Confirmation Report

An Indonesian Resource Grammar (INDRA) in the Framework of Head-Driven Phrase Structure Grammar (HPSG) and its Application to Machine Translation

David Moeljadi
Division of Linguistics and Multilingual Studies
School of Humanities and Social Sciences
Nanyang Technological University

Thesis Advisory Committee:
1. Dr Francis Bond (Supervisor)
2. Dr I Wayan Arka (Co-Supervisor)
3. Dr František Kratochvíl

This report contains 21,063 words

25 May 2015

Contents

Abstract

The present report is a proposal of my PhD topic "An Indonesian Resource Grammar (INDRA) in the Framework of Head-Driven Phrase Structure Grammar (HPSG) and its Application to Machine Translation". This report describes the creation and the initial stage development of a broad-coverage Indonesian grammar, using the framework of HPSG (Sag et al., 2003) with Minimal Recursion Semantics (MRS) (Copestake et al., 2005). The approach taken is a corpus-driven approach. The scope is on the analysis and computational implementation of Indonesian text in the Nanyang Technological University Multilingual Corpus (NTU-MC) (Tan & Bond, 2012). Previous work on the computational grammar of Indonesian are mainly done in the framework of Lexical-Functional Grammar (LFG) such as Arka (2010a) and Musgrave (2001). To the best of my knowledge, no work on Indonesian HPSG has been done. Thus, the development of INDRA can also function as an investigation of the cross-linguistic potency of HPSG and MRS.

At the present stage, INDRA covers the basic phenomena in the Indonesian grammar and focuses on verbal constructions and subcategorization since they are fundamental for argument and event structure. The lexicon was semi-automatically acquired from the English Resource Grammar (ERG) (Copestake & Flickinger, 2000) via Wordnet Bahasa (Nurril Hirfana et al., 2011; Bond et al., 2014). The coverage, i.e. the quality and the quantity of parsed sentences in the corpus by the grammar, is evaluated using test-suites. The parsed sentences are treebanked for the next development of the grammar. The last chapter in this report addresses the plan for future work. In the future, INDRA will be used in the development process of multilingual machine translation.

List of Abbreviations and Conventions

-	affix boundary
=	clitic boundary
1	first person
2	second person
3	third person
ACT	active
ADJ	adjective
ADV	adverb
AUX	auxiliary verb
CLF	classifier
COP	copula
DEF	definite
DET	determiner
EXCL	exclusive
FUT	future
INCL	inclusive
NEG	negation
NP	noun phrase
PASS	passive
PL	plural
PP	prepositional phrase
PRF	perfect
PROG	progressive
RED	reduplication
REL	relativizer
SG	singular
VP	verb phrase

List of Tables

List of Figures

Chapter 1

Introduction

> "Languages are objects of considerable complexity, which can be studied scientifically."
>
> Sag et al. (2003: 2)

The present report describes some preliminary work done on the creation and the initial stage development of a grammar[1] of Indonesian (Indonesian Resource Grammar or INDRA)[2] up to the present time and the future development. INDRA will be the first Indonesian Head-Driven Phrase Structure Grammar (HPSG) (Sag et al., 2003), implemented as a computational grammar which can parse[3] and generate[4] Indonesian text. For my PhD purpose, INDRA will parse the Indonesian text in the Nanyang Technological University Multilingual Corpus (NTU-MC) (Tan & Bond, 2012). The present state of INDRA is available to be examined and can be downloaded from GitHub.[5] The computational tools employed to create INDRA and INDRA itself are open source.

[1] The term "grammar" in this report follows the definition of grammar in Wasow (2004) which focuses on syntax and morphosyntax and includes semantics.

[2] INDRA stands for INDonesian Resource grAmmar. The word *indra* in Indonesian is polysemous: **[1]in·dra** *n* organ to taste, smell, hear, see, touch and feel something instinctively (intuitively); **[2]in·dra** *n* **1** king; **2** *Hinduism* (written with a capital letter) name of a god who rules the sky (Alwi et al., 2008). It is also a common Indonesian male name.

[3] Parsing is the act of determining the syntactic structure of a sentence. The goal is typically to represent "who did what to whom" in the sentence (Sproat et al., 2004: 608).

[4] Generating is the act of defining in a formally precise way a set of sequences (strings over some vocabulary of words) that represent the well-formed sentences of a given language (Sag et al., 2003: 525) (see also Section 1.2 on page 3).

[5] `https://github.com/davidmoeljadi/INDRA`

This report is divided into five chapters: Chapter 1 contains brief information about the linguistic part which includes the Indonesian language and the grammar theory; Chapter 2 provides a brief explanation of the computational part, i.e. the development environment, the tools employed to build and develop INDRA as well as machine translation; Chapter 3 explains the research method; Chapter 4 discusses some preliminary work; finally, Chapter 5 describes my research plan.

In this chapter, Section 1.1 explains the statement of research; Section 1.2 contains a brief description about grammar engineering; Section 1.3 provides the historical, sociolinguistic, typological, morphological and syntactic aspects of the Indonesian language as well as some previous work done on the documentation of the Indonesian grammar or reference grammars; Section 1.4 and Section 1.5 introduce the background theory to analyze the Indonesian grammar as well as some previous work done within the Lexical Functional Grammar (LFG) framework for Indonesian grammar.

1.1 Statement of research

The aim of this research is to build and develop an Indonesian resource grammar (INDRA) implemented within the framework of Head-Driven Phrase Structure Grammar (HPSG) (Sag et al., 2003) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005) using tools developed by The Deep Linguistic Processing with HPSG Initiative (DELPH-IN) research consortium.[6] INDRA will parse and generate the Indonesian text of Singaporean tourism corpus in the Nanyang Technological University Multilingual Corpus (NTU-MC) (Tan & Bond, 2012)[7] and will be applied for multilingual machine translation.

[6] `http://www.delph-in.net`
[7] `http://compling.hss.ntu.edu.sg/ntumc/`

1.2 Grammar engineering

The common practice of language documentation or descriptive grammar may include the following areas: phonology (the study of sound systems), morphology (the study of word structure), syntax (the study of sentence structure), semantics (the study of language meaning) and pragmatics (the study of language use). Grammar engineering is similar to grammar documentation because it tries to describe the language as used by native speakers but it is particularly focused on syntax. In addition, grammar engineering allows us to have the computer do the work of checking the models for consistency and to test against a much broader range of examples (Bender & Fokkens, 2010).

Sag et al. (2003: 13, 21) note that syntax plays a crucial role in human language processing because it imposes constraints on how sentences can or cannot be construed and develops a set of rules that will predict the acceptability of (a large subset of) sentences in a language. Some of the goals of grammar engineering are:

1. to tell for any arbitrary string of words whether it is a well-formed sentence or not and to give a possible range of syntactic and semantic representations (interpretations)

2. to consider how the grammar of one language differs from the grammar of other languages

3. to consider what our findings might tell us about human linguistic abilities in general

The well-formedness of sentences can be tested by asking native speakers for their judgments of acceptability. However, considering variation across speakers, linguistic and nonlinguistic context, the use of multiple sources such as data from actual usage, i.e. written and spoken corpora, is always a good idea (Sag et al., 2003: 2-3).

Flickinger et al. (2010) mention that the necessary components in grammar engineering are as follows:

1. Linguistic theory. A solid linguistic theory which has rigid mathematical foundation, tractable computational model and universal to different languages (see Section 1.4 on page 11).

2. Grammar engineering platform, which is used for implementation of the formalism (description language). It should have grammar editor, processor which includes parser and generator, graphical user interface, profiling system and treebanking tools (see Section 2.2 on page 20).

3. Linguistic resources, such as corpora, test-suites, treebanks and reference grammars which include existing grammars for other languages on the same platform and existing grammars for one language on other platforms (see Section 2.1 on page 18, Section 2.3 on page 24 and Section 2.5 on page 26).

4. Methodology (see Chapter 3 on page 32)

INDRA aims to be a computational grammar which has the goals of telling the well-formedness of sentences in Indonesian, giving a possible range of interpretations, as well as considering the similarities and differences between Indonesian and other languages. INDRA also aims to have all the necessary components in grammar engineering field mentioned above.

1.3 Indonesian language

This section provides a brief description of the Indonesian language. Firstly, the historical and sociolinguistic background of Indonesian are introduced. Afterwards, the typological, morphological and syntactic aspects are briefly explained.

1.3.1 Historical and sociolinguistic background

Indonesian (ISO 639-3: ind), called *bahasa Indonesia* (lit. the language (*bahasa*) of Indonesia) by its speakers, is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Malayic branch with Standard Malay spoken in Malaysia, Brunei Malay in Brunei, Local

Figure 1.1: Malay dialects (Adelaar, 2010: 203)

Malay in Singapore and other Malay varieties spoken at various places in Indonesia such as Minangkabau and Makassar Malay (Lewis, 2009) (see Figure 1.1). The Indonesian language 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: 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. It is over 80% cognate with Standard Malay (Lewis, 2009).

The history of the Indonesian language cannot be separated from its diglossic[8] nature which exists from the very beginning of the historical record when it is called Old Malay around the seventh century A.D. to the present day (Paauw, 2009: 3). With the coming of Islam and during the era of the Malay kingdoms

[8]Diglossia is a relatively stable language situation in which there are two types of language variety: "High" language variety which is learned by formal education and is used for most written and formal spoken purposes and "Low" language variety which is used for ordinary conversation (Ferguson, 1959: 336).

from the twelfth to the nineteenth century A.D., a literary variety of Malay, known as Classical Malay, was codified and spread throughout the Malay world as a court language (Paauw, 2009: 14). During this period, the literary Malay in the Riau and Lingga group of islands, under Dutch influence, became the basis of the present-day standard Indonesian, while the one in the Malay peninsula, under British influence, is nowadays known as Standard Malay. The Dutch colonial used the literary language as a language of colonial administration and (to a limited extent) education and tried to develop and standardize the language. Classical Malay which was generally written in Arabic script, began to be uniformly written in Latin script after Charles Adrian van Ophuijsen introduced his system of spelling in 1901 (Abas, 1987: 81-86).

One significant milestone that marks the adoption of Malay as the national language in the present-day Indonesia is the Pledge of Youth by members of youth organizations in the first All Indonesian Youth Congress on 28 October 1928. The language was named Indonesian in the pledge. On the following day after the declaration of independence on 17 August 1945, the 1945 Constitution of the Republic of Indonesia was promulgated, in which it is stated in Section XV, Article 36 that the language of the state is Indonesian (*Bahasa Negara ialah Bahasa Indonesia*). Since then, many efforts have been done to standardize the Indonesian language.

Two years after the independence, in 1947 the Faculty of Letters and Philosophy of the University of Indonesia set up an institute for language and cultural studies, which became the Division of Language and Culture (*Lembaga Bahasa dan Budaya*) in 1952 (Montolalu & Suryadinata, 2007: 44). After changing its name several times, in 2010 it became the Institution for Language Development and Cultivation (*Badan Pengembangan dan Pembinaan Bahasa*), under the Ministry of Education and Culture. Montolalu et al. (2007: 44) note that the main responsibility of this institution is to ensure that Indonesian becomes a national language in its true sense. One of the domains that has received much attention is the spelling reform. The Soewandi spelling system, released in 1947, replaced the 1901 Van Ophuijsen spelling system (Abas, 1987: 83-88). It was then replaced

by the Perfected Spelling System (*Ejaan Bahasa Indonesia yang Disempurnakan*) which was released in 1972 (Abas, 1987: 99-103). The Perfected Spelling System was then revised in 1987 and 2009.

The language institute also publishes some dictionaries, creates new terms and provides support for the standardization and propagation of the language. Pusat Bahasa published in 1988 a standard reference dictionary called "The Great Dictionary of the Indonesian Language of the Language Center" (*Kamus Besar Bahasa Indonesia Pusat Bahasa* or KBBI). This dictionary contains approximately 62,000 lemmas. Since then it has been revised and expanded three times: the second edition in 1991 contains approximately 72,000 lemmas, the third edition in 2001 contains approximately 78,000 lemmas, and the current fourth edition in 2008 contains approximately 90,000 lemmas. The third edition has been made online to public and has an official site (Alwi et al., 2008).[9] Another important decision made is to publish a good standard grammar entitled "A Standard Grammar of the Indonesian Language" (*Tata Bahasa Baku Bahasa Indonesia*) in 1988. This reference grammar has been revised and expanded two times in 1993 and in 1998 (Alwi et al., 2014).

While much attention has been paid to the development and cultivation of the standard "High" variety of Indonesian, little attention has been particularly paid to describing and standardizing the "Low" variety of Indonesian. Sneddon (2006: 4-6) calls this variety "Colloquial Jakartan Indonesian" and states that it is the prestige variety of colloquial Indonesian in Jakarta, the capital city of Indonesia, and is becoming the standard informal style. Paaw (2009: 40) mentions that Colloquial Jakartan Indonesian is a variety which has only been recognized as a separate variety recently. Historically, it developed from the Low Malay varieties spoken in Java by Chinese immigrant communities, which have been termed "Java Malay". It has also been influenced by the Betawi language of Jakarta, a Low Malay variety which is thought to have been spoken in the Jakarta region for over one thousand years.

