Building an HPSG-based Indonesian Resource Grammar (INDRA)

David Moeljadi **Francis Bond** **Sanghoun Song**
Division of Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
{D001,fcbond,sanghoun}@ntu.edu.sg

Abstract

This paper presents the creation and the initial stage development of a broad-coverage Indonesian Resource Grammar (INDRA) within the framework of Head Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005). At the present stage, INDRA focuses on verbal constructions and subcategorization since they are fundamental for argument and event structure. Verbs in INDRA were semi-automatically acquired from the English Resource Grammar (ERG) (Flickinger, 2000) via Wordnet Bahasa (Nurril Hirfana Mohamed Noor et al., 2011; Bond et al., 2014). In the future, INDRA will be used in the development process of machine translation. A preliminary evaluation of INDRA on the MRS test-suite shows promising coverage.

1 Introduction to Indonesian

Indonesian (ISO 639-3: ind) is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Malayic branch with Standard Malay in Malaysia and other Malay varieties (Lewis, 2009). It is spoken mainly in the Republic of Indonesia as the sole official and national language and as the common language for hundreds of ethnic groups living there (Alwi et al., 2014, pp. 1-2). In Indonesia it is spoken by around 22.8 million people as their first language and by more than 140 million people as their second language. The lexical similarity is over 80% with Standard Malay (Lewis, 2009).

Morphologically, Indonesian is a mildly agglutinative language, compared to Finnish or Turkish where the morpheme-per-word ratio is higher (Larasati et al., 2011). It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, and reduplication. Most of the affixes are derivational. Two important inflectional affixes are the prefix *meN-* which marks active voice and *di-* which denotes passive voice (Sneddon et al., 2010, pp. 29, 72).

Indonesian has a strong tendency to be head-initial (Sneddon et al., 2010, pp. 26-28). In a noun phrase with an adjective, a demonstrative or a relative clause, the head noun precedes the adjective, the demonstrative or the relative clause. There is no agreement in Indonesian. In general, grammatical relations are only distinguished in terms of word order. As is often the case with Austronesian languages of Indonesia, Indonesian has a basic word order of SVO with a nominative-accusative alignment pattern. Argument alternations are triggered by passive and applicative constructions.

2 Background

This section introduces the background theory, as well as an overview of the Deep Linguistic Processing with HPSG Initiative (DELPH-IN) and the tools to build and develop INDRA.

2.1 Frameworks

INDRA uses the theoretical framework of HPSG (Pollard and Sag, 1994). HPSG is monostratal, handling orthography, syntax, semantics and pragmatics in a single structure (sign), modeled through typed feature structures. HPSG is unification- and constraint-based. The words and phrases are combined according to constraints of the lexical entries based on the type hierarchy. INDRA uses MRS (Copestake et al., 2005) as its semantic framework because it is adaptable for HPSG typed-feature structure and suitable for parsing and generation. The semantic structures in MRS are underspecified for scope and thus suitable for representing ambiguous scoping.

Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop, 53rd Annual Meeting of the ACL and 7th IJCNLP, pages 9–16,
Beijing, China, July 26-31, 2015.

There is no previous work done on Indonesian HPSG but much has been done using Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982), e.g. Arka and Manning (2008) on active and passive voice and Arka (2000) on control constructions. In addition, Arka (2012) and Mistica (2013) have worked on the computational grammar "IndoGram" which is a part of the ParGram (Sulger et al., 2013).[1] However, it is not open-source or very broad in its coverage. Further, it does not produce MRS, so cannot be easily incorporated into our machine translation system. Thus, there is a need to build and develop a broad-coverage open-source HPSG of Indonesian.

