (PASs) of a word and the syntactic alternations it participates in, but ignore large semantic differences in the component words; and (b) syntactico-semantic classes that capture PAS and alternation properties, but are also semantically coherent (a la Levin classes).

We focus on Indonesian as our case study, a language that is spoken by more than 165 million speakers, but is nonetheless relatively under-resourced in terms of NLP. In particular, we focus on the syntactic variation that arises with the affixing of the Indonesian suffix *-kan*, which varies according to the kind of stem it attaches to.

1 Introduction

This research was motivated by the desire to semi-automatically develop a lexicon for a wide-coverage, precision grammar of Indonesian. Although these linguistically-motivated grammars are invaluable resources to the NLP community, the biggest drawback is the time required for the manual creation and curation of the lexicon. Our work aims to expedite this process by automatically assigning syntactic information to stems that make up the verbal elements, on the basis of predicting syntactico-semantic clusters based on distributional similarity.

However, one minor point becomes one major obstacle in this task: Indonesian is a relatively under-resourced language in terms of NLP. Therefore, many of the techniques that have been deemed successful in the inferring of syntactic information or inducing syntactico-semantic classes are not available to us. Even studies that are considered lightweight minimally employ a part-of-speech (POS) tagger and chunker (Joanis et al., 2008), with many studies benefiting from the richness of the features that a syntactic parser provides (Schulte im Walde, 2006). In the case of Indonesian, there exist POS taggers (Pisceldo et al., 2009; Wicaksono and Purwarianti, 2010)[1] but no chunker or syntactic parser, and the reliance on such pre-processing tools is unrealistic.

We adhere to the notion that semantic similarity begets syntactic similarity as per Levin (1989), and so employ a distributional similarity method to learn our syntactic classes, based on a non-parametric Bayesian model. We experiment with learning both: (a) syntactic classes that capture the range of predicate argument structures (PASs) of a word and the syntactic alternations it participates in, but ignore large semantic differences in the component words; and (b) syntactico-semantic classes that capture PAS and alternation properties, but are also semantically coherent (a la Levin classes).

Here, we focus on the syntactic variation that arises with the affixing of the Indonesian suffix *-kan*. The specific morpho-syntactic behaviour of the *kan*-affixed verb is very much determined by the type of stem it attaches to, and its resulting behaviour varies from stem type to stem type (Kroeger, 2007; Vamarasi, 1999; Arka, 1993). The spectrum of variation induced by the affixing of *-kan* is not observed on all types of stems, and so being able to identify these superordinate types, representing the same morpho-syntactic variation, would assist greatly in accelerating lexicon development. It has been shown that Levin classes can be successfully induced employing unsupervised methods (Schulte im Walde, 2006; Kipper et al., 2006). We investigate the viability of automati-

[1] Although no POS tagger has been released for public use.

International Joint Conference on Natural Language Processing, pages 685–691,
Nagoya, Japan, 14-18 October 2013.

cally inducing coarser-grained types that represent morpho-syntactic variation, and we test whether the method we define is suited to such a task. Specifically, we focus on a case study detailed in Section 2 on the syntactic and semantic variation that arises with the affixing of the Indonesian suffix *-kan*. In Section 3 we outline our criteria in creating our gold standard data. Section 4 gives technical details of our methodology, and our interpretation of distributional similarity expressed in soft clusters derived using the hierarchical Dirichlet process (HDP). We present our results comparing our method employing HDP with a simpler benchmark system using hierarchical agglomerative clustering in Section 5, and also find that the method we employ in this study is better suited to discovering Levin-style classes rather than detecting morpho-syntactic variation, even though we had accommodated for syntactic structure in our model, by including functional words as structural indicators. We finally conclude with how we may extend this preliminary investigation.