In addition to this "Low" variety, the regional languages spoken in various

[9] `http://bahasa.kemdiknas.go.id/kbbi/index.php`

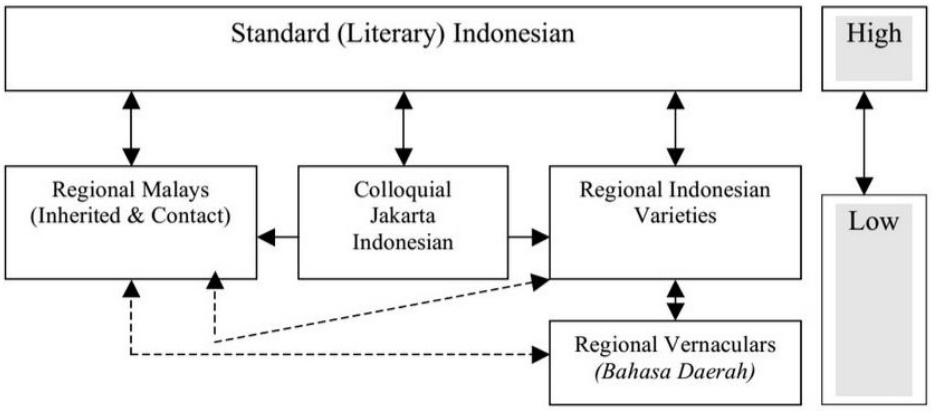


Figure 1.2: Diglossic situation in Indonesia (Paauw, 2009: 16)

places in Indonesia which count more than 500 add the complexity of sociolinguistic situation in Indonesia. The "High" variety of Indonesian is used in the context of education, religion, mass media and government activities. The "Low" variety of Indonesian is used for everyday communication between Indonesians. The regional vernaculars (*bahasa daerah*) are used for communication at home with family and friends in the community. In some areas, Indonesian coexists with yet another regional lingua franca, which is often a Malay variety. For example, in the city of Larantuka of Flores Island, locals speak Indonesian, Larantuka Malay, and their local language Lamaholot in different contexts. This complex situation is well described in Paauw (2009) and shown in Figure 1.2.

1.3.2 Indonesian grammar and reference grammars

Because of its historical and sociolinguistic aspects, the "High" variety of Indonesian is perhaps the most well-described and studied Austronesian language of Indonesia with several comprehensive grammar available such as Macdonald (1976), Mintz (2002), Liaw (2004), Sneddon et al. (2010) and Alwi et al. (2014). Hammarström et al. (2013) provide a comprehensive collection of 127 references of descriptive work in Indonesian such as reference grammars written in English,

Table 1.1: Table of consonants in Indonesian (Soderberg & Olson, 2008)

	Bi-labial	Labio-dental	Dental	Alveo-lar	Post-alveolar	Palatal	Velar	Glottal
Plosive & affricate	p b		t̪	d	tʃ dʒ		k g	(ʔ)
(Orthography)	p b		t	d	c j		k/q g	k/q
Nasal	m			n		ɲ	ŋ	
(Orthography)	m			n		ny	ng	
Flap/trill				r				
(Orthography)				r				
Fricative		(f)		s (z)	(ʃ)			h
(Orthography)		f/v		s z	sy			h
Approximant	w					j		
(Orthography)	w					y		
Lateral approximant				l				
(Orthography)				l				

Indonesian, Dutch, German, French and Russian, dictionaries and texts.

Indonesian has 22 consonants, 6 vowels (monophthongs) and 3 diphthongs (Soderberg & Olson, 2008; Alwi et al., 2014: 55-80). Table 1.1 shows all 22 consonants in Indonesian and the orthography. Figure 1.3 shows all 6 vowels and 3 diphthongs in Indonesian. The orthography for /e/ and /ə/ is <e>. Bracketed phonemes only appear in loan words or morphological boundaries. 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 in Indonesian are derivational (Sneddon et al., 2010: 29). When affixes combine with roots or stems, a number of phonetic or phonological alternations through morphophonemic processes occur. Indonesian has a strong tendency to be head-initial (Macdonald, 1976: 24-25; Sneddon et al., 2010: 26-28). In a noun phrase with an adjective or a demonstrative or

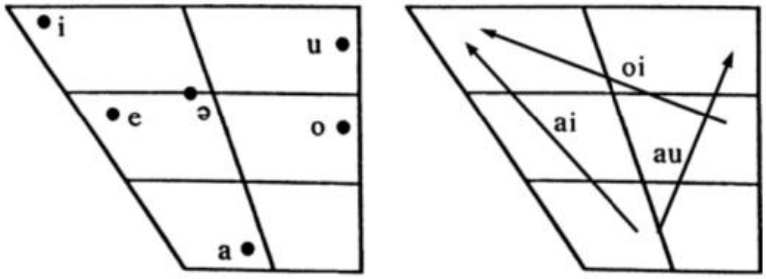


Figure 1.3: Monophthongs and diphthongs in Indonesian (Soderberg & Olson, 2008)

a relative clause, the head noun precedes the adjective or 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 changing operations are triggered by passive and applicative constructions.

Macdonald (1976) covers many important phenomena such as affixes, reduplication, conjunction, passive voice, existential sentence, topic-comment sentence and various sentence patterns. Mintz (2002) compares Indonesian and standard Malay grammar and gives examples in both varieties. It emphasizes the order of words and phrases in a sentence. Liaw (2004) provides compact information on various grammatical phenomena with many examples. Compared with other reference grammars, he discusses more on prepositions, adverb-forming affixes, subordinative compounds and some minor sentences such as those used in block language and formulaic language. Sneddon et al. (2010) is perhaps the best reference grammar for Indonesian. It contains many descriptions of demonstratives, classifiers, partitives, articles, proper nouns and common nouns, count nouns and mass nouns, indefinite pronouns, aspect auxiliaries, particles, conjunction, derivative affixes, types of reduplication, relative clause and various sentence patterns. Alwi et al. (2014) has more analyses and examples of verbs based on its morpholog-

ical and syntactic types. It distinguishes verb based on the morphology, whether it should be used without affixes, with optional affixes, or obligatory affixes, as well as the number of arguments in a sentence, i.e. intransitive, semitransitive, monotransitive and ditransitive.

Some details or phenomena not covered in one reference grammar may be covered in another one and vice versa. However, some disagreement about a particular phenomenon may also occur. For example, many reference grammars agree that *adalah* and *ialah* are copulas (Alwi et al., 2014: 358-359; Liaw, 2004: 191; Macdonald, 1976: 155-156; Mintz, 2002: 22; Sneddon et al., 2010: 246). In addition, some reference grammars seem to agree that *merupakan* can be considered as a copula or expressing the concept of 'to be' (Alwi et al., 2014: 359; Mintz, 2002: 425), while others such as Sneddon et al. (2010: 247) do not. Moreover, some reference grammars include *menjadi* as copulative or functioning as a copula (Macdonald, 1976: 94, 132) or expressing the concept of 'to be' (Mintz, 2002: 423-425), while Sneddon et al. (2010: 247) regards it as a full verb and not a copula. Because the reference grammars complement one another and comparing their analyses is essential, I have started compiling an extensive bibliography or summary of language phenomena covered in Indonesian reference grammars mentioned above and linguistics papers in order to give the best analysis for the development of INDRA. Table 4.5 on page 46 shows a small extract of the extensive bibliography of language phenomena.

1.4 Head-Driven Phrase Structure Grammar and Lexical Functional Grammar

Many formal models of the syntax of natural language are Context-Free Grammars (CFG), also called Phrase-Structure Grammars. These grammars are based on constituent structure, which was formalized in 1956 by Chomsky. CFG consists of a set of rules and a lexicon of symbols (parts-of-speech) and words. The set of rules groups and orders the symbols. The lexicon connects the symbols with the words. Head-Driven Phrase Structure Grammar (HPSG) adds the idea that the

lexical head of each constituent or the grammatically most important word in a phrase is passed up the parse tree (Jurafsky & Martin, 2009).

The analysis of the Indonesian grammar in INDRA uses the theoretical framework of HPSG (Sag et al., 2003). HPSG is surface oriented, positing no additional abstract structures and thus providing a reasonably simple structure that is directly associated with the string of words that form each sentence. HPSG is mono-stratal, i.e. orthography, syntax, semantics, pragmatics are all handled in a single structure or the sign. The sign is the primary or elementary unit in HPSG, modeled through typed feature structures. Signs in HPSG include words, phrases, sentences and utterances (Sag et al., 2003). Types are classes of linguistic entities. Each type is associated with a particular feature structure. Feature structures are sets of feature or attribute and value pairs which represent objects. Features or attributes are unanalyzable atomic symbols from some finite set and values are either atomic symbols or feature structures themselves. Feature structures are usually illustrated with an Attribute-Value Matrix (AVM) or a Directed Acyclic Graph (DAG) (Jurafsky & Martin, 2009: 524-526; Fokkens, 2014: 27-28). Figure 1.4 on the following page illustrates the AVM of *anjing* "dog". In this AVM, the feature PNG (person-number-gender) has feature structures as its value: the feature PERNUM (person-number) has a value *3rd* (third person) and the feature ANIMACY has *non-human* as its value. The singular or plural number is not specified because a bare noun phrase in Indonesian such as *anjing* "dog" can either be singular, plural or generic as follows:

(1.1) (a) *Ini* *anjing.*
this dog
"This is a dog." (own data)

(b) *Ini* *semua* *anjing.*
this all dog
"These are dogs." (own data)

(c) *Anjing* *menggonggong.*
dog bark
"Dogs bark." (own data)

$$\begin{bmatrix} \textit{noun-relation} & \\ \text{PRED} & \textit{\_anjing\_n\_rel} \\ \text{ARG0} & \begin{bmatrix} \textit{ref-ind} & \\ \text{PNG} & \begin{bmatrix} \textit{png} & \\ \text{PERNUM} & \textit{3rd} \\ \text{ANIMACY} & \textit{non-human} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 1.4: AVM of *anjing* "dog"

The ANIMACY is *non-human* because Indonesian uses different classifiers for inanimate, human, and non-human (see Figure 4.1 on page 48 in Section 4.2.1).

Lexical entries, lexical rules and phrase structure rules are all feature structures. The feature structures in HPSG are typed and hierarchical, i.e. in an ontology of linguistic objects or types, the supertypes of the lexical types subsume the subtypes and pass down the constraints to the subtypes. Subtypes inherit all properties from their supertypes and a subtype can have more than one supertype. HPSG is unification- and constraint-based. In unification, the words and phrases are combined into larger expressions according to constraints of the lexical entries based on the type hierarchy. HPSG is lexicalist, i.e. the bulk of syntactic and semantic properties are defined in the lexicon. HPSG embodies the claim that determining which strings of words constitute well-formed sentences and specifying the linguistic meaning of sentences depend mostly on the nature of words (Sag et al., 2003: 168). Lexical items are like words and phrases, contain information about phonology, syntax and semantics. Constructions (phrase-rules) are also modeled as feature structures. This allows constructions to be analyzed via multiple inheritance hierarchies, modeling the fact that constructions cluster into groups with a 'family resemblance' that corresponds to a constraint on a common supertype.

Feature structures are also found in Lexical Functional Grammar (LFG) (Ka-

plan & Bresnan, 1982) and other formalisms. However, LFG takes a modular approach using at least two levels or aspects of syntactic structure: functional-structure (f-structure) and constituent-structure (c-structure). F-structure is the abstract functional syntactic organization, representing predicate-argument structure and functional relations like subject and object in Attribute-Value Matrices (AVM). On the other hand, c-structure is the overt, more concrete level of linear and hierarchical organization of words into phrases, represented in a tree diagram. C-structure nodes and f-structures are linked by a φ function that maps every c-structure node to exactly one f-structure. The sign in HPSG is similar to a tuple of LFG structures, consisting of at least f-structure and c-structure. Although feature structures in LFG are not typed, templates express generalizations over functional descriptions which can be defined in terms of other templates, similar to inheritance hierarchy in HPSG (Dalrymple et al., 2004).