2.2 DELPH-IN

The DELPH-IN consortium (Deep Linguistic Processing with HPSG Initiative, `http://www.delph-in.net`) is a research collaboration between linguists and computer scientists which builds and develops open source grammar, tools for grammar development and applications using HPSG and MRS. More than fifteen grammars have been created and developed within DELPH-IN, e.g. English Resource Grammar (ERG) (Copestake and Flickinger, 2000) and Japanese grammar Jacy (Siegel and Bender, 2002). DELPH-IN grammars define typed feature structures using Type Description Language (TDL) (Copestake, 2002).

We make extensive use of several open-source tools for grammar development provided by DELPH-IN: Linguistic Knowledge Builder (LKB) (Copestake, 2002), a grammar and lexicon development environment for typed feature structure grammars; The LinGO Grammar Matrix (Bender et al., 2010), a web-based questionnaire for writing new DELPH-IN grammars, providing a wide range of phenomena and basic files to make the grammars compatible with DELPH-IN parsers and generators; Answer Constraint Engine (ACE) (`http://sweaglesw.org/linguistics/ace/`), an efficient processor for DELPH-IN grammars; ITSDB or [incr tsdb()] (Oepen and Flickinger, 1998), a tool for testing, profiling the performance of the grammar and treebanking; Full Forest Treebanker (FFTB) (`http://moin.delph-in.net/FftbTop`), a treebanking tool for DELPH-IN grammars, allowing the selection of an arbitrary tree from the "full forest" without enumerating all analyses in the parsing stage; and LOGON (Oepen et al., 2007), a collection of software, grammars, and other linguistic resources for transfer-based machine translation.

[1] `http://iness.uib.no/iness/xle-web`

3 INDRA

This section describes some preliminary work as well as the methodology.

3.1 Methodology

The methodology used in INDRA follows Bender et al. (2008). We model our analysis in HPSG and implement it by editing some TDL files after analyzing a phenomenon based on reference grammars and other linguistic literatures. Afterwards, we compile the grammar and test it by parsing sample sentences or test-suites. The grammar is debugged and developed further if some gaps or problems are found according to the parse results. Afterwards, the sample sentences in test-suites will be parsed again and treebanked. This process goes repetitively. If problems are not found or the debugging process has finished with a good result, the grammar will be updated in GitHub (`https://github.com/davidmoeljadi/INDRA`).

3.2 Grammar Development

INDRA was created firstly by filling in the required sections of the online page of LinGO Grammar Matrix questionnaire which covers basic grammar phenomena such as word order, tense-aspect-mode, coordination, morphology, subcategorization of nouns and verbs (`http://www.delph-in.net/matrix/customize/matrix.cgi`). INDRA subcategorizes nouns into three groups: common noun, pronoun and proper name. Common nouns are subcategorized into inanimate, non-human and human based on three main classifiers in Indonesian: the classifier *buah* (lit. fruit) for inanimate nouns, *ekor* (lit. tail) for non-human animate nouns and *orang* (lit. person) for human nouns (Sneddon et al., 2010, p. 139; Alwi et al., 2014, p. 288).

Verbs are subcategorized into three groups: intransitive which has one argument, transitive which has two arguments and optional transitive which has one obligatory subject argument and one optional object argument as in *Adi makan (nasi)* "Adi eats (rice)". The verb subcategorization here follows Alwi et al. (2014, pp. 95-98). Besides the number of arguments, the possibil-

ity of passivization with morphological inflection plays an important role in distinguishing intransitives from transitives in Indonesian. Examples [1] and [2a] show intransitive and transitive sentences respectively.

(1) *Adi tidur.*
Adi sleep

"Adi sleeps."

(2) a. *Adi mengejar Budi.*
Adi ACT-chase Budi

"Adi chases Budi."

b. *Budi dikejar Adi.*
Budi PASS-chase Adi

"Budi is chased by Adi."

c. *Budi saya kejar.*
Budi 1SG chase

"Budi is chased by me."