Our contributions in this work are: (1) the demonstration that hierarchical Dirichlet processes are a highly effective way of modelling word similarity, outperforming simpler strategies; (2) the successful application of the syntax–semantics hypothesis of Levin to an under-resourced language based on distributional similarity analysis; (3) the finding that conflating semantic classes into superordinate types may be useful for annotating the lexicon, but when performing clustering tasks that employ distributional semantics, having a more semantically-oriented classification, such as Levin classes, are best suited for such methods, even when approximations are made to account for syntactic information; and (4) the demonstration that clustering based on semantic properties is a relatively strong predictor of deep syntactic lexical properties, and can be of great assistance in semi-automatically constructing a deep lexical resource for an under-resourced language

2 Background

Indonesian is an Austronesian language spoken by more than 165 million speakers in Indonesia (where it is the national language) and around the world (Gordon, 2005). Even with this status it still is an under-resourced language when it comes to NLP. For our case study, we aim to discover groups of like stems that, when used predicatively in the same morphological context, give rise to the same syntactic behaviour. That is, we aim to induce classes of stems that exhibit the same syntactico-semantic behaviour when they have the same morphological marking.

Predictions on syntactico-semantic properties of stems via morphological processes have also been explored for English (Grimshaw, 1990). Although Grimshaw's account of nominalisation restrictions with the English suffix *-ing* can be explained with a more general theory of argument structure, she also shows that the nominalisation of certain predicates in this way exclude certain lexical classes, namely psychological predicates as shown in Example (1) .

(1) a. *The (movie's) depress**ing** of the audience.
b. *The worry**ing** of the public.

The morpho-syntactic study presented in this paper is specific to Indonesian, but these lexical changes initiated by morphological processes can be a source of investigation into syntactico-semantic properties of lexemes for a variety of languages including English. For our case study we look into the Indonesian suffix *-kan*, which is generally described as a morpheme that triggers a lexical rule that increases valency. It can introduce a benefactive object, form a causative construction, or apply other semantic changes. Examples (2) and (3) show the benefactive, and causative uses, respectively:

(2) a. *Dia membeli buku itu untuk Mary.*
s/he AV+buy book this for M
"(S)he bought a book for Mary."
b. *Dia membelikan Mary buku itu.*
s/he AV+buy+KAN M book this
"(S)he bought Mary a book."

(3) a. *Orang-orang mengungsi.*
person-person AV+take-refuge
"The people took refuge."
b. *PBB mengungsikan orang-orang.*
U.N. AV+refuge+KAN person-person
"The U.N. evacuated the people."

In the second line of each of these glossed examples, AV stands for *actor voice*, which means that the verb is active. This is marked by the prefix *me-* plus a homorganic nasal, which can be realised

as *m*, *n(g|y)* or ∅. This verb behaves in a similar fashion to English verbs in an active sentence. We limit the examination of verbs in this study to those that exhibit the actor voice (AV) marking.

Linguists have tried to characterise stems according to their behaviour when affixed with *-kan* (Dardjowidjojo, 1971; Arka, 1993; Vamarasi, 1999). In particular, Vamarasi (1999) claims that *kan* is a good diagnostic for separating unaccusative from unergative stems, which predicts their morphosynatic behaviour. However the facts of *-kan* seem more intricate than this characterisation. Even though the causative and benefactive constructions uses of *kan* are the most commonly cited, its usage is much more varied and nuanced, as shown by Kroeger (2007), which is why we chose this morpho-syntactic construction as our case study.

Since the early '90s, the tools and resources employed in valency acquisition tasks have become increasingly sophisticated and lingusitically-rich. One of the earlier examples of this is by Brent (1993), who employs a system based on deterministic morphological cues to identify predefined syntactic patterns from the Brown Corpus. Manning (1993) employs a shallow parser or chunker in order to acquire subcategorisation frames from the New York Times. Schulte im Walde (2002) induces subcategorisation information for German with the use of a lexicalised probabilistic context free grammar (PCFG), and O'Donovan et al. (2005) employ the richly-annotated Penn Treebank in achieving this endeavour. In terms of resources, our work most closely resembles Brent (1993), in that we rely mainly on linguistic knowledge based on simple lexical features. However, the way linguistic knowledge is learned and applied is quite different, as we will see in Section 3