To the best of my knowledge, there is no previous work done on Indonesian HPSG but much has been done on Indonesian LFG. Arka & Meladel (2011) discuss the syntactic and the functional-semantic/pragmatic constraints associated with all types of negation in Indonesian. Arka (2011) and Arka (2013b) examine the relation between the expressions of tense, aspect, and mood (TAM) and possessive/definite nominalization *=nya* in Indonesian. Both papers demonstrate that *=nya* encodes past time reference. Arka & Manning (2008) deal with active and passive voice in Indonesian. Arka (2010a) deals with categorial multifunctionality, i.e. the phenomenon of the same form of word or phrase which can appear in different functions in a sentence: as a predicate, a modifier, and an argument. Arka et al. (2009) present an analysis of the suffix *–i* in Indonesian, focusing on the issues of applicative-causative polysemy of the suffix and its alternation with suffix *–kan*. Meladel et al. (2009) investigate reduplication in Indonesian, focusing on verb reduplication that has the agentive voice affix *meN-*. Arka (2013a) discusses the nonverbal predication, showing that it is related to the future/nonfuture tense and the nominal predication is syntactically different from adjectival and prepositional predication. He argues that the absence of copula is associated with nonfuture tense, while an inchoative predication must be used for the nominal

predication in the future tense. His finding that Indonesian does show a grammatical tense in relation to syntactic restriction uncovered a phenomenon which many grammarians fail to see. Arka (2014) works on syntactic, semantic and pragmatic properties of Indonesian control constructions with an interplay among the voice system, morpho-semantic-lexical properties of the control verbs, animacy of the argument and information structure of argument focusing. Arka (2010b) reports his analysis on two passives in Indonesian with *di-* and *ter-*. Musgrave (2001) in his Phd thesis concludes that the contrasting properties of non-subject arguments of prefixed and bare verbs suggest a divide between two types of clause structure in Indonesian. He argues that the LFG theory as formulated is unable to give an adequate account of this idea. Because of the similarity between HPSG and LFG, the above-mentioned previous work done in LFG can be employed as the starting point of grammar analysis.

1.5 Minimal Recursion Semantics

INDRA uses Minimal Recursion Semantics (MRS) (Copestake et al., 2005) as its semantic framework. MRS is adaptable for HPSG typed-feature structure cross-linguistically and suitable for parsing and generation. The semantic structures in MRS are underspecified for scope and thus suitable for representing ambiguous scoping. The primary interest in MRS is in finding the correct lexemes and the relationships between them that are licensed by the syntax. Thus, a representation language where scope can be ignored when it is irrelevant but retrieved when it is needed is required (Copestake et al., 2005: 286).

The MRS basic semantic units of lexemes, called elementary predicates (EP), indicate single relations with the associated arguments, e.g. beyond(x, y). The structure of EPs is flat (no embedding) and labelled, i.e. the scoping information is indicated with handles (h). In HPSG, EP is represented as a feature structure with arguments and label. Figure 1.5 shows the MRS feature structure of Example 1.1 c along with its non-feature structure equivalent.

In Figure 1.5, the overall structure is of type *mrs* and it has features TOP,

$$\begin{bmatrix} mrs & \\ \text{TOP} & \boxed{0}\ h \\ \text{INDEX} & \boxed{2}\ e \\ \text{RELS} & \left\langle \begin{bmatrix} \_anjing\_n\_rel \\ \text{LBL} & \boxed{4}\ h \\ \text{ARG0} & \boxed{3}\ x \end{bmatrix}, \begin{bmatrix} exist\_q\_rel \\ \text{LBL} & \boxed{5}\ h \\ \text{ARG0} & \boxed{3}\ x \\ \text{RSTR} & \boxed{6}\ h \\ \text{BODY} & \boxed{7}\ h \end{bmatrix}, \begin{bmatrix} \_menggonggong\_v\_rel \\ \text{LBL} & \boxed{1}\ h \\ \text{ARG0} & \boxed{2}\ e \\ \text{ARG1} & \boxed{3}\ x \end{bmatrix} \right\rangle \\ \text{HCONS} & \left\langle \begin{bmatrix} qeq \\ \text{HARG} & \boxed{0}\ h \\ \text{LARG} & \boxed{1}\ h \end{bmatrix}, \begin{bmatrix} qeq \\ \text{HARG} & \boxed{6}\ h \\ \text{LARG} & \boxed{4}\ h \end{bmatrix} \right\rangle \\ \text{ICONS} & \langle\ \rangle \end{bmatrix}$$

$\langle h0,\ h5{:}\text{exist\_q}\{(x3,h6,h7),\ h4{:}\text{\_anjing\_n}(x3),\ h1{:}\text{\_menggonggong\_v}(e2,x3)\},$
$\{h0{=}_q h1,\ h6{=}_q h4\}\rangle$

Figure 1.5: MRS feature structure of *anjing menggonggong* "dogs bark"

INDEX, RELS, HCONS and ICONS. RELS is a feature that introduces a list of EPs, in which all types denoting relations end in *_rel*. In EPs, all lexical items have features LBL and ARG0. The value of LBL is a handle (h), which is used to express scope relations. The value of ARG0 is an individual (x) or an event (e). For common nouns such as *anjing* "dog", the value of ARG0 will be a referential index which serves as a pointer to the entity referred to by the NP and as the value of the ARG1 feature of the lexical predicate *menggonggong* "bark" which selects the NP *anjing* "dog" as its semantic argument. The ARG0 in the semantic head daughters *menggonggong* "bark" is equated with the INDEX which has the value $e2$.

The lexical type for quantifiers like some, every and exist introduces a quant-relation type which introduces two additional features: RSTR (restrictor) and BODY. As scopal features, they have handles as their values. The RSTR is

related to the top handle of the quantifier's restriction ($h6$). The BODY is left unbound: this is what allows quantifier to have varied scoping possibilities. The value of ARG0 ($x3$) is the referential index that the quantifier binds.

The HCONS (handle constraint) is a set of handle constraints which reflect syntactic limitations on possible scope relations among the atomic predications. Handles which appear as the value of scopal arguments are called holes. The *qeq* (equality modulo quantifiers) relation (or $=_q$) indicates that either the value of LARG label l fills the hole h in HARG (i.e. $h{=}l$), or l is indirectly linked to h via one or more "floating quantifiers". In this case, l is either the body of an argument filling h, or an argument that is the body of an argument directly or indirectly filling h (Copestake et al., 2005: 10).

The ICONS (individual constraint) deals with anaphors and information structure. It constraints two referential entities and says that the the two variables refer to the same thing. In Figure 1.5, there is no anaphor and thus the ICONS is empty.

Chapter 2

Computational Background

This chapter provides the computational background of building INDRA. Section 2.1 contains an overview of the Deep Linguistic Processing with HPSG Initiative (DELPH-IN) consortium and the grammars developed within it as well as some aspects in which analyses in DELPH-IN grammars differ from the standard theoretical HPSG. Section 2.2 provides a brief explanation of the tools employed to build and develop INDRA. Section 2.3 describes the Nanyang Technological University Multilingual Corpus (NTU-MC). For my PhD purpose, INDRA is developed to be able to parse Indonesian sentences in that corpus. Section 2.4 gives a short introduction of machine translation. INDRA will be implemented for multilingual machine translation. Section 2.5 mentions several work done on computational linguistics related to Indonesian.

2.1 Deep Linguistic Processing with HPSG Initiative (DELPH-IN)

The Deep Linguistic Processing with HPSG Initiative (DELPH-IN)[1] consortium is a research collaboration between linguists and computer scientists which builds and develops open source grammar, tools for grammar development and Natural Language Processing (NLP) applications using HPSG and MRS.

[1] `http://www.delph-in.net`

```
anjing := nonhuman-noun-noun-lex &
  [ STEM < "anjing" >,
    SYNSEM.LKEYS.KEYREL.PRED "_anjing_n_rel" ].
```

Figure 2.1: TDL of a lexical item *anjing* "dog"

More than fifteen grammars of various languages with different typological characteristics have been created and developed within DELPH-IN,[2] some of them are as follows:

1. English Resource Grammar (ERG) (Flickinger, 2000). The first implemented HPSG grammar and a robust English grammar, used for several applications such as machine translation and language education. It has 38,294 lexical items, 81 lexical rules, 212 grammar rules, 215 features, 9,086 types, and a large treebank[3] called `Redwoods` (Oepen et al., 2002).[4]

2. Jacy (Siegel & Bender, 2002). A medium sized Japanese grammar, used for Japanese-English machine translation. It has 56,277 lexical items, 69 lexical rules, 51 grammar rules, 183 features, 2,519 types, and a treebank called `Hinoki` (Bond et al., 2004).

3. Zhong.[5] A grammar for Chinese languages, including simplified and traditional Mandarin Chinese as well as Cantonese and Min Nan. As of 12 May 2015, it has 43,070 lexical items, 2 lexical rules, 51 grammar rules, 197 features and 3,004 types.

DELPH-IN grammars define typed feature structures using Type Description Language (TDL) (Copestake, 2000). An example of TDL type definition from INDRA is shown in Figure 2.1.

[2] see `http://moin.delph-in.net/GrammarCatalogue` for the DELPH-IN grammar catalogue

[3] Treebank is a syntactically annotated corpus (see also Footnote 9 in this chapter).

[4] `http://lingo.stanford.edu/redwoods/`

[5] `http://wiki.delph-in.net/moin/ZhongTop`

The type identifier (in this case, a lexical item *anjing* "dog") stands left of the symbol `:=` which is followed by at least one supertype (in this case, a lexical type *nonhuman-noun-noun-lex*). Square brackets `[ ]` define further constraints on the type. A symbol `&` connects different supertypes or supertypes and constraints. Inside the square brackets, features and their values are separated from each other by white spaces. Dots can be used to define paths of features. `< >` is a notation for lists. If the lexical item contains more than one word, e.g. *anjing hutan* "dhole" (lit. dog forest), the value of the STEM would be `< "anjing", "hutan" >`.

There are some differences between the standard HPSG theory and its implementation in DELPH-IN grammars due to the computational efficiency (Fokkens, 2014). One of the differences is most rules in DELPH-IN grammars are either unary or binary rules.

2.2 The Development Environment

The DELPH-IN community provides several open-source tools for grammar development as well as an on-line wiki[6] for the documentation. Linguistic Knowledge Builder (LKB) (Copestake, 2002) is mainly used in the grammar development environment. The LKB system was initially developed at the University of Cambridge Computer Library. Its first version was implemented in 1991 and has been updated at the Center for the Study of Language and Information (CSLI), Stanford University (Copestake, 2002: ix). It is now part of the Linguistic Grammars Online (LinGO) project, partially supported by the National Science Foundation of the United States of America. The LKB system is a grammar and lexicon development environment for typed feature structure grammars which enables grammar developers, even linguists which have very little knowledge in computer science, to write grammars and lexicons for natural languages to be parsed and generated (Copestake, 2002: 6-7). LKB is particularly good for debugging process.

[6] `http://moin.delph-in.net/`

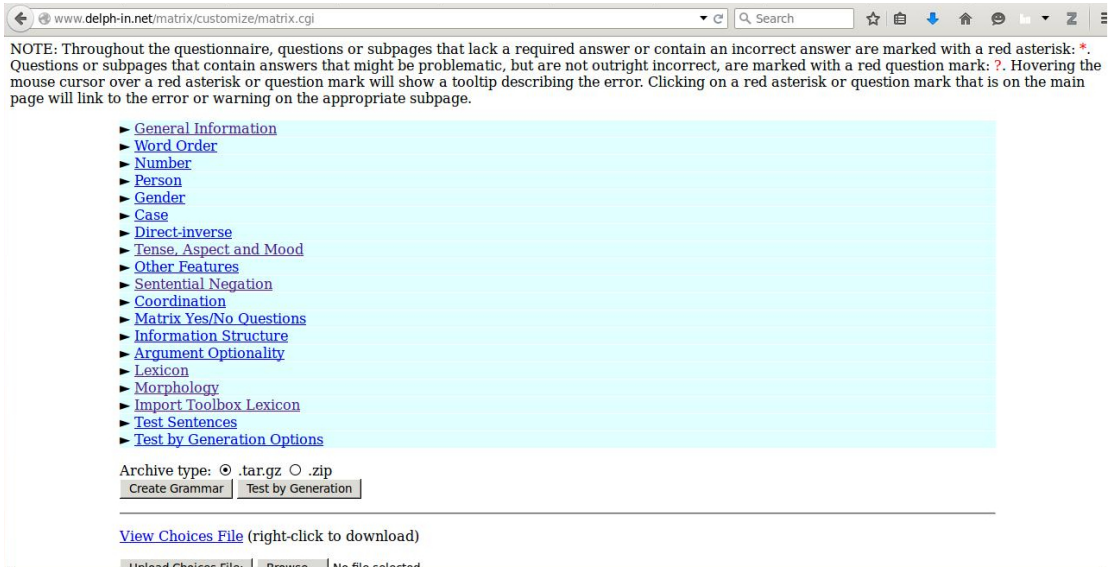


Figure 2.2: Screenshot of the LinGO Grammar Matrix's main page, taken on 22 April 2015

The LinGO Grammar Matrix (Bender et al., 2002; Bender et al., 2010)[7] is a web-based questionnaire for writing new DELPH-IN grammars, developed at the University of Washington (see Figure 2.2). The documentation for answering the questionnaire is provided in Fokkens et al. (2012). The matrix is originally based on ERG and Jacy (Fokkens, 2014). As a starter kit, it was created to be useful cross-linguistically, providing a wide range of phenomena and basic files to make the grammars compatible with DELPH-IN parsers and generators. After a list of questions in the questionnaire is filled out, the matrix generates a choices file (see Figure 2.3) and builds a new grammar. The new grammar can be manually edited and extended afterwards.