In Example (2a), the verb *mengejar* is formed from an active prefix *meN-* and the base *kejar* (the initial sound *k* undergoes nasalization; see Section 4.2). The active prefix *meN-* is changed to a passive prefix *di-* in passive type one (Sneddon et al., 2010, pp. 256-257) in Example (2b) and without affix in passive type two (Sneddon et al., 2010, pp. 257-258) in Example (2c). Sneddon et al. (2010, pp. 256-257) states that in passive type one, the actor is third person or a noun, while in passive two, the agent is a pronoun or pronoun substitute and it comes before the unprefixed verb.

The more detailed verb subcategorization into other groups such as ditransitive will be mentioned in the next subsection. The lexical items for each noun and verb subcategory were added and the affixes to support the active-passive voice were included. However, the Matrix does not handle morphology as in the nasalization process of *meN-* and thus has to be manually added (see Section 4.2).

3.3 Lexical Acquisition

The lexicon is important in the robustness of the grammar. Since inputting words or lexical entries manually into the grammar is labor intensive and time consuming, doing lexical acquisition semi-automatically is vital. In order to do this, we need good lexical resources. We attempted to extract Indonesian verbs from Wordnet Bahasa (Nurril Hirfana Mohamed Noor et al., 2011; Bond et al., 2014) and group them based on syntactic types in the ERG, such as intransitive, transitive, and ditransitive, using Python 3.4 and Natural Language Toolkit (NLTK) (Bird et al., 2009). The grouping of verbs (verb frames) in Wordnet (Fellbaum, 1998) is employed to be the bridge between the English and Indonesian grammar.

Each verb synset in Wordnet (also Wordnet Bahasa) contains a list of sentence frames specified by the lexicographer illustrating the types of simple sentences in which the verbs in the synset can be used (Fellbaum, 1998). There are 35 verbal sentence frames in Wordnet, some of them are shown as follows with their frame numbers:

(3) 1 Something ----s
8 Somebody ----s something
21 Somebody ----s something PP

Frame 1 is a typical intransitive verbal sentence frame, as in *the book fell*; frame 8 is a typical (mono)transitive verbal sentence frame, as in *he chases his friend*; and frame 21 is a typical ditransitive verbal sentence frame, as in *she put a book on a table*. A verb may have more than one synset and each synset may have more than one verb frame, e.g. the verb *eat* has six synsets with each synset having different verb frames. Three of the six synsets, together with their definition and verb frames, are presented in Table 1. These verb frames can be employed as a bridge between the verb types (also verb lexical items) in ERG and those in INDRA.

Synset	Definition	Verb frame
01168468-v	Take in solid food	8 Somebody ----s something
01166351-v	Eat a meal, take a meal	2 Somebody ----s
01157517-v	Use up (resources or materials)	11 Something ----s something 8 Somebody ----s something

Table 1: Three of six synsets of the verb "eat" and their verb frames in Wordnet

Out of 354 verb types in ERG, the top eleven most frequently used types in the corpus were chosen, excluding the specific English verb types such as *be*-type verbs (e.g. *is*, *be* and *was*), *have*-type verbs, verbs with prepositions (e.g. *depend on*, *refer to* and *look after*) and modals (e.g. *would*, *may* and *need*). The chosen eleven verb types are given in Table 2. The third, fifth and eighth type (*v_-_unacc_le*, *v_-_le* and *v_pp_unacc_le* all written in

bold in Table 2) are regarded as the same type, i.e. intransitive verb type, in INDRA.