In terms of the methodology, the studies that we look to are those systems that are built to disambiguate and/or discover syntactico-semantic Levin-style classes, rather than systems that aim to induce valency or syntactic frame information from corpora. These can be built in a supervised fashion as in Lapata and Brew (2004) or tackled as a clustering task as in Schulte im Walde (2006) or Bonial et al. (2011). Lapata and Brew (2004) develop a semi-supervised system that generates, for a given verb and its syntactic frame, a probability distribution over the Levin verb classes. They then use this system to disambiguate tokens using collocation information. Our system, like Schulte im Walde (2006), uses an unsupervised clustering approach. In her approach, Schulte Im Walde employs hierarchical agglomerative clustering over parse features to discover word classes in German, and evaluates using manually-created gold-standard data.

3 Evaluation Data

This section describes how we arrive at the two evaluation sets we use in our experiments.

3.1 Forming Levin Classes

We use VerbNet 3.2[2] as our guide for forming Levin classes for Indonesian, and rely on their translation to determine membership for the class, for a particular sense of that verb.

We have 30 stems that we group into 16 Levin classes. Unlike the types we form in Section 3.2, which have unique membership, a lexical item can appear in multiple classes as appropriate. For example *baca* "read" has membership in both VerbNet classes **say-27.7** and **learn-14**. We show a subset of Indonesian Levin classes we develop based on VerbNet 2.3 in Table 1.

3.2 Forming Superordinate Levin Types

These superordinate types combine Levin classes to form groups of stems that behave in the same way syntactically, but may not all be synonyms of each other. In determining the coarse-grained superordinate types, we did not simply want to group stems according to intuition. Rather, we were after an explicit description of the syntax and semantics of grouped stems that all behave in the same way when affixed with *-kan*. Stems that are grouped together should exhibit the same semantic shifts. That is, if affixing *-kan* to a stem gives rise to a causative meaning, then its corresponding group member will also produce a causative meaning when *-kan* is applied to the stem. Also, if adding a *-kan* does not increase the valency for a stem in a particular group, then its corresponding group member will also exhibit the same syntatic behaviour.

In order to achieve this, we map out the different behaviour of verb stems when they occur in the morphological patterns (a) and (b): [3]

[2] `http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html`

[3] As mentioned earlier in Section 2, AV stands for *actor voice*, and can be likened to an English verb in an active sentence.

Indonesian Members	VerbNet Class
beri "give" *jaja* "hawk/sell", *pinjam* "lend"	give-13.1
kenang "think" *kenal* "know" *ingat* "remember"	consider-29.9
mati "die", *tewas* "perish"	disappearance-48.2
susup "duck down", *singkir* "get out of way"	avoid-52
timpa "hit" *hantam* "hit/blow" *tabrak* "hit"	hit-18.1
baca "read" *tulis* "write"	say-37.7
baca "read" *hafal* "memorize"	learn-14

Table 1: Subset of the mapping of Levin classes into Indonesian

(a) ME *N*+stem
AV+stem

(b) ME *N*+stem+KAN
AV+stem+KAN

We map out the variation of arguments for pattern (a) with only the AV prefix, i.e. ME *N*+stem, and then note the changes when the stem has both the actor AV and *-kan* affixes, i.e. pattern (b) ME *N*+stem+KAN. We also track the semantic changes relative to the stem for these two patterns and found that 25 verb stems found their way into 8 verb types.[4] This formed one of our evaluation sets in our experiments (see Mistica (2013) for further details on forming these superordinate types).

In the interests of space, we only present two out of the 8 manually-induced verb types in Table 2. Below each of the types, we show the syntactic and semantic changes that determine our verb types or subclasses.

4 Method

We define our features in terms of the context of occurrence of our target lexeme, and employ hierachical agglomerative clustering (HAC) over these features in two ways: (1) directly over the raw word frequencies; and (2) over extracted semantic features learned via the contexts of occurrence, which are represented as topic probabilities.

We use Indonesian Wikipedia[5] as our text collection, and remove mark-up with Wikiprep,[6] then tokenise with the English-trained models of OpenNLP.[7] The total word count of the text collection is approximately 26 million words. In the next section we summarise the features we use in our experiments, in addition to outlining our clustering method.