The fundamental element of the matrix is a file called `matrix.tdl`. It defines the top of the hierarchy of typed feature structures and general rules for combining subjects, complements or modifiers with a head as well as coordination structures.

[7] `http://www.delph-in.net/matrix/customize/matrix.cgi`

```
section=word-order          section=number
word-order=svo               number1_name=sg
has-dets=yes                  number1_supertype1_name=number
noun-det-order=noun-det      number2_name=pl
has-aux=yes                   number2_supertype1_name=number
aux-comp-order=before
aux-comp=vp                 section=person
multiple-aux=yes            person=1-2-3
                            first-person=incl-excl
```

Figure 2.3: A small extract of a choices file (see also Table 4.5 on page 46)

It also provides the basics to construct semantics based on MRS (Fokkens, 2014).

In addition to LKB and LinGO Grammar Matrix, several open-source tools provided by DELPH-IN for grammar development are as follows:

1. Parsers and generators:

   (a) Platform for Experimentation with efficient HPSG processing Techniques (PET) (Callmeier, 2000). PET reads the same source files as LKB and produces identical results but allows a high-efficiency batch processing.

   (b) Answer Constraint Engine (ACE) (Packard, 2013a).[8] An efficient processor for DELPH-IN HPSG grammars. Its parsing and generation performance are both around 15 times faster than LKB. In certain common configurations, ACE is significantly faster than PET.

   (c) another grammar engineering environment (agree) (Slayden, 2012). agree can work with DELPH-IN style TDL grammars configured for LKB or PET within Mono and Microsoft's .NET framework.

[8] http://sweaglesw.org/linguistics/ace/

2. Treebanking:[9]

   (a) ITSDB or [incr tsdb()] (pronounced *tee ess dee bee plus plus*) (Oepen & Flickinger, 1998). A tool for testing, profiling the performance of the grammar (analyzing the coverage and performance) and treebanking.

   (b) Full Forest Treebanker (FFTB) (Packard, 2014). A treebanking tool for DELPH-IN grammars, allowing the selection of an arbitrary tree from the "full forest" without enumerating/unpacking all analyses in the parsing stage. It is partly integrated with [incr tsdb()] and the LOGON tree.

3. Machine translation engine:

   (a) LOGON (Oepen et al., 2007). The LOGON infrastructure is a collection of software, grammars and other linguistic resources for transfer-based machine translation (MT).

   (b) Answer Constraint Engine (ACE) (Packard, 2013a)[10]

I make extensive use of ACE, ITSDB, FFTB and LOGON for the development of INDRA.

As a part of the grammar development process, in order to make tracking of modifications possible, the latest version of INDRA is regularly saved or backed up in GitHub[11] and contains the following items:

1. The grammar (including the lexicon)
2. Some test-suites (see Section 3.1) and treebanks
3. Python scripts for integrating the Indonesian POS Tagger (see Section 2.5.2) to INDRA

[9] Treebanking is a process of making a syntactically annotated corpus by annotating each sentence with a parse tree (Jurafsky & Martin, 2009: 438).

[10] `http://sweaglesw.org/linguistics/ace/`

[11] `https://github.com/davidmoeljadi/INDRA`

It is licensed under the MIT license, a free software license originating at the Massachusetts Institute of Technology (MIT), which grants any person to download, use, copy, modify and distribute the grammar if they include the copyright and permission notice. Information on how to download is on the GitHub page. In the future, the development of the grammar will be documented in DELPH-IN wiki.

2.3 Nanyang Technological University Multilingual Corpus (NTU-MC)

A corpus-driven approach is taken to the selection of phenomena to be worked on in my PhD. The target corpus for identifying Indonesian structures will be from the Nanyang Technological University Multilingual Corpus (NTU-MC) (Tan & Bond, 2012).[12] NTU-MC is developed by Division of Linguistics and Multilingual Studies at Nanyang Technological University in Singapore. It is a linguistically diverse, parallel corpus with information of part of speech (POS tagged) and the meaning or sense for each word. Each subcorpus in the NTU-MC is sense-tagged using a large lexical database called Wordnet (see 2.5.1 on page 27) (Bond et al., 2013).

NTU-MC is constructing a semantically annotated and linguistically diverse corpus with many Asian languages. It contains eight languages: English, Mandarin Chinese, Japanese, Indonesian, Korean, Arabic, Vietnamese, and Thai. The data are from four genres: Singapore Tourism Board (STB) websites,[13] Sherlock Holmes short stories, open source essays (the Cathedral and the Bazaar), and news (Kyoto Corpus translated by NICT) (Tan & Bond, 2012; Bond et al., 2013). The Indonesian text data from STB which contains 2,197 sentences will be employed as the target corpus.[14] My PhD work is to develop INDRA to parse at

[12] `http://compling.hss.ntu.edu.sg/ntumc/`

[13] `www.yoursingapore.com` and Singapore Medical Tourism pages

[14] Apart from the Indonesian text data, the STB parallel corpus contains 2,988 sentences in English, 2,332 sentences in Mandarin Chinese, and 2,723 sentences in Japanese (Bond et al., 2013: 152).

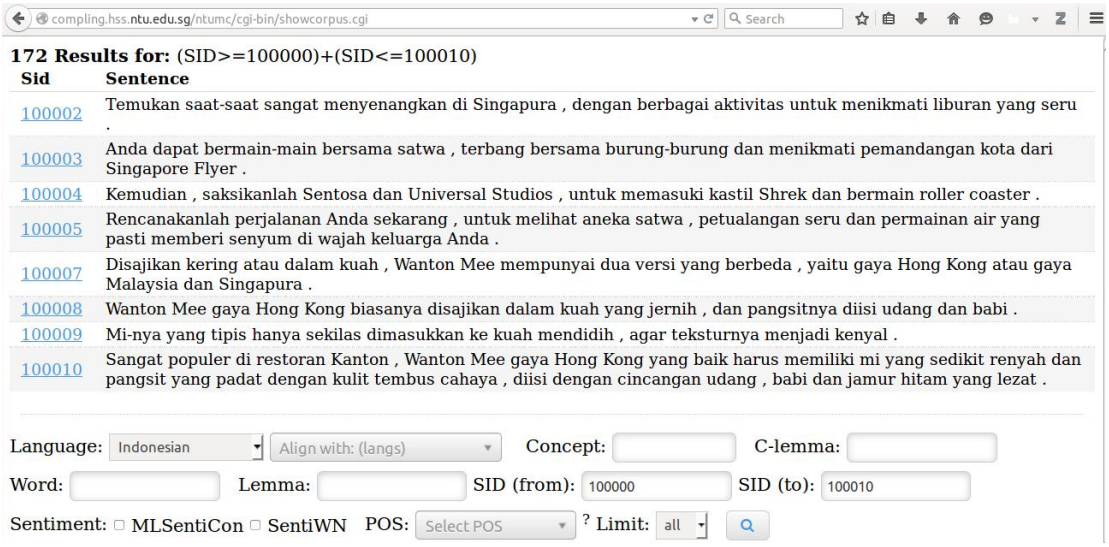


Figure 2.4: Screenshot of Indonesian text (sentence ID 100002 to 100010) in NTU-MC

least 60% of the Indonesian corpus. In addition to the Indonesian text data from STB, there is an ongoing process of adding Sherlock Holmes short stories to the Indonesian corpora.

Although parallel corpus such as NTU-MC with English as the source language is useful for contrastive analysis in particular, naturally occurring text is vital in order to make findings as verifiable as possible (Bond et al., 2015). In the future, INDRA aims to parse naturally occurring text such as the one collected in Moeljadi (2014b).

2.4 Machine translation

INDRA will be implemented for automatic translation from Indonesian to English and vice versa using computerized system or machine translation (MT). There are three classic MT models or rule-based MT systems (RBMT), i.e. the direct, transfer, and interlingua approaches; and modern statistical MT (SMT)

(Jurafsky & Martin, 2009: 898). Direct translation employs a large bilingual dictionary and translate word-by-word. Transfer model parses the source language sentence and applies rules to transfer into a target language parse then generates the target language sentence. It needs rules for syntactic transfer and lexical transfer. Interlingua MT uses abstract meaning representation or an interlingua which can generate the target language sentence. Statistical MT (SMT) employs probabilistic models trained from parallel corpora and chooses the most probable translation. RBMT may produce correct translation but it needs considerable amount of linguistic knowledge while SMT requires little linguistic knowledge but may not produce accurate translation.

INDRA uses transfer grammars for Indonesian to English MT and vice versa. The initial transfer grammars `inen` for Indonesian to English MT and `enin` for English to Indonesian MT were made by Sanghoun Song and made available in GitHub.[15] The transfer grammars use MRS which represents the meaning of the source language sentence in order to generate the correct translation for the target language. In other words, a source sentence is parsed to give a semantic representation and a transfer component converts this into a target semantic representation (Copestake et al., 2005: 283). In addition, lexical transfer is also employed to map the lexicon from the source language to the target language. In order to achieve a better result, SMT model may be integrated to generate a hybrid MT system.

2.5 Previous work on Indonesian computational linguistics

Previous work on computational linguistics related to Indonesian are divided into two parts: the lexical resource from Wordnet Bahasa (Nurril Hirfana et al., 2011; Bond et al., 2014) and the Indonesian morpheme segmentation and part-of-speech tagger using Indonesian POS Tagger (Rashel et al., 2014).

[15] `https://github.com/sanghoun/tm`

2.5.1 Wordnet Bahasa

Lexicon or a repository of words is vital in building and developing a grammar. Thus, a dictionary or a lexical database with comprehensive data of words is very important as a lexical source for a computational grammar. INDRA uses Wordnet Bahasa (Nurril Hirfana et al., 2011; Bond et al., 2014) as its lexical source. Wordnet Bahasa was created based on Princeton Wordnet (PWN) (Fellbaum, 2005), a large English lexical database.

PWN was built at the Cognitive Science Laboratory of Princeton University in 1985. Different from other dictionaries and lexical databases, PWN groups nouns, verbs, adjectives, and adverbs into sets of cognitive synonyms or synsets, each expressing a distinct concept. They are interlinked through a number of semantic relations (Fellbaum, 2005). Since its creation, many other wordnets in different languages have been built based on PWN (Bond & Paik, 2012; Bond & Foster, 2013). They are available online at the Extended Open Multilingual Wordnet (1.2),[16] hosted at Nanyang Technological University in Singapore. One of them, Wordnet Bahasa, is built as a lexical database in the Malay language (at present it contains data from Indonesian and Standard Malay).[17] It combines data from several lexical resources: the French-English-Malay dictionary (FEM), the KAmus Melayu-Inggeris (KAMI), and wordnets for English, French and Chinese (Nurril Hirfana et al., 2011: 258).

The initial version of Wordnet Bahasa in 2011 contains 19,210 synsets, 48,110 senses, and 19,460 unique words. Data from the Malaysian and Indonesian Wordnet were subsequently merged and its version 1.0 released in 2012 contains 49,668 synsets, 145,696 senses, and 64,431 unique words. Although many synsets and lemmas were included, errors in the language classification (e.g. Standard Malay words listed as Indonesian lemmas) and bad lemmas (English lemmas, lemmas containing numbers or characters other than alphabets, unlexicalized lemmas and lemmas in passive form) were found (Moeljadi, 2014a). I cleaned up the version

[16] `http://compling.hss.ntu.edu.sg/omw/`

[17] Wordnet Bahasa will be extended to accommodate other varieties of Malay language such as the Local Malay variety spoken in Singapore and Jakarta Malay.

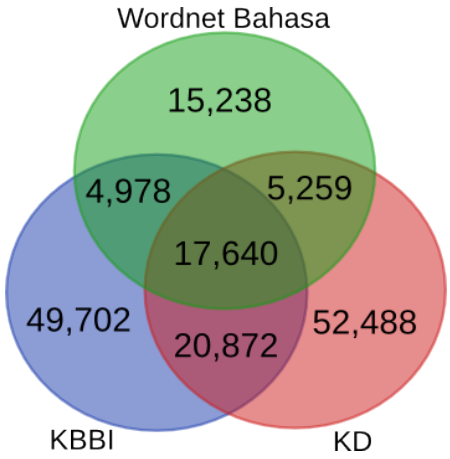


Figure 2.5: The number of lemmas in Wordnet Bahasa, KBBI and KD

1.0 using the data from The Great Dictionary of the Indonesian Language of the Language Center (*Kamus Besar Bahasa Indonesia Pusat Bahasa* or KBBI) and The Institute Dictionary (*Kamus Dewan* or KD).[18] In addition, more data were contributed: `Kateglo`[19] from Ruli Manurung and Wikipedia data from Lian Tze Lim. Definitions from Asian Wordnet project (Riza et al., 2010) were also added. After the data addition and the cleaning up process, the latest version (version 1.1) was released in 2015 and contains 40,493 synsets, 118,903 senses, and 43,113 unique words. The latest raw data is released under the MIT license.[20] Lemmas and synsets from KBBI and KD not yet included in Wordnet Bahasa will be further added (see Figure 2.5).