Verb type	Freq		Examples of verb
	Corp	Lex	
v_pp*_dir_le	7079	204	go, come, hike
v_vp_seq_le	3921	105	want, like, try
_-_unacc_le	3144	334	close, start, end
v_np_noarg3_le	2723	5	make, take, give
v_-_le	2666	486	arrive, occur, stand
v_np-pp_e_le	2439	334	compare, know, relate
v_pp*-cp_le	2360	154	think, add, note
v_pp_unacc_le	2307	44	rise, fall, grow
v_np-pp_prop_le	1861	135	base, put, locate
v_cp_prop_le	1600	80	believe, know, find
v_np_ntr_le	1558	10	get, want, total

Table 2: The ten most frequently used ERG verb types in the corpus

The first type contains verbs expressing movement or direction with optional PP complements, as in *B crept into the room*. The verbs in the second type are subject control verbs, as in *B intended to win*. The third type consists of unaccusative verbs without complements as in *The plate gleamed*. The fourth type contains verbs having two arguments (monotransitive) although they have a potential to be ditransitive as in *B took the book*. The fifth type contains intransitive (unergative) verbs as in *B arose*. The verbs in the sixth type have obligatory NP and PP complements as in *B compared C with D*. The verbs in the seventh type are verbs with optional PP complements and obligatory subordinate clauses as in *B said to C that D won*. Unaccusative verbs with optional PP complements as in *The seed grew into a tree* belong to the eighth type. Ditransitive verbs with obligatory NPs and PPs with state result as in *B put C on D* belong to the ninth type. The tenth type consists of verbs with optional complementizers as in *B hoped (that) C won* and the eleventh type consists of verbs with obligatory NP complements which cannot be passivized as in *B remains C*.

Based on the syntactic information of each verb type mentioned above, the corresponding verb frames in Wordnet were manually chosen. For example, the first type contains intransitive verbs with optional PP; thus, the verb frames should be Sb ----s and Sb ----s PP. The intransitive verbs without complements should correspond to the verb frames Sth ----s or Sb ----s, regardless of whether the subject is a thing or a person. Table 3 shows the eleven verb types in ERG and their corresponding Wordnet verb frames.

First, we checked for each verb in each verb type in Table 2 whether it is in Wordnet or not. If it could be found in Wordnet, the next step was to check whether the verb includes the verb frames mentioned in Table 3 or not. This step had to be done in order to find out the right synset since a verb can have many synsets but different verb frames as shown in Table 1. After the right synset was found, the corresponding Indonesian lemmas or translations were checked. One synset may have more than one Indonesian lemma or may not have Indonesian lemmas at all.

Verb type	Verb frame
v_pp*_dir_le	2 Sb ----s & 22 Sb ----s PP
v_vp_seq_le	28 Sb ----s to INFINITIVE
v_-_unacc_le v_-_le v_pp_unacc_le	1 Sth ----s || 2 Sb ----s
v_np_noarg3_le	8 Sb ----s sth || 11 Sth ----s sth
v_np-pp_e_le	15 Sb ----s sth to sb || 17 Sb ----s sb with sth || 20 Sb ----s sb PP || 21 Sb ----s sth PP || 31 Sb ----s sth with sth
v_pp*-cp_le	26 Sb ----s that CLAUSE
v_np-pp_prop_le	20 Sb ----s sb PP || 21 Sb ----s sth PP
v_cp_prop_le	26 Sb ----s that CLAUSE
v_np_ntr_le	8 Sb ----s sth || 11 Sth ----s sth

Table 3: The eleven most frequently used ERG verb types in the corpus and their corresponding Wordnet verb frames (sb = somebody, sth = something, & = AND, || = OR

The next important step is to check one by one the Indonesian lemmas belonging to the same synset and verb frames whether each can be grouped in the same verb type or not. This manual step has to be done because grouping verbs in a particular language into types is a language-specific work. Arka (2000) states that languages vary with respect to their lexical stock of "synonymous" verbs that may have different argument structures, e.g. the verb *know* can be both intransitive and transitive in Indonesian *tahu* and *ketahui* respectively, transitive only with an obligatory NP in Balinese[2] *tawang*, and transitive with optional NP in English *know*. Lastly, after the Indonesian verbs were extracted and grouped into their cor-

[2]Balinese (ISO 639-3: ban) is a Western Malayo-Polynesian language of the Austronesian language family. It belongs to the Malayo-Sumbawan branch. It is mainly spoken in the island of Bali in the Republic of Indonesia as a regional language (Lewis, 2009).

responding verb types, a new lexicon file for INDRA was made, in which the verbs are alphabetically sorted. The result is, in total, 939 Indonesian verbs were extracted and grouped into nine verb types as presented in Table 4. One verb may belong to more than one verb type.