4.1 Feature Engineering

Our features determine how we collect unigrams from the text collection. We collect these unigram features from 735 lexemes that we were able to identify as possible *-kan* hosts. These 735 lexemes had stems that belonged to any of the open class categories in Indonesian (noun, adjective or verb).

In our preparation of the Wikipedia data, we include function words as a means to infer structural information. Because we do not use a parser to explicitly obtain syntactic features, this is how we approximate this kind of information.

We use three main feature types in our task: (1) **morph** ∈ 'k', 'mk', 'smk'; (2) **win** ∈ 1 to 5; and (3) **context** ∈ '+' (forward), '−' (backward).

Morphological features (morph): These are contextual features for different morphological forms of the target lexeme, where: *'s'* stands for *stem*, i.e. the unaffixed lexeme; *'m'* stands for the AV variant of the lexeme, based on pattern (a) from Section 3.2; and *'k'* stands for the KAN suffixed form of the AV variant of the lexeme, based on pattern (b) from Section 3.2. An example of the *'s'*, *'m'* and *'k'* variants of *beli* "buy" are *beli*, *membeli*, and *membelikan*, respectively. These morphological features determine whether the unigram features we collect for a lexeme are based on instances of ((s)m)k forms found in the text. We experiment with the context features based on these morphological variants in isolation and also in combination. For example, 'mk' would capture context features for the *membeli* and *membelikan* variants of the stem *beli* "buy".

Window Size (win): This stipulates the context window size, relative to individual occurrences of the target lexeme, and can take a value of 1–5.

[4] We had also manually grouped stems from other word classes: 48 noun stems were grouped into 13 subclasses; and 27 adjective stems were grouped into 5, giving us a total of 100 stems with the 25 verbs, but we only report on the verb experiments.

[5] http://dumps.wikimedia.org/idwiki/

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

[7] http://opennlp.apache.org/ Our experiments showed that OpenNLP's English models performed better than a rule-based Malay sentence tokeniser (Baldwin and Awab, 2006).

MORPHOLOGY	VERB FRAME	DECOMPOSITION
Example Type A: *acuh* "to heed", *terjemah* "translate", *mandi* "bathe"		
MEN+V_1	–	–
MEN+V_1+KAN	$<NP_a, NP_b>$	DO_{to}([NP_a], [V_1 TO([NP])])
Example Type B: *dengar* "hear", *kenang* "think of"		
MEN+V_3	$<NP_a, NP_b>$	$HAPPEN_{to}$([NP_b], [V_3 TO([NP_a])])
MEN+V_3+KAN	$<NP_a, NP_b>$	DO_{to}([NP_a], [V_3 TO([NP_b])])

Table 2: Manually generated verb Types ('–' = no attested word form in the text; '{...}' = optional)

Context Features (context): We look at backward ('−') or forward ('+') context unigrams.

4.2 Clustering Stems

We employ hierarchical agglomerative clustering (HAC) in two ways: (1) over the raw frequencies of words based on the feature representations defined in Section 4.1; and (2) over the output of the distributional semantic modelling (HDP) discussed in Section 4.3. The output of this step produces topic models. In other words, we perform HAC over raw unigram frequencies and induced topic models from these raw frequencies to ascertain the usefulness of the HDP step.

To compute the distance between a pair of patterns, we use Squared Euclidean, and for the linkage criterion for merging clusters we use weighted linkage clustering (WPGMA). We compare the output of HAC with the flat-structured gold-standard classes. In order to induce flat clusters from the hierarchical output of HAC, we apply a similarity threshold $t = 0.825$ to determine which instances should be grouped together.