[18]The Institute Dictionary (*Kamus Dewan*) is a dictionary for Standard Malay published by The Institute of Language and Literature (*Dewan Bahasa dan Pustaka*). The fourth edition, which is the newest edition published in 2005, contains 75,387 lemmas including proverbs (Lian Tze Lim p.c.).

[19]`http://kateglo.com/`

[20]`https://sourceforge.net/p/wn-msa/tab/HEAD/tree/trunk/`

2.5.2 Indonesian Part-Of-Speech Tagger

Part-of-speech (POS) tagging is a process of assigning parts-of-speech to words. It takes words and tagsets as the input and produces a POS tag for each word as the output. It is useful for resolving ambiguity, i.e. to specify the correct part-of-speech of a word having more than one part-of-speech in a given context. In the development process of INDRA, POS tags are important particularly for dealing with unknown words. INDRA uses Indonesian Part-Of-Speech (POS) Tagger (Rashel et al., 2014),[21] a tool developed by a team from the Faculty of Computer Science of the University of Indonesia for POS tagging in Indonesian. It receives Indonesian text as an input and gives lines of words with the respective POS tags as an output.

Indonesian POS Tagger is based on Indonesian Morphological Analyzer (MorphInd) (Larasati et al., 2011),[22] a tool which handles both morphological analysis and lemmatization for a given surface word form. Indonesian POS Tagger applies a rule-based approach by employing hand-written disambiguation rules. It has six main modules: multi-word expression tokenizer, name entity recognizer, closed-class word tagging, open-class word tagging, rule-based disambiguation, and resolver for unknown words. It applies tags to every token, starting from closed-class words to open-class words and disambiguates the tags based on a set of manually defined rules. It has 23 tagsets of parts-of-speech, such as NNP for proper nouns, NN for common nouns, VB for verbs, MD for modals and auxiliary verbs, and JJ for adjectives. Figure 2.6 on the following page shows a result of POS tagging an Indonesian sentence in Example 2.1 using Indonesian POS Tagger.[23]

(2.1) *Adi bisa membawa es krim-es krim itu ke rumah sakit.*
Adi can bring ice cream-RED that to hospital
"Adi can bring those ice creams to hospital." (own data)

[21] `http://bahasa.cs.ui.ac.id/postag/tagger`
[22] `http://septinalarasati.com/work/morphind/`
[23] Because Indonesian verbs and modals do not denote tense, the English translation for examples in this report uses simple present tense in default.

```
Adi      VB
bisa     MD       MD,NN    rule-11
membawa  VB
es       NN
krim-es  NN
krim     NN
itu      PR
ke       IN
rumah sakit       NN
.        Z
```

Figure 2.6: A POS-tagged Indonesian sentence in Example 2.1 on the previous page

Figure 2.6 on page 30 shows us that Indonesian POS Tagger correctly disambiguates the word *bisa* with hand-written rule-11. The word *bisa* in Indonesian is ambiguous, as a noun it means "poison" and as a modal it means "can". A compound noun *rumah sakit* "hospital" (lit. *sick house*, *rumah* means "house" and *sakit* means "sick") is correctly tagged as a common noun (NN). However, it incorrectly tags Andi, a proper noun, as a verb although there is no such verb in Indonesian and segments a reduplicated compound noun *es krim-es krim* "ice creams" incorrectly. The reason is because of the analysis result in MorphInd. The team from the University of Indonesia plans to make a better morphological analyzer than MorphInd to improve the Indonesian POS Tagger.

2.6 Summary

My PhD research focuses on building and developing a robust Indonesian grammar (INDRA) for parsing and generating Indonesian sentences in NTU-MC within the framework of HPSG and MRS using tools developed by DELPH-IN community, focusing on syntax. The Indonesian language here is the standard, "High" variety of Indonesian written in the present-day official Perfected Spelling System.

In spite of the rich literature of the Indonesian grammar, syntactic and semantic analysis in the framework of HPSG and MRS have not yet received much attention. As an open-source computational grammar, INDRA is an implementation of Indonesian HPSG and MRS. Wordnet Bahasa is employed as a lexical source and Indonesian POS Tagger is used for unknown word handling. For machine translation application, INDRA uses transfer grammars. The latest version of INDRA is stored in GitHub and can be downloaded under the MIT license.

Chapter 3

Research method

> "Grammar engineering method is spiraling upward."
> Francis Bond (p.c.)

This chapter illustrates the method employed for grammar engineering or building and developing a computational grammar. In general, the research method is a mixture of linguistic analysis and computational implementation. By doing grammar engineering, we must consider every detail of language phenomena which might not come to our mind when we document or describe a language on paper. Using the LinGO Grammar Matrix (see Section 2.2 on page 20) as a starter kit for a new grammar is convenient because we do not have to create and write the TDL files from scratch one by one although it is possible that the analyses we made in the questionnaire have to be manually edited or deleted at a later stage.

This chapter consists of two sections: Section 3.1 outlines the grammar development mechanism including a short description of how the grammar acquires the lexicon and the process of building a syntactically annotated corpus or a treebank and Section 3.2 on page 38 summarizes the grammar evaluation procedure.

3.1 Grammar development

The aim of grammar development is to parse and generate text. A sample of the text illustrating a particular language phenomenon or construction in the form of grammatical and ungrammatical sentences is selected and formatted in

```
Language: Indonesian
Language code: ind
Author: David Moeljadi
Date: 2014-12-05

#1                      #2
#word order             #word order
Source: author          Source: author
Vetted: t               Vetted: t
Judgment: g             Judgment: u
Phenomena: {wo}         Phenomena: {wo}
Saya makan kue          Saya kue makan
saya makan kue          Saya kue makan
1sg eat cake            1sg cake eat
'I eat cake(s).'        'I eat cake(s).'
#grammatical            #ungrammatical because the word order is SOV
```

Figure 3.1: A small extract of `lab3` test-suite

interlinearized glossed text according to Leipzig glossing rules[1] into one or several sample files called test-suites.[2] Test-suite can be divided into two types: phenomena based test-suite which contains particular phenomena and natural test-suite which is taken from a parallel corpus such as NTU-MC or naturally occurring text. Figure 3.1 shows a small extract of a phenomena based test-suite.

The test-suite file in Figure 3.1 consists of a header which contains the information of the language name, language code, author(s) of test-suite, the date and the source if the sentences are taken from reference grammars or web pages. Each example includes a number, optional comment(s) begin with `#`, source, vetted (`t` means checked with a native speaker, otherwise `f`), judgment (`g` for grammatical and `u` for ungrammatical), phenomenon, sentence in three layers: in standard or-

[1] `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`

[2] The general guidelines and formatting of test-suites can be checked at `http://compling.hss.ntu.edu.sg/courses/hg7021/testsuites.html`

thography, with morpheme boundaries and morpheme-by-morpheme glosses, and translation.

After analyzing the phenomenon based on reference grammars and other linguistics literatures, the analysis is modeled in HPSG and implemented by manually adding or editing some TDL files. The implementation of similar phenomenon in ERG, Jacy and Zhong may also be referred to. Afterwards, the grammar is compiled and tested by parsing sample sentences or test-suites, both the previous and the recent test-suites. The grammar is debugged if some gaps or problems are found according to the parse results until both the new and the previous phenomena are covered correctly. Sometimes the debugging process takes a long time because a new analysis has to be made and modeled in HPSG and implemented subsequently. Afterwards, the sample sentences in test-suites will be parsed again and treebanked. The test-suites can be extended if needed. This process goes repetitively (see Figure 3.2). If problems are not found or the debugging process has finished with a good result, the grammar will be updated in GitHub.

Lexical acquisition

In order to build a robust and broad-coverage grammar, the lexicon plays an important role. Increasing the amount of words in the lexicon will develop the grammar extensively. Lexical acquisition can be manually done, i.e. inputting the lexical items one by one, or semi-automatically done via linguistic resources such as wordnet and annotated corpus, i.e. extracting and mapping the lexicon to the computational grammar automatically using a computer program and checking the correctness with native speaker(s) because it is anticipated that there will be many cases in which exact matching is not possible. Preliminary work on lexical acquisition is described in Section 4.1 on page 39.

In addition, because the number of open class words (e.g. nouns) grows, it is almost impossible to know how many there are and acquire all of them to the computational grammar. Thus, the task of automatically identifying the lexical class of unknown words or unknown word handling is important to make the grammar more robust. DELPH-IN grammars such as Jacy and Zhong use ACE's

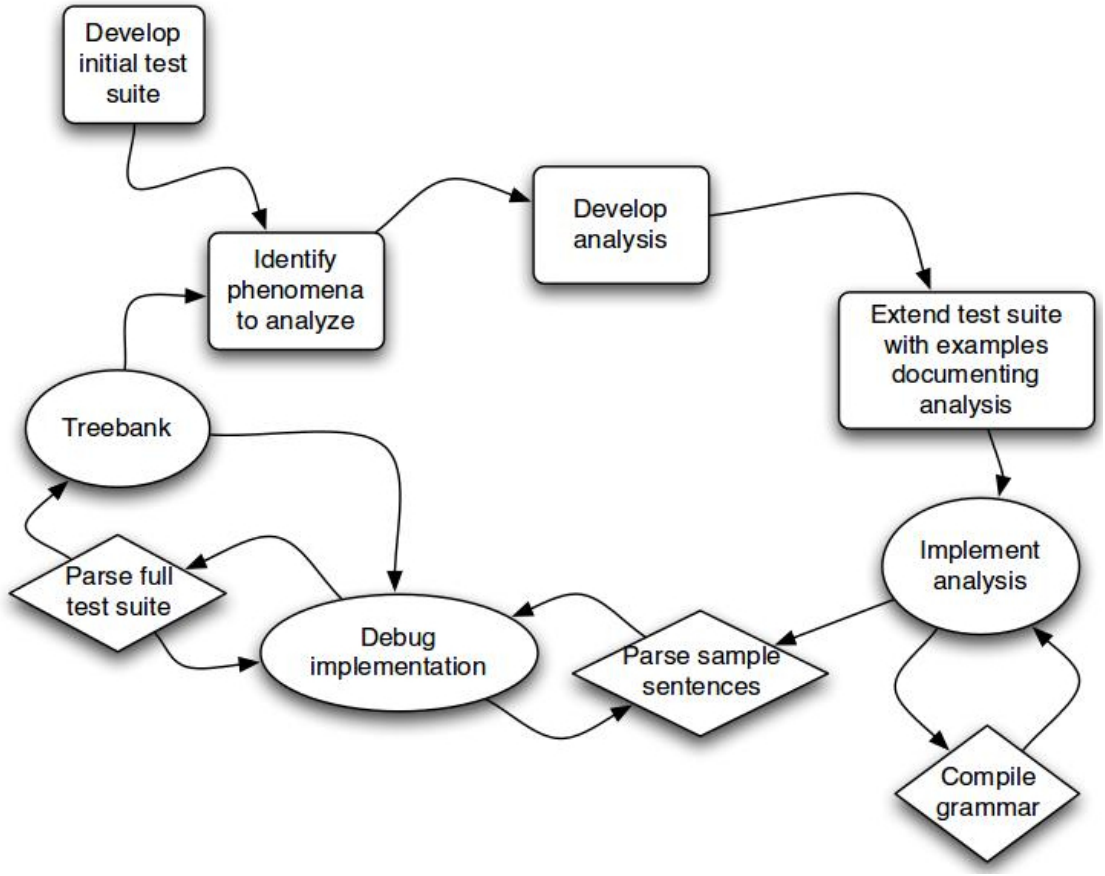


Figure 3.2: The process of grammar development (Bender et al., 2011: 10)

Figure 3.3: Screenshot of FFTB main page for `MRS` test-suite treebanking

YY-mode in order to deal with this issue (Song, 2015a,b).

Treebanking

A treebank is a syntactically annotated corpus of sentences with parse trees. Treebanking is included into the grammar development mentioned above. In my PhD work, it will be semi-automatically built using Full Forest Treebanker (FFTB). Figure 3.3 shows the FFTB page of MRS test-suite. MRS test-suite contains a representative set of sentences designed to show some of the semantic phenomena (see also Section 4.3). [3]

FFTB is a tool for treebanking with DELPH-IN grammars that allows the users to select a tree from the "full forest" of possible trees manually without listing/specifying all analyses in the parsing stage and store it into database for statistical ranking of candidate parses, transfers and translations. Using FFTB, we can note some interesting findings or linguistic analyses item by item. The left figure in Figure 3.4 shows that the grammar has two candidate trees of the

[3] `http://moin.delph-in.net/MatrixMrsTestSuite`

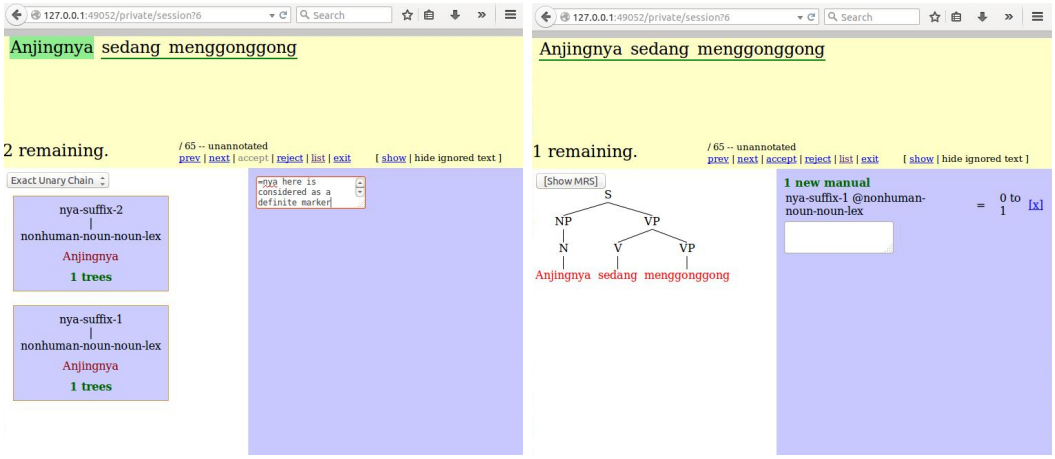


Figure 3.4: Screenhot of treebanking process of *anjingnya sedang menggonggong* "the dog is barking"

following sentence:

(3.1) *Anjingnya sedang menggonggong.*
dog=DEF PROG bark
"The dog is barking." (MRS Test-suite no. 65)

The enclitic *=nya* in the grammar has two representations: as a definite marker (*nya-suffix-1*) like "the" and as a 3SG possessive marker (*nya-suffix-2*) like "his" or "her" in English. Because the intended meaning is "the dog" not "his/her dog", *nya-suffix-1* was chosen and the intended tree was parsed and treebanked in the right figure.

FFTB is partly integrated with [incr tsdb()] and the LOGON tree (Packard, 2014). If the sentence is ungrammatical, no parse tree will be found. However, if the sentence is grammatical and no correct tree is found, all the possible trees are rejected and the grammar has to be modified or debugged (see Figure 3.2).

A multilingual corpus such as NTU-MC (Tan & Bond, 2012) can be employed to reduce the number of generated parse trees of each sentence that have to be checked since constructions which are ambiguous in one language are often not in another (Frermann & Bond, 2012).

3.2 Grammar evaluation

During the grammar development process, regression testing is important when we make some changes to the grammar. The coverage and overgeneration of the modified grammar should be compared to the gold standard.

INDRA will be regularly evaluated using a regression testing tool called [incr tsdb()] (Oepen & Flickinger, 1998). The tool supports grammar profiling which checks the parsed sentences which can be clicked to view the structures or sent to LKB for interactive parsing. It evaluates coverage and overgeneration as well as competence and performance of the grammar over different stages of development. It also allows the comparison of profiles between different versions of grammar to check the progress in terms of coverage and overgeneration as well as to detect and analyze any unanticipated degrading.

Chapter 4

Preliminary work

This chapter describes some preliminary work done to create and develop the Indonesian Resource Grammar (INDRA). A "baby" grammar of INDRA was created using the Linguistic Grammars Online (LinGO) grammar matrix in January 2014. Since its creation, the grammar components such as lexicon, grammar rules and constraints have been gradually added. Section 4.1 explains the process and the results of acquiring lexicon semi-automatically from ERG via cleaned Wordnet Bahasa. Section 4.2 mentions the grammar phenomena covered at the present state. Section 4.3 shows the evaluation.

4.1 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. This section describes my attempts to extract Indonesian verbs from Wordnet Bahasa (Nurril Hirfana Mohamed Noor et al., 2011: 258) and group them based on syntactic types in the English Resource Grammar (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. A verb frame or a subcategorization

Table 4.1: Three of six synsets of the verb *eat* and their verb frames in Wordnet

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

frame for the verb (Jurafsky & Martin, 2009) is the possible sets of complements. The assumptions are words with similar meaning have similar subcategorization frames and words with the same translations have similar meanings (Fujita & Bond, 2008). In other words, the number of arguments of verbs with similar meaning should be the same across languages.

Each verb synset in PWN (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, preceded by their frame numbers:

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 synsets and each synset may have more than one verb frames, for example the verb *eat* has six synsets with each synset having different verb frames. Three of the six synsets, together with each definition and verb frames, are presented in Table 4.1.

These verb frames can be employed as a bridge between the verb types (also

Table 4.2: The eleven most frequently used ERG verb types in the corpus

Verb type	Freq		Examples of verb
	Corp	Lex	
v_pp*_dir_le	7,079	204	go, come, hike
v_vp_seq_le	3,921	105	want, like, try
v_-_unacc_le	3,144	334	close, start, end
v_np_noarg3_le	2,723	5	make, take, give
v_-_le	2,666	486	arrive, occur, stand
v_np-pp_e_le	2,439	334	compare, know, relate
v_pp*-cp_le	2,360	154	think, add, note
v_pp_unacc_le	2,307	44	rise, fall, grow
v_np-pp_prop_le	1,861	135	base, put, locate
v_cp_prop_le	1,600	80	believe, know, find
v_np_ntr_le	1,558	10	get, want, total

verb lexical items) in ERG and those in INDRA so that automatic lexical acquisition can be done. Out of 354 verb types in ERG,[1] the top eleven most frequently used types in the `Redwoods` 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 4.2. The third, fifth and eighth type (*v_-_unacc_le*, *v_-_le* and *v_pp_unacc_le* all written in bold in Table 4.2) are regarded as the same type, i.e. intransitive verb type, in INDRA.

The first type contains verbs expressing movement or direction having optional prepositional phrase (PP) complements expressing direction, as in *B crept into the room*. The verbs in the second type are subject control verbs, the subject of which in the main clause is the same as in the subordinate verb phrase (VP) complement clause, 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 only or monotransitive verbs although they have a potential

[1] `http://compling.hss.ntu.edu.sg/ltdb/cgi/ERG_1214//ltypes.cgi`

to be ditransitive as in *B took the book*. The fifth type contains intransitive (unergative) verb as in *B arose*. The verbs in the sixth type have obligatory noun phrase (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. Table 4.3 shows the eleven verb types in ERG and their corresponding Wordnet verb frames.

The first type includes verb frame 2 and verb frame 22 since it has optional PP complements and contains verbs expressing movement or direction. The intransitive verbs include verb frame 1 in which the subject is a thing or verb frame 2 in which the subject is a human. Monotransitive verbs with obligatory NP complements, whether they can be passivized or not, include verb frame 8 in which the subject is a human or verb frame 11 in which the subject is a thing. Ditransitive verbs have verb frame 20 in which the direct NP object is a human or verb frame 21 in which the direct NP object is a thing. Verbs such as *think* and *believe* with optional or obligatory complementizer include verb frame 26 with *that* clause.

Each verb in each verb type in Table 4.2 was firstly checked 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 4.3 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 4.1. After the right synset was found, the corresponding Indonesian lemmas or translations were checked. One synset may have more than one Indonesian lemma and also may not have Indonesian lemmas at all.

The next important step is to check one by one the Indonesian lemmas belong-

Table 4.3: The eleven most frequently used ERG verb types in the corpus and their corresponding Wordnet verb frames (sb = somebody, sth = something, & = AND, || = OR)

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	1 Sth ----s ||
v_-_le	2 Sb ----s
v_pp_unacc_le	
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-	20 Sb ----s sb PP ||
pp_prop_le	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

ing 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. He gives an example that 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 corresponding verb types, a new lexicon file for INDRA was made, in which the verbs are alphabetically sorted.

A few verbs in ERG (240 verbs) could not be found in Wordnet, such as *basejump*, *bird feed*, *carpool*, *counter attack*, *defuel*, *entwist*, *flip flop*, *fuzz*, *gust*, *ice skate*, *increment*, *misfeed*, *multitask*, *roller skate*, *self insure*, *tap dance*, *unfrost* and *water ski*. 181 synsets do not have Indonesian lemmas, such as 01948659-v *balloon*, 00883635-v *gloat*, 01525295-v *malfunction*, 01977421-v *plop* and 00420549-v *tauten*. 372 verbs in Wordnet were found not having the same verb frames as in Table 4.3. For example, there are three types of *afford* in ERG, one of them belongs to the subject control verbs *v_vp_seq_le*, which is supposed to have a verb frame 28 "Somebody ----s to INFINITIVE" in Wordnet. However, among the four synsets found in Wordnet for *afford*, none of them has the verb frame 28. In total, 939 Indonesian verbs were extracted and grouped into nine verb types as presented in Table 4.4. 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. Future work is to further extract verbs, nouns, adjectives and adverbs from Wordnet Bahasa and other resources such as KBBI.

[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).

Table 4.4: New verb types and the corresponding number of verbs in INDRA

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

4.2 Grammar development

This section describes the development process of INDRA since its creation. LinGO Grammar Matrix which covers basic grammar phenomena such as word order, tense-aspect-mode (TAM), sentential negation, coordination and morphology was employed for creating the initial stage of INDRA. Phenomena which cannot be handled using the Grammar Matrix were covered by modifying the TDL files manually. These include morphophonology, definiteness, adverbs, raising and control. The grammar was evaluated via test-suites.

4.2.1 Using LinGO Grammar Matrix

INDRA was created firstly by filling in the required sections of the online page of LinGO Grammar Matrix customization questionnaire,[3] hosted and developed at the University of Washington (Bender et al., 2002; Bender et al., 2010). Table 4.5 summarizes the phenomena covered, the options chosen for Indonesian and the evaluation source from reference grammars.

As mentioned in Section 1.4 on page 11, Arka (2013a) states that there is

[3] `http://www.delph-in.net/matrix/customize/matrix.cgi`

Table 4.5: Phenomena covered by filling in LinGO Grammar Matrix questionnaire (M = Mintz (2002), L = Liaw (2004), S = Sneddon et al. (2010), A = Alwi et al. (2014), the number following each letter is the page number)

Phenomena	Chosen options for Indonesian	Source
Basic word order	SVO	L186, S265, A329
Word order of DET and nouns	Noun-DET	M104, L32, S133-134, A267
Word order of AUX and verbs	AUX-VP	M45-46, L62-68, S204-211, A163-167
Number	SG and PL	M281-285, L4, S20-22, A290-292
Person	1, 2, 3 (INCL and EXCL in 1PL)	M86-94, L29, S165, A256
Tense	FUT *akan* and underspecified	M72-73, L67, Arka (2013a)
Aspect	PRF *sudah* and PROG *sedang*	M(75, 82), L63, S204-205, A165
Sentential negation	*tidak* as an ADV modifying VP	M299, S202, A391
Coordination	monosyndeton *dan* (last coordinand is marked, e.g. “A B and C", also allows "A and B and C")	M63-65, L165-167, S346-347, A303
Yes/no question	sentence initial question word *apakah*	M262, L244, S320, A366
Lexical types	Noun subcategorization (see Figure 4.1) and verb subcategorization (see Figure 4.2)	L111-114, S(65, 72, 139), A(95-98, 288-290)
Lexicon	nouns, verbs, adjectives, auxiliaries and determiners	own data
Morphology	active *meN-* and passive *di-* voice inflection	S29

future tense in Indonesian while others such as Mintz (2002: 72-73) and Liaw (2004: 67) mention that Indonesian verbs do not denote tense, tense is indicated by temporal noun phrases or auxiliary verbs; thus, tense is underspecified. My analysis is that Indonesian denotes future tense by auxiliary verb *akan* but at the same time, tense is underspecified if there is no overt temporal marker or auxiliary *akan* (see Example 4.1 a and 4.1 b).

(4.1) (a) *Anjingnya akan menggonggong.*
dog=DEF FUT bark
"The dog will bark." (MRS Test-suite No. 392)

(b) *Anjingnya menggonggong.*
dog=DEF bark
"The dog barked/barks/will bark." (based on MRS Test-suite No. 392)

Nouns are subcategorized 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) is for inanimate nouns, *ekor* (lit. tail) for non-human animate nouns and *orang* (lit. person) for human nouns (Sneddon et al., 2010: 139; Alwi et al., 2014: 288). Figure 4.1 illustrates the noun subcategorization made using LinGO Grammar Matrix.