This lexical acquisition is useful to extract lexical items (semi-)automatically through linguistic resources such as Wordnet Bahasa. The generated lexicon can be used to improve the grammar's coverage. We plan to further extract more verbs as well as other parts-of-speech such as nouns, adjectives and adverbs.

Verb type	Number of verb
v_pp*_dir_le	76
v_vp_seq_le	49
v_-_unacc_le	594
v_np_noarg3_le	5
v_np-pp_e_le	41
v_pp*-cp_le	23
v_np-pp_prop_le	85
v_cp_prop_le	53
v_np_ntr_le	13
Total	939

Table 4: New verb types and the corresponding number of verbs in INDRA

4 Analyzing Indonesian Phenomena

After creating INDRA via the Grammar Matrix customization system, some additions and changes were done to the TDL files. Pronouns, proper names and adjectives which were formerly added via the Grammar Matrix customization system, were subsequently constrained so that they cannot parse phrases such as **saya kaya* "rich I". In addition, besides the new verb types which had been acquired from ERG, more verb rules such as control and raising were manually added. In total, there are 49 lexical types/categories in the lexicon.

The next subsections discuss some phenomena, e.g. decomposing words and morphology, analyzed and implemented in INDRA.

4.1 Decomposed Words

Following Seah and Bond (2014) who state that pronouns can be analyzed componentially, some words such as *sini* "here" can be mapped to multiple predicates, e.g. *sini* "here" can be thought of as *tempat ini* "this place". The way to model this is by defining type hierarchies for the head (e.g. *tempat* "place") and the demonstrative (e.g. *ini* "this"). Figure 1 and 2 show the type hierarchy for heads and demonstratives respectively.

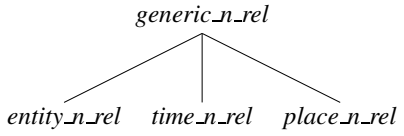

Figure 1: Type hierarchy for heads

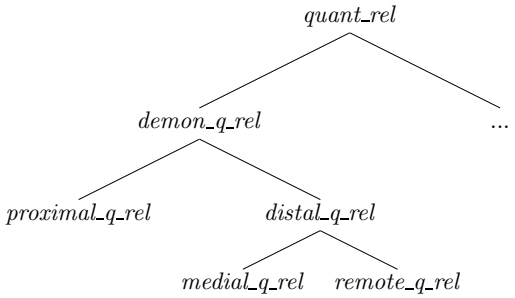

Figure 2: Type hierarchy for demonstratives

Indonesian has two demonstratives: *ini* "this" and *itu* "that" but three locative pronouns: *sini* "here (near speaker)", *situ* "there (not far off)" and *sana* "there (far off)" (Sneddon et al., 2010, pp. 133, 195). These can be modeled using the type hierarchy for demonstratives. The demonstrative *itu* "that" has the predicate *distal_q_rel*; the locative pronouns *situ* and *sana* has the predicate *medial_q_rel* and *remote_q_rel* respectively, which are the daughters of the predicate *distal_q_rel*. Figure 3 shows the implementation in TDL.

Figure 4 shows the MRS representation of the decomposed word *situ* "there" which is preceded by a preposition *di* "at". The ARG0 in the semantic head daughter *di* "at" is equated with the INDEX which has the value *e*2. The value of the ARG2 (*x*4) is coindexed with the ARG0 of *place_n_rel* and *medial_q_rel*. The *medial_q_rel* introduces RSTR which is related to the top handle of the quantifier's restriction (*h*7) and linked to the LBL of *place_n_rel* ($h7 =_q h5$).

Decomposing words is important to get more refined semantics. We will expand this to other heads and demonstratives such as *kini* "at present" which can be decomposed into *time_n_rel* and

```
situ := n+det-lex &
 [STEM < "situ" >,
  SYNSEM.LKEYS [ KEYREL.PRED "place_n_rel",
                 ALTKEYREL.PRED "medial_q_rel"]].
```

Figure 3: Decomposed predicates of *situ* "there"

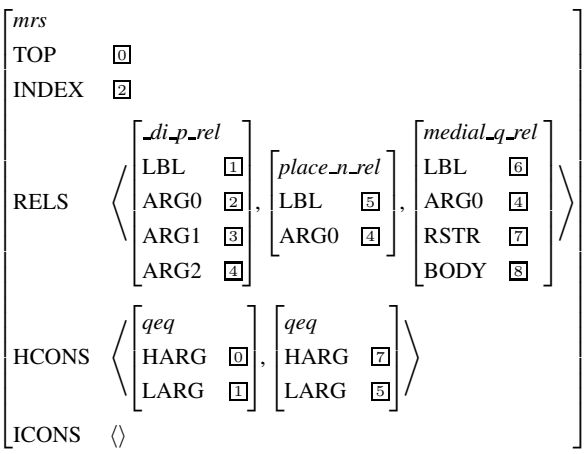

Figure 4: MRS representation of *di situ* (lit. "at there")

proximal_q_rel.

4.2 Morphology

As mentioned in Section 3.2, a number of nasalization (sound changes) or morphology process occur when *meN-* combines with bases. Table 5 shows us that a number of sound changes occur when *meN-* combines with a base. A base loses its initial consonant if the consonant is one of the following voiceless consonants: *p*, *t*, *s* and *k*. It retains its initial consonant otherwise. The sound changes of every possible combination of consonant clusters in Alwi et al. (2014, pp. 67-68) was manually examined using an online Indonsian dictionary (KBBI Daring) (Alwi et al., 2008). In addition, when the base consists of only one syllable, *meN-* becomes *menge-* with no sound changes in the base. Every possible combination of one syllable word with *meN-* which forms a transitive verb in KBBI Daring was listed up. There were 44 one syllable words in total. All 24 possible consonant clusters and 44 one syllable words were added to the inflectional rules in INDRA.

Moreover, besides the consonant clusters and one syllable words, a manual extension was also done for the exceptions. The sound *p* is usually lost when combined with *meN-* but it is retained when it is a derivational prefix *per-* as in *pertinggi* (from *per-* and *tinggi* "high"). At the present stage, all transitive bases with *per-* are being listed up and will be added in INDRA. There are also bases such as *punyai* "have" and *syairkan* "compose a poem" (Sneddon et al., 2010, pp. 16-17) which do not undergo the common sound changes.

At the present stage, this morphology process

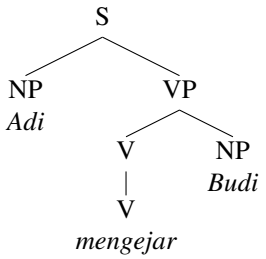

Figure 5: Parse tree of *Adi mengejar Budi* "Adi chases Budi"

applies to all transitive verbs in INDRA with a constraint stating that objects are obligatory. Other verb types such as ditransitives, control and raising which can be passivized will be further included in the inflectional rules. At present, INDRA can parse the example (2a) as shown in Figure 5. The MRS representation is exactly the same as the MRS representation for transitive sentences (see Figure 6). The value of `ARG0` of the semantic head daughter *_kejar_v_rel* is an event (*e*2) which is equated with the `INDEX`. The value of `ARG0` of *named_rel* "*adi*" (*x*3) and *named_rel* "*budi*" (*x*9) refer to the value of the `ARG1` and `ARG2` feature of the semantic head daughter respectively.

We intend to cover all the exceptions in the inflectional rules, particularly dealing with words having *per-* and to expand the rules to other verb types such as ditransitives. Passive type one and type two rules also need to be analyzed and implemented. As Sneddon et al. (2010, pp. 256, 263-264) pointed out, passive constructions in Indonesian are far more frequent than in English; an Indonesian passive is often naturally translated into English by an active construction. Thus, dealing with passive constructions will increase the grammar coverage. We anticipate that translating Indonesian passive constructions into English will be a challenge for machine translation.

5 Associated Resources

In order to make INDRA more robust, the following resources have been set up: Indonesian POS Tagger (Rashel et al., 2014) with ACE's YY-mode for unknown word handling (`http://moin.delph-in.net/ZhongYYMode`) which can parse sentences with unknown words and transfer grammar for machine translation. At present, INDRA can translate some simple sentences such as the ones in example (1) and (2a) using the `inen` (Indonesian-English) transfer grammar.

Allomorph of *meN-*	Initial orthography of the base		Example
mem-	p	(L)	*me**mp**akai* “use”
	pl, pr, ps, pt, b, bl, br, f, fl, fr, v	(R)	*me**mb**eli* “buy”
men-	t	(L)	*me**nt**anam* “plant”
	tr, ts, d, dr, c, j, sl, sr, sy, sw, sp, st, sk, sm, sn, z	(R)	*me**nc**ari* “seek”
meny-	s	(L)	*me**nys**ewa* “rent”
meng-	k	(L)	*me**ngk**irim* “send”
	kh, kl, kr, g, gl, gr, h, q, a, i, u, e, o	(R)	*me**ngg**anti* “replace”
me-	m, n, ny, ng, l, r, w, y	(R)	*me**l**empar* “throw”
menge-	(base with one syllable)		*me**ngecek*** ”check”

Table 5: Morphology process of *meN-* (L = lost, R = retained; Sneddon et al., 2010: 13-18)

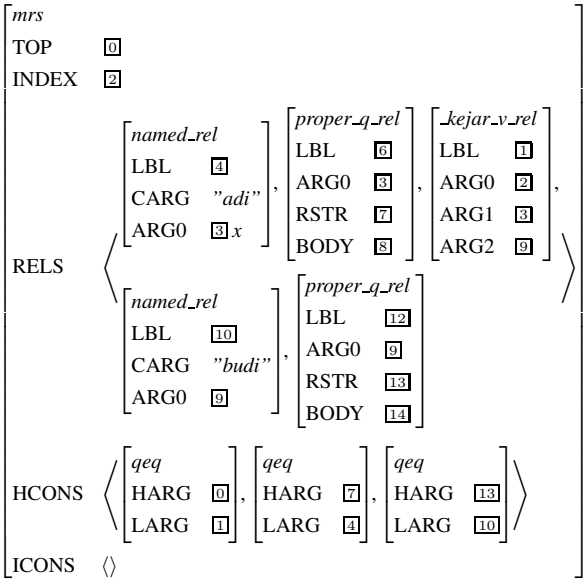

Figure 6: MRS representation of *Adi mengejar Budi* “Adi chases Budi”

6 Evaluation

A test-suite designed to show various semantic phenomena for Indonesian (MRS test-suite) was created based on the original set of 107 sentences in English. The [incr tsdb()] tool (Oepen and Flickinger, 1998) is employed for grammar testing and profiling. Out of 172 sentences, INDRA can parse 55 of them (overall coverage 32%). We got this 32% coverage after the lexical acquisition described in Section 3.3. Table 6 shows the coverage before and after lexical acquisition.

As of 18 June 2015, INDRA contains 1,235 lexical items, 939 of which are verbs extracted from ERG via Wordnet Bahasa; 6 lexical rules; 20 grammar rules; 135 features and 1,596 types. In addition to the phenomena in the Grammar Matrix customization system, INDRA also covers proper names, definiteness, possessive enclitics, adverbs, control and raising, decomposed words and morphology. Phenomena which are planned to be covered in the next two years are relative clauses, numbers, quantifiers, classifiers, copula constructions, passives, topic-comment constructions, particles, interrogatives and imperatives. We estimate that 15% of the MRS test-suite would be covered once passives and relative clauses were added.

	results / items	coverage
before	52 / 172	30.2%
after	55 / 172	32.0%

Table 6: Comparison of coverage in MRS test-suite before and after lexical acquisition

7 Summary and Future Work

The lexical acquisition has proved that by acquiring more lexical items, the grammar’s coverage can be improved. We plan to do more lexical acquisition for verbs, nouns, adjectives and adverbs in the future. At the same time, lexical types, rules and constraints for new lexical items will be added. Our plan in the next two years is to cover at least 60% of Indonesian text in the Nanyang Technological University — Multilingual Corpus (NTU-MC) (Tan and Bond, 2012). The latest version of INDRA is regularly backed up in GitHub.

Acknowledgments

Thanks to Michael Wayne Goodman and Dan Flickinger for teaching us how to use GitHub and FFTB. Thanks to Fam Rashel for helping us with POS Tagger and to Lian Tze Lim for helping us improve Wordnet Bahasa. This research was supported in part by the MOE Tier 2 grant *That’s what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13).

References

Hasan Alwi, Dendy Sugono, and Sri Sukesi Adiwimarta. 2008. *Kamus Besar Bahasa Indonesia Dalam Jaringan (KBBI Daring)*. 3 edition.

Hasan Alwi, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 2014. *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, 3 edition.

I Wayan Arka and C. D. Manning. 2008. Voice and grammatical relations in Indonesian: a new perspective. In *Voice and Grammatical Relations in Austronesian Languages*, pages 45–69. CSLI Publications, Stanford.

I Wayan Arka. 2000. Control and argument structure: explaining control into subject in Indonesian. In *Fourth International Symposium on Malay/Indonesian Linguistics*, Jakarta.

I Wayan Arka. 2012. Developing a deep grammar of indonesian within the pargram framework: Theoretical and implementational challenges. In *26th Pacific Asia Conference on Language, Information and Computation*, pages 19–38.

Emily M Bender, Dan Flickinger, and Stephan Oepen. 2008. Grammar engineering for linguistic hypothesis testing. In *Proceedings of the Texas Linguistics Society X conference: Computational linguistics for less-studied languages*, pages 16–36.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. In *Research on Language and Computation*, pages 23–72. Springer, Netherlands.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined wordnet bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC-2000)*, Athens.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(4):281–332.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge.

Dan Flickinger. 2000. On Building a More Efficient Grammar by Exploiting Types. 6(1):15–28.

Ronald Kaplan and Joan Bresnan. 1982. Lexical Functional Grammar: A formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*, pages 173–281. the MIT Press, Cambridge.

Septina Dian Larasati, Vladislav Kuboň, and Daniel Zeman. 2011. Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Springer CCIS proceedings of the Workshop on Systems and Frameworks for Computational Morphology*, pages 119–129, August.

M. Paul Lewis. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16 edition.

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

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267, Singapore.

Stephan Oepen and Daniel Flickinger. 1998. Towards systematic grammar profiling: Test suite technology ten years after. 12(4):411–436.

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation: On linguistics and probabilities in MT. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 144–153, Skövde, Sweden.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Fam Rashel, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung. 2014. Building an Indonesian Rule-Based Part-of-Speech Tagger. Kuching.

Yu Jie Seah and Francis Bond. 2014. Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese.

Melanie Siegel and Emily M. Bender. 2002. Efficient Deep Processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, Taipei.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*. Allen & Unwin, New South Wales, 2 edition.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, et al. 2013. Pargrambank: The pargram parallel treebank. In *ACL (1)*, pages 550–560.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). 22(4):161–174.