4.3 Modelling Distributional Similarity

Distributional semantic models are commonly employed in the induction and disambiguation of word senses (McCarthy and Carroll, 2003; Lapata and Brew, 2004; Brody and Lapata, 2009; Lau et al., 2012), and to a lesser extent, in learning syntactic classes and diathesis alternation behaviour (Parisien and Stevenson, 2011; Bonial et al., 2011). We infer lexical similarity and soft word clusters using topic modelling, based on a hierarchical Dirichlet process (HDP: Teh et al. (2006)), a non-parametric extension of latent Dirichlet allocation (LDA: Blei et al. (2003)). LDA is a Bayesian generative topic model that learns *latent* topics for a collection of documents based on the *observable* words. Our definition of a document is a target lexeme and the observable words that surround the target lexeme (based on the window size in the parameter settings).

Formally, in LDA a topic is associated with a multinomial distribution of words, and each document (i.e. lexeme) in the collection is associated with a multinomial distribution of topics. HDP relaxes the constraint in LDA where the number of topics T is fixed, and learns T based on the training data using Dirichlet processes (DPs).

4.4 Evaluation

We develop two baseline systems to compare our results against: (1) majority class; and (2) random class assignment based on a uniform class distribution. The random scores reported are based on the median of 11 random assignments.

We use pairwise precision (pP), recall (pR), and F-score (pF_1) to evaluate our generated clusters, relative to the gold-standard word classes, as described by Schulte im Walde (2006).

5 Results

We perform two experiments. First, we apply the hierarchical Dirichlet process (HDP) to produce topic probabilities, over which we perform HAC. Second, we perform HAC over the raw unigram features (NoHDP), as our benchmark system, a method also employed by systems such as Schulte im Walde (2002) for German and Jurgens and Stevens (2010) for English word sense induction. In both cases, we base our experiments on the 735 lexemes identified as being able to be affixed with *-kan*, and the unigram features from Section 4.1. Note, however, that evaluation is based on the subset of the 735 lexemes which were manually classified into classes and types in Section 3.

We employ a *bagging* approach (sampling with

System	Maj.	Rand.	ON-ALL	ON-VERBS
LEVIN-HDP	.114	.065	**.174**	**.367**
LEVIN-NOHDP			.057	.111
TYPES-HDP	.271	.140	**.281**	.261
TYPES-NOHDP			.026	.152

Table 3: pF_1 score comparing benchmark system NOHDP with our HDP system for Levin Classes (LEVIN) and our coarser-grained TYPES

A	*main* "play", *nyanyi* "sing", *gesek* "scrape"
B	*kirim* "send", *hantar* "place"
C	*dapat* "get", *menang* "win", *terima* "receive"

Table 4: Induced groups with no known categorised words

replacement) to ascertain the best parameters to apply to our 735 lexemes in terms of the unigram features we define in Section 4.1.

Given the discovered parameters, we report our results in Table 3. The label ON-ALL for all HDP systems are sytems that have had topics induced from all 735 stems (made up of not only verbs, but also nouns and adjectives), while ON-VERBS only induces topics from a subset of the 735 lexemes whose stems are also verbs, even though we only evaluate on verbs in these experiments.

We observe in Table 3 that HDP consistently outperforms NO-HDP systems. Furthermore, the LEVIN-HDP system outperforms the Random ("Rand.") and the Majority Class ("Maj.") baselines, as well as the benchmark NOHDP system. The TYPES-HDP system, on the other hand, barely exceeds the Majority Class baseline with the ON-ALL experiment, and fails to do so with the ON-VERBS experiment.

6 Discussion

For our error analysis, we examine a sample of the resulting stem groups from the Levin Class experiments. Table 4 shows membership of all stems found in four separate clusters. The lexemes from these particular groups do not have membership into any of the gold standard Levin classes, unlike the groups formed in Table 5. In this table, the top half are groups that match our Levin classes, part of which is presented in Table 1, and the bottom half are groups that do not match Levin classes.

Group A from Table 4 has 3 verbs — *main* "play", *nyanyi* "sing", and *gesek* "scrape" — which may initially seem not to form a semantically coherent group, however they all are associated with producing music: *main* "play" is used to describe the playing of most musical instruments, and *gesek* "scrape/rub" is used for string instruments, such as violins or cellos. Group B has members that describe movement from one place to another, as does Group C.

D	*singkir* "get out of way", *susup* "duck down"
E	*baca* "read" *hafal* "memorise"
F	*terjemah* "translate" *tulis* "write", *muat* "insert/contain"
G	*paksa* "force" *pinjam* "lend" *hapus* "wipe off/vanish/blot out"

Table 5: Induced groups with known categorised words

Groups D and E in Table 5 faithfully replicate the Levin Classes **avoid-52**, adn **learn-14** from Table 1. However, Groups F and G seem to not form coherent semantic groups.

7 Conclusion

We have explored the question of whether distributional similarity models can be used to learn deep syntactic features for an under-resourced language, namely Indonesian. Our results demonstrate that hierarchical Dirichlet processes are a highly effective way of modelling word similarity, and outperform a simpler strategy of simply applying HAC over raw frequencies. We have also shown that learning classes geared toward the potential morpho-syntactic alternations of stems, while conflating the semantics of the stem are too coarse for this particular method. The experiments that used true Levin classes to evaluate against performed much better in comparison to the baselines, than did the experiments where we induced our manually constructed coarse-grained types. Although resources and tools are limited for Indonesian NLP, we would need to model syntactic structure more effectively to gain success in predicting lexical types rather than Levin classes.

References

I Wayan Arka. 1993. Morphological aspects of the -kan causative in indonesian. Master's thesis, The University of Sydney, Sydney, Australia, November.

Timothy Baldwin and Suád Awab. 2006. Open source corpus analysis tools for Malay. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 2212–5, Genoa, Italy.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Claire Bonial, Susan Windisch Brown, Jena D. Hwang, Christopher Parisien, Martha Palmer, and Suzanne Stevenson. 2011. Incorporating coercive constructions into a verb lexicon. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 72–80, Portland, USA.

Michael R. Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 103–111, Athens, Greece.

Soenjono Dardjowidjojo. 1971. The meN-, meN-kan, and meN-i verbs in Indonesian. *Philippine Journal of Linguistics*, 2:71–84.

Raymond Gordon. 2005. *Ethnologue: Languages of the World*. SIL International, Dallas, USA.

Jane Grimshaw. 1990. *Argument Structure*. The MIT Press, Cambridge, USA.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

David Jurgens and Keith Stevens. 2010. HERMIT: Flexible clustering for the SemEval-2 WSI task. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 359–362, Uppsala, Sweden.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC 2006*, pages 1027–1032, Genoa, Italy.

Paul R. Kroeger. 2007. Morphosyntactic vs. morphosemantic functions of indonesian '-kan. In Joan Bresnan, Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Christopher D. Manning, editors, *Architectures, rules, and preferences: variations on themes*, CSLI Lecture Notes, pages 229–251. CSLI Publications.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, pages 591–601, Avignon, France.

Beth Levin. 1989. *English Verb Classes and Alternations: A preliminary investigation*. The University of Chicago Press, Chicago, USA.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceeding of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, USA.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Meladel Mistica. 2013. *An Investigation into Deviant Morphology: Issues in the Implementation of a Deep Grammar for Indonesian*. Ph.D. thesis, The Australian National University, Canberra, Australia.

Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. 2005. Large-scale induction and evaluation of lexical resources from the Penn-II and Penn-III treebanks. *Computational Linguistics*, 31(3):229–365.

Chris Parisien and Suzanne Stevenson. 2011. Generalizing between form and meaning using learned verb classes. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Boston, USA.

Femphy Pisceldo, Ruli Manurung, and Mirna Adriani. 2009. Probabilistic part-of-speech tagging for Bahasa Indonesia. In *Proceedings of the Third International MALINDO Workshop*, Singapore.

Sabine Schulte im Walde. 2002. A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Marit Kana Vamarasi. 1999. *Grammatical relations in Bahasa Indonesia*, volume 93 of *Series D*. Pacific Linguistics, Canberra, Australia.

Alfan Farizki Wicaksono and Ayu Purwarianti. 2010. HMM based part-of-speech tagger for Bahasa Indonesia. In *Proceedings of the 4th International MALINDO Workshop (MALINDO2010)*, Depok, Indonesia.