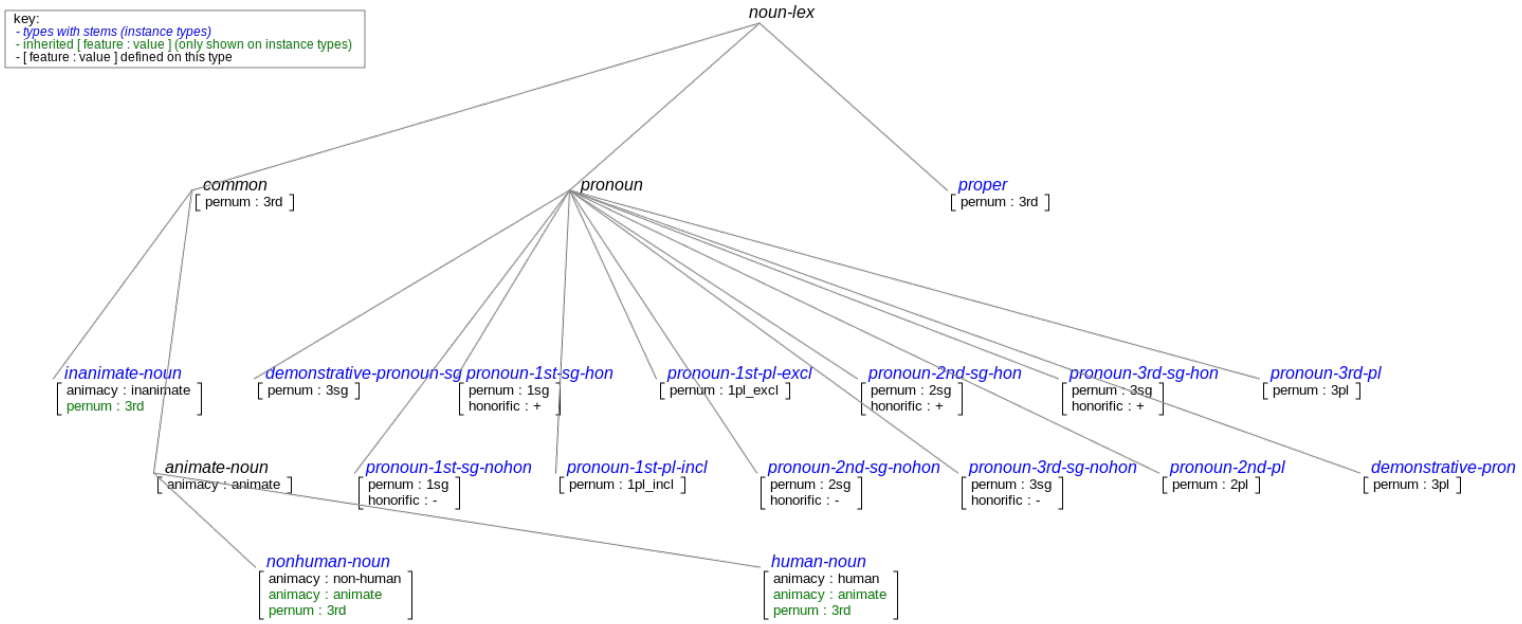


Figure 4.1: Noun hierarchy created via LinGO Grammar Matrix

Verbs are subcategorized into three groups: intransitive (*intr*) which has one argument, transitive (*tr*) which has two arguments and optional transitive (*opt-tr*) which has one obligatory subject argument and one optional object argument as in *Adi makan (nasi)* "Adi eats (rice)". The verb subcategorization here follows Liaw (2004: 111-114) and Alwi et al. (2014: 95-98).

Besides the number of arguments, the possibility of passivization with morphological inflection plays an important role in distinguishing intransitives from transitives in Indonesian. Example 4.2 and Example 4.3 a show intransitive and transitive sentence respectively. In Example 4.3 a, the verb *mengejar* is formed from an active prefix *meN-*[4] and the base *kejar* (the initial sound *k* undergoes nasalization; see Section 4 on page 57). The active prefix *meN-* is changed to a passive prefix *di-* in passive type one (Sneddon et al., 2010: 256-257) in Example 4.3 b and without affix in passive type two (Sneddon et al., 2010: 257-258) in Example 4.3 c. Sneddon et al. (2010: 256-257) state 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.

(4.2) *Adi tidur.*
Adi sleep
"Adi sleeps." (own data)

(4.3) (a) *Adi mengejar Budi.*
Adi ACT-chase Budi
"Adi chases Budi." (MRS Test-suite No. 41)

(b) *Budi dikejar Adi.*
Budi PASS-chase Adi
"Budi is chased by Adi." (based on MRS Test-suite No. 41)

(c) *Budi saya kejar.*
Budi 1SG chase
"Budi is chased by me." (own data, based on MRS Test-suite No. 41)

[4]It is not the case that all verbs having a prefix *meN-* are transitive verbs. The intransitive verb *menggonggong* "bark" has a prefix *meN-* and it cannot be passivized.

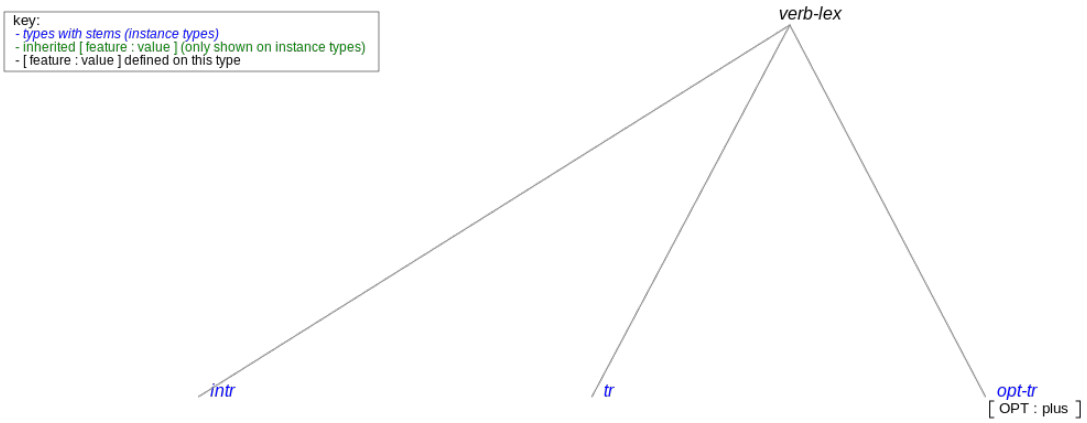


Figure 4.2: Verb hierarchy created via LinGO Grammar Matrix

Figure 4.2 illustrates the verb subcategorization made using LinGO Grammar Matrix. The more detailed subcategorization into other groups such as ditransitive was done manually (manual extension) and is mentioned in the following subsection.

The lexical items for each noun and verb subcategory were added and the morphological inflection affixes to support the active-passive voice were also included in the Grammar Matrix. However, the Matrix does not handle morphophonology as in the nasalization process of *meN-* and thus has to be manually added. Section 4 on page 57 discusses this manual extension.

Regarding the morphology part, Indonesian has many affixes for derivation but we do not deal with these because derivational morphology is computationally expensive, i.e. it needs a lot of memories. Rather than combining roots with derivational affixes using rules, all derived words or lexical items are listed up in the lexicon.

Finally, a small test-suite called `lab3`, containing 66 positive (grammatical) and 40 negative (ungrammatical) examples, was created in Lab 3 of HG7021 at NTU for each phenomenon to evaluate the behaviour of the grammar.

Table 4.6: Phenomena covered by editing TDL files

Proper names	Definiteness	Adjectives	Subj/Obj control
Pronouns	Possessive enclitics	Adverbs	Subj/Obj raising
Plural nouns	Decomposed words	Morphophonology	

4.2.2 Manual extension

After creating INDRA via Grammar Matrix, some additions and changes were done to the TDL files for phenomena mentioned in Table 4.6. Pronouns, proper names and adjectives were formerly added via Grammar Matrix and were subsequently constrained so that they cannot parse phrases such as *saya kaya* "rich I". Indonesian uses reduplication to mark plurality of nouns, e.g. *buku-buku* "books". For each type of nouns, a new lexical type which inherits the type of that noun with constraint on the number by adding a feature `PERNUM` and the value `pl` was manually added. Definiteness with enclitic *=nya* "the" and possessive enclitics *=ku* "my", *=mu* "your" and *=nya* "his/her" were added as inflectional rules.

In addition, besides the new verb types which had been acquired from ERG (see Section 4.1), more verb types and rules such as control and raising were manually added. The following table summarizes the total 49 lexical types/categories in the lexicon and the number of items within each category.

Lexical category	Explanation	Example	Number of items
inanimate-noun-noun-lex	Inanimate noun	*buku* "book"	52
inanimate-plural-noun-lex	Inanimate noun PL	*buku-buku* "books"	52
nonhuman-noun-noun-lex	Non-human animate noun	*anjing* "dog"	14
nonhuman-plural-noun-lex	Non-human animate noun PL	*anjing-anjing* "dogs"	14
human-noun-noun-lex	Human animate noun	*manusia* "human being"	9

Lexical category	Explanation	Example	Number of items
human-plural-noun-lex	Human animate noun PL	*manusia-manusia* "human beings"	9
proper-name-lex	Proper name	*Adi*	9
pronoun-1st-sg-nohon-noun-lex	1SG non-honorific	*aku* "I"	1
pronoun-1st-sg-hon-noun-lex	1SG honorific	*saya* "I"	1
pronoun-1st-pl-incl-noun-lex	1PL.INCL	*kita* "we"	1
pronoun-1st-pl-excl-noun-lex	1PL.EXCL	*kami* "we"	1
pronoun-2nd-sg-nohon-noun-lex	2SG non-honorific	*kamu* "you"	1
pronoun-2nd-sg-hon-noun-lex	2SG honorific	*Anda* "you"	1
pronoun-3rd-sg-nohon-noun-lex	3SG non-honorific	*dia* "s/he"	2
pronoun-3rd-sg-hon-noun-lex	3SG honorific	*beliau* "s/he"	1
pronoun-2nd-pl-noun-lex	2PL	*kalian* "you"	1
pronoun-3rd-pl-noun-lex	3PL	*mereka* "they"	1
n+det-lex	Decomposed word of noun and DET (see the following subsection)	*sini* "here"	5
intr-verb-lex (includes v_-_unacc_le)	Intransitive verb	*tidur* "sleep"	595
tr-verb-lex	Transitive verb	*kejar* "chase"	44
opt-tr-verb-lex	Optional transitive verb	*makan* "eat"	8
ques-comp-verb-lex	Verb with interrogative clause	*tahu* "know"	3
locative-verb-lex	Locative verb	*ada* "exist"	1

Lexical category	Explanation	Example	Number of items
prop-comp-verb-lex (includes v_pp*-cp_le)	Verb with complementizer *bahwa* "that"	*berkata* "say"	23
trans-first-arg-raising-verb	Subject raising verb	*tampak* "seem"	2
ditrans-second-arg-raising-verb	Object raising verb	*mengharapkan* "wish"	1
trans-first-arg-control-verb (includes v_vp_seq_le)	Subject control verb	*ingin* "want"	52
ditrans-second-arg-control-verb	Object control verb	*menyuruh* "order"	1
ditrans-first-arg-control-verb	Subject control verb	*berjanji* "promise"	1
v_pp*_dir_le	Motion verb with optional PP	*pergi* "go"	76
v_np_noarg3_le	Ditransitive with optional third argument	*membayarkan* "pay"	5
v_np-pp_e_le	Ditransitive with PP	*membandingkan* "compare"	41
v_np-pp_prop_le	Ditransitive with PP with state result	*meletakkan* "put"	85
v_cp_prop_le	Verb with optional complementizer	*mafhum* "understand"	53
v_np_ntr_le	Intransitive verb with obligatory complement	*berjumlah* "amount"	13
aspect_perf-aux-lex	Perfect aspect	*sudah* "already"	1
aspect_prog-aux-lex	Progressive aspect	*sedang* "in the process of"	1
modal-future-aux-lex	Future auxiliary	*akan* "will"	1
modal-aux-lex	Modal auxiliary	*bisa* "can"	4
determiner-determiner-lex	Determiner	*ini* "this"	2
adposition-lex	Adposition	*ke* "to"	10

Lexical category	Explanation	Example	Number of items
qpart-lex-item	Question word	*apakah*	1
conj-lex	Conjunction	*dan* "and"	1
propcomp-lex-item	Complementizer	*bahwa* "that"	1
adjective-lex	Adjective	*cepat* "fast"	18
neg-adv-lex	Negation	*tidak* "not"	1
adverb-pre-lex	Adverb before predicate	*sering* "often"	6
adverb-post-lex	Adverb after predicate	*pelan-pelan* "slowly"	4
adverb-lex	Adverb before or after predicate	*sekarang* "now"	5
Total			1,235

Decomposed words

There are many cases that some words such as *sini* "here" are decomposable or can be mapped to multiple elementary predicates (EPs), e.g. *sini* "here" can be thought of as *tempat ini* "this place" (Seah & Bond, 2014). 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 4.3 shows the type hierarchies in TDL file. Figure 4.4 and 4.5 show the type hierarchy for heads and demonstratives respectively.

```
demon_q_rel := quant_rel.
proximal_q_rel := demon_q_rel.
distal_q_rel := demon_q_rel.
medial_q_rel := distal_q_rel.
remote_q_rel := distal_q_rel.
```

```
entity_n_rel := generic_n_rel.
orang_n_rel := entity_n_rel.
benda_n_rel := entity_n_rel.
hal_n_rel := entity_n_rel.
time_n_rel  := generic_n_rel.
place_n_rel  := generic_n_rel.
```

Figure 4.3: Defining type hierarchies in TDL

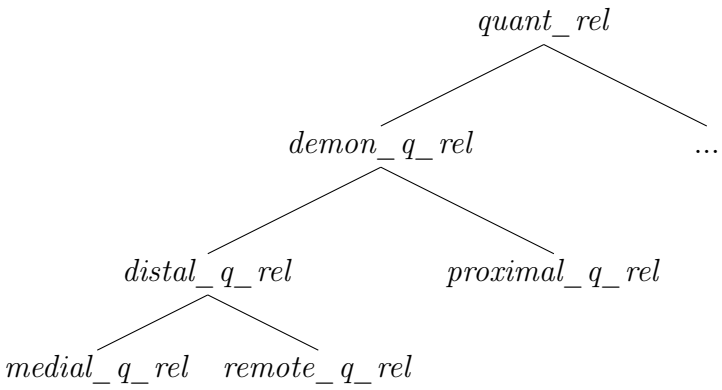


Figure 4.5: Type hierarchy for demonstratives

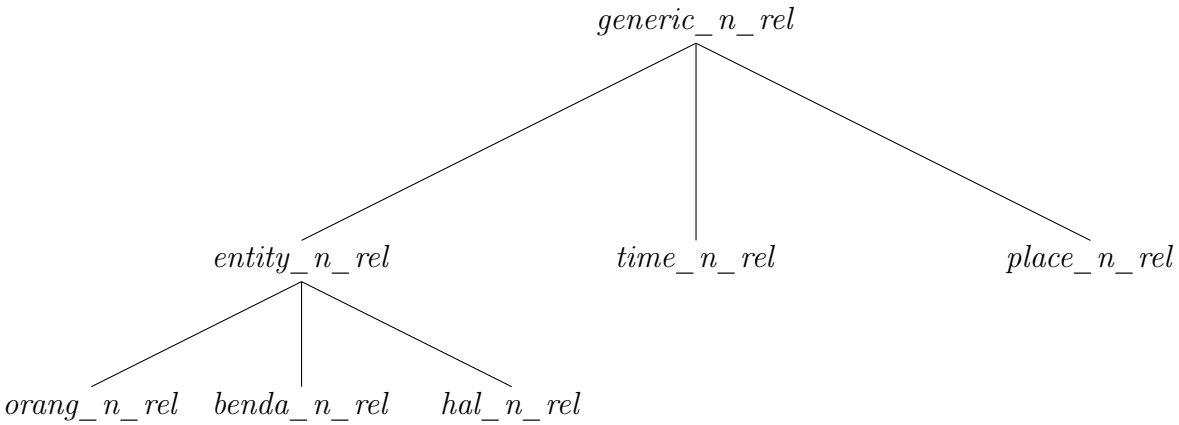


Figure 4.4: Type hierarchy for heads

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: 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 4.6 shows the implementation in TDL. The `LKEYS.KEYREL` is used for the heads and `LKEYS.ALTKEYREL` for the demonstratives.

```
itu_2 := determiner-determiner-lex &
  [ STEM < "itu" >,
    SYNSEM.LKEYS.KEYREL.PRED "distal_q_rel",
    TRAITS native_token_list ].

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

sana := n+det-lex &
  [ STEM < "sana" >,
    SYNSEM.LKEYS [KEYREL.PRED "place_n_rel",
                  ALTKEYREL.PRED "remote_q_rel" ],
    TRAITS native_token_list].
```

Figure 4.6: Decomposed predicates of the word *situ* "there" and *sana* "over there"

Figure 4.7 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 $e2$. The value of the ARG2 ($x4$) 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 ($h7$) and linked to the LBL of *place_ n_ rel* ($h7 =_q h5$). Decomposing words is important to get more refined semantics. Future work is to expand this to other heads and demonstratives such as *kini* "at present" which can be decomposed into *waktu* "time" (*time_ n_ rel*) and *ini* "this" (*proximal_ q_ rel*).

$$\begin{bmatrix} mrs & \\ \text{TOP} & \boxed{0}\ h \\ \text{INDEX} & \boxed{2}\ e \\ \text{RELS} & \left\langle \begin{bmatrix} \_di\_p\_rel & \\ \text{LBL} & \boxed{1}\ h \\ \text{ARG0} & \boxed{2}\ e \\ \text{ARG1} & \boxed{3}\ i \\ \text{ARG2} & \boxed{4}\ x \end{bmatrix}, \begin{bmatrix} place\_n\_rel & \\ \text{LBL} & \boxed{5}\ h \\ \text{ARG0} & \boxed{4}\ x \end{bmatrix}, \begin{bmatrix} medial\_q\_rel & \\ \text{LBL} & \boxed{6}\ h \\ \text{ARG0} & \boxed{4}\ x \\ \text{RSTR} & \boxed{7}\ h \\ \text{BODY} & \boxed{8}\ h \end{bmatrix} \right\rangle \\ \text{HCONS} & \left\langle \begin{bmatrix} qeq & \\ \text{HARG} & \boxed{0}\ h \\ \text{LARG} & \boxed{1}\ h \end{bmatrix}, \begin{bmatrix} qeq & \\ \text{HARG} & \boxed{7}\ h \\ \text{LARG} & \boxed{5}\ h \end{bmatrix} \right\rangle \\ \text{ICONS} & \langle\ \rangle \end{bmatrix}$$

Figure 4.7: MRS representation of *di situ* (lit. at there)

Morphophonology

A number of nasalization (sound changes) or morphophonology process occur when *meN-* combines with bases.[5] Table 4.8 shows us that a number of sound changes occur when *meN-* combines with a base. A base losts 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: 67-78) was manually examined using KBBI (Alwi et al., 2008; Sugono et al., 2014) and summarized. In addition, when the base consists of only one syllable, which usually are loanwords, *meN-* becomes *menge-* with no sound changes in the base. Every possible combination of one syllable word with *meN-* which forms a transitive verb (can be passivized) in KBBI Daring (Alwi et al., 2008) was listed up. There were 44 one syllable words in total. All possible consonant clusters and one syllable words were added to the file `irules.tdl` (see Figure 4.8).

[5]Unlike *meN-*, the passive voice prefix *di-*, however, does not undergo sound changes.

Table 4.8: Morphophonology process of *meN-* (Liaw, 2004: 74-76, 79-80; Sneddon et al., 2010: 13-18)

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

```
act-prefix :=
%prefix (p mem) (pl mempl) (pr mempr) (ps memps) (pt mempt)
(pertinggi mempertinggi) (punyai mempunyai) (b memb) (m mem)
(f memf) (v memv) (w mew) (t men) (tr mentr) (ts ments)
(d mend) (n men) (r mer) (s meny) (sl mensl) (sr mensr)
(sp mensp) (sm mensm) (sn mensn) (sk mensk) (sy mensy)
(syairkan menyairkan) (sw mensw) (st menst) (z menz) (l mel)
(c menc) (cek mengecek) (cekah mencekah) (j menj) (ny meny)
(y mey) (k meng) (kh mengkh) (kl mengkl) (kr mengkr) (g mengg)
(ng meng) (q mengq) (h mengh) (bel mengebel) (bom mengebom)
(bon mengebon) (bor mengebor) (buk mengebuk) (cas mengecas)
(cat mengecat) (cor mengecor) (dab mengedab) (dep mengedep)
(dor mengedor) (dot mengedot) (drel mengedrel) (dril mengedril)
(drop mengedrop) (dub mengedub) (dup mengedup) (kir mengekir)
(kol mengekol) (kop mengekop) (lap mengelap) (las mengelas)
(lem mengelem) (pak mengepak) (pas mengepas) (pel mengepel)
(poskan mengeposkan) (pres mengepres) (rem mengerem)
(sahkan mengesahkan) (set mengeset) (sir mengesir)
(sol mengesol) (som mengesom) (sun mengesun) (tap mengetap)
(tem mengetem) (tes mengetes) (tik mengetik) (tim mengetim)
(tip mengetip) (top mengetop) (tos mengetos) (trek mengetrek)
(tum mengetum) (a menga) (i mengi) (u mengu) (e menge) (o mengo)
act-lex-rule &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.OPT - ].
```

Figure 4.8: Inflectional rules for the active prefix *meN-*

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 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 `irules.tdl`. There are also bases such as *punyai* "have" and *syairkan* "compose a poem" (Sneddon et al., 2010: 16-17) which do not undergo the common sound changes.

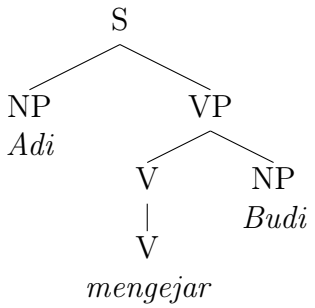


Figure 4.9: Parse tree of *Adi mengejar Budi* "Adi chases Budi"

At the present stage, this morphophonology process applies to all transitive verbs in *tr-verb-lex* 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 *act-lex-rule*. At present, INDRA can successfully parse the example 4.3 a as shown in Figure 4.9 on page 60.

The MRS representation is exactly the same as the MRS representation for transitive sentences. The value of ARG0 of the semantic head daughter *_kejar_v_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. The lexical type for proper names introduces a quant-relation type. The RSTR of each *proper_q_rel* is *qeq*-ed to the LBL of the respected *named_rel*.

Future work is to cover all the exceptions in `irules.tdl`, particularly dealing with words having *per-* and to expand *act-lex-rule* 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: 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 and translating Indonesian passive constructions into English will be a big challenge for

$$
\begin{bmatrix}
\textit{mrs} & \\
\text{TOP} & \boxed{0}\ h \\
\text{INDEX} & \boxed{2}\ e \\
\text{RELS} & \left\langle
\begin{bmatrix}
\textit{named\_rel} & \\
\text{LBL} & \boxed{4}\ h \\
\text{CARG} & \textit{"adi"} \\
\text{ARG0} & \boxed{3}\ x
\end{bmatrix},
\begin{bmatrix}
\textit{proper\_q\_rel} & \\
\text{LBL} & \boxed{6}\ h \\
\text{ARG0} & \boxed{3}\ x \\
\text{RSTR} & \boxed{7}\ h \\
\text{BODY} & \boxed{8}\ h
\end{bmatrix},
\begin{bmatrix}
\textit{\_kejar\_v\_rel} & \\
\text{LBL} & \boxed{1}\ h \\
\text{ARG0} & \boxed{2}\ e \\
\text{ARG1} & \boxed{3}\ x \\
\text{ARG2} & \boxed{9}\ x
\end{bmatrix},
\right. \\
 & \left.
\begin{bmatrix}
\textit{named\_rel} & \\
\text{LBL} & \boxed{10}\ h \\
\text{CARG} & \textit{"budi"} \\
\text{ARG0} & \boxed{9}\ x
\end{bmatrix},
\begin{bmatrix}
\textit{proper\_q\_rel} & \\
\text{LBL} & \boxed{12}\ h \\
\text{ARG0} & \boxed{9}\ x \\
\text{RSTR} & \boxed{13}\ h \\
\text{BODY} & \boxed{14}\ h
\end{bmatrix}
\right\rangle \\
\text{HCONS} & \left\langle
\begin{bmatrix}
\textit{qeq} & \\
\text{HARG} & \boxed{0}\ h \\
\text{LARG} & \boxed{1}\ h
\end{bmatrix},
\begin{bmatrix}
\textit{qeq} & \\
\text{HARG} & \boxed{7}\ h \\
\text{LARG} & \boxed{4}\ h
\end{bmatrix},
\begin{bmatrix}
\textit{qeq} & \\
\text{HARG} & \boxed{13}\ h \\
\text{LARG} & \boxed{10}\ h
\end{bmatrix}
\right\rangle \\
\text{ICONS} & \langle\ \rangle
\end{bmatrix}
$$

Figure 4.10: MRS representation of *Adi mengejar Budi* “Adi chases Budi”

machine translation task.

Associated resources

In order to make INDRA more robust, the following resources have been set up:

1. Bridging rules (Flickinger, 2014). Linguistically grounded grammars usually lack some coverage, i.e. there are sentences which the grammar fails to parse. However, partial analyses of these unparsