<table>
<thead>
<tr>
<th>Title</th>
<th>Building an HPSG-based Indonesian resource grammar (INDRA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Moeljadi, David; Bond, Francis; Song, Sanghoun</td>
</tr>
<tr>
<td>Date</td>
<td>2015</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/40509">http://hdl.handle.net/10220/40509</a></td>
</tr>
<tr>
<td>Rights</td>
<td>© 2015 Association for Computational Linguistics. This paper was published in Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop, 53rd Annual Meeting of the ACL and 7th IJCNLP and is made available as an electronic reprint (preprint) with permission of Association for Computational Linguistics. The published version is available at: [<a href="https://www.semanticscholar.org/paper/Building-an-HPSG-based-Indonesian-Resource-Grammar-Moeljadi-Bond/253af4a99614dd1ae93cf511b2db79e58bb6430d">https://www.semanticscholar.org/paper/Building-an-HPSG-based-Indonesian-Resource-Grammar-Moeljadi-Bond/253af4a99614dd1ae93cf511b2db79e58bb6430d</a>]. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper is prohibited and is subject to penalties under law.</td>
</tr>
</tbody>
</table>
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 monostatal, 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.
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). 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.

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-

http://inest.uib.no/inest/xle-web
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.

<table>
<thead>
<tr>
<th>Synset</th>
<th>Definition</th>
<th>Verb frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>01168468-v</td>
<td>Take in solid food</td>
<td>8 Somebody ----s something</td>
</tr>
<tr>
<td>01166351-v</td>
<td>Eat a meal, take a meal</td>
<td>2 Somebody ----s something</td>
</tr>
<tr>
<td>01157517-v</td>
<td>Use up (resources or materials)</td>
<td>11 Something ----s something</td>
</tr>
<tr>
<td>01157517-v</td>
<td>Use up (resources or materials)</td>
<td>8 Somebody ----s something</td>
</tr>
</tbody>
</table>

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_fe, v_fe and v_pp_unacc_fe all written in
bold in Table 2) are regarded as the same type, i.e. intransitive verb type, in INDRA.

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Freq Corp</th>
<th>Examples of verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>v_pp^*_dir_le</td>
<td>7079 204</td>
<td>go, come, hike</td>
</tr>
<tr>
<td>v_vp_seq_le</td>
<td>3921 105</td>
<td>want, like, try</td>
</tr>
<tr>
<td>v uninacc_le</td>
<td>3144 334</td>
<td>close, start, end</td>
</tr>
<tr>
<td>v_pp_noarg3_le</td>
<td>2723 5</td>
<td>make, take, give</td>
</tr>
<tr>
<td>v_le</td>
<td>2666 486</td>
<td>arrive, occur, stand</td>
</tr>
<tr>
<td>v_pp_pp_e_le</td>
<td>2439 334</td>
<td>compare, know, relate</td>
</tr>
<tr>
<td>v_pp^*cp_le</td>
<td>2360 154</td>
<td>think, add, note</td>
</tr>
<tr>
<td>v_pp_unacc_le</td>
<td>2307 44</td>
<td>rise, fall, grow</td>
</tr>
<tr>
<td>v_pp_prop_le</td>
<td>1861 135</td>
<td>base, put, locate</td>
</tr>
<tr>
<td>v_cp_prop_le</td>
<td>1600 80</td>
<td>believe, know, find</td>
</tr>
<tr>
<td>v_npptr_le</td>
<td>1558 10</td>
<td>get, want, total</td>
</tr>
</tbody>
</table>

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.

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 Balinese2 tawang, and transitive with optional NP in English know. Lastly, after the Indonesian verbs were extracted and grouped into their cor-

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

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Verb frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>v_pp^*_dir_le</td>
<td>2 Sb ----- s &amp; 22 Sb ----- s PP</td>
</tr>
<tr>
<td>v_vp_seq_le</td>
<td>28 Sb ----- s to INFINITIVE</td>
</tr>
<tr>
<td>v uninacc_le</td>
<td>1 Sth ----- s</td>
</tr>
<tr>
<td>v_le</td>
<td>2 Sb ----- s</td>
</tr>
<tr>
<td>v_pp_unacc_le</td>
<td>8 Sb ----- s sth</td>
</tr>
<tr>
<td>v_pp_noarg3_le</td>
<td>15 Sb ----- s sth to sb</td>
</tr>
<tr>
<td>v_pp_pp_e_le</td>
<td>20 Sb ----- s sb PP</td>
</tr>
<tr>
<td>v_pp^*cp_le</td>
<td>26 Sb ----- s that CLAUSE</td>
</tr>
<tr>
<td>v_pp_prop_le</td>
<td>20 Sb ----- s sb PP</td>
</tr>
<tr>
<td>v_cp_prop_le</td>
<td>21 Sb ----- s sth PP</td>
</tr>
<tr>
<td>v_npptr_le</td>
<td>26 Sb ----- s that CLAUSE</td>
</tr>
<tr>
<td>v_npptr_le</td>
<td>8 Sb ----- s sth</td>
</tr>
<tr>
<td></td>
<td>12 Sth ----- s sth</td>
</tr>
</tbody>
</table>

---

2Balinese (ISO 639-3: ban) is a Western Malayo-Polynesian language of the Austronesian language family. It belongs to the Malay-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 IN-
DRA was made, in which the verbs are alphabeti-
cally 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 be-
long to more than one verb type.

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

<table>
<thead>
<tr>
<th>Verb type</th>
<th>Number of verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>v_pp*/dir_le</td>
<td>76</td>
</tr>
<tr>
<td>v_vp.seq_le</td>
<td>49</td>
</tr>
<tr>
<td>v._unace_le</td>
<td>594</td>
</tr>
<tr>
<td>v_np_noarg3_le</td>
<td>5</td>
</tr>
<tr>
<td>v_np_pp.e_le</td>
<td>41</td>
</tr>
<tr>
<td>v_pp*/cp_le</td>
<td>23</td>
</tr>
<tr>
<td>v_np_pp_prop_le</td>
<td>85</td>
</tr>
<tr>
<td>v_cp_prop_le</td>
<td>53</td>
</tr>
<tr>
<td>v_np_itr_le</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>939</td>
</tr>
</tbody>
</table>

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

4 Analyzing Indonesian Phenomena

After creating INDRA via the Grammar Ma-
trix 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 sys-
tem, 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, ana-
lyzed 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 ini “this” can be mapped to mul-
tiple predicates, e.g. ini “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

```
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”

“This”). Figure 1 and 2 show the type hierarchy
for heads and demonstratives respectively.

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 loca-
tive pronouns situ and sana has the predicate me-
dial_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 ARGO in the se-
matteric head daughter di “at” is equated with the
INDEX which has the value e2. The value of the
ARG2 (x4) is coindexed with the ARGO of
place_n.rel and medial_q.rel. The medial_q.rel in-
troduces RSTR which is related to the top handle
of the quantifier’s restriction (h7) and linked to the
LBL of place_n.rel (h7=x, 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
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 Indonesian 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 menN- 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 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 kejar._rel is an event (e2) which is equated with the INDEX. The value of ARG0 of named_rel “adi” (x3) and named_rel “budi” (x9) 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://main.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.
Table 5: Morphology process of \textit{meN} (L = lost, R = retained; Sneddon et al., 2010: 13-18)

<table>
<thead>
<tr>
<th>Allomorph of meN-</th>
<th>Initial orthography of the base</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>mem-</td>
<td>p</td>
<td>(L) \textit{menggunakan} “use”</td>
</tr>
<tr>
<td>mem-</td>
<td>pl, pr, ps, pt, b, bl, t, ll, tr, v</td>
<td>(R) \textit{membeli} “buy”</td>
</tr>
<tr>
<td>men-</td>
<td>t</td>
<td>(L) \textit{mentanam} “plant”</td>
</tr>
<tr>
<td>meny-</td>
<td>tr, ts, d, dr, e, j, sl, sr, sy, sw, sp, st, sk, sm, sn, z</td>
<td>(R) \textit{mencairi} “seek”</td>
</tr>
<tr>
<td>meng-</td>
<td>s</td>
<td>(L) \textit{menyebrang} “rent”</td>
</tr>
<tr>
<td>meng-</td>
<td>k</td>
<td>(L) \textit{mengirim} “send”</td>
</tr>
<tr>
<td>me-</td>
<td>m, n, ny, ng, l, r, w, y</td>
<td>(R) \textit{melanggar} “break”</td>
</tr>
<tr>
<td>menge-</td>
<td>(base with one syllable)</td>
<td>\textit{mengecek} “check”</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>results / items coverage</th>
<th>before 52 / 172</th>
<th>after 55 / 172</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>30.2%</td>
<td>32.0%</td>
</tr>
</tbody>
</table>

Figure 6: MRS representation of \textit{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 \texttt{[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.

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 \textit{That’s what you meant: a Rich Representation for Manipulation of Meaning} (MOE ARC41/13).

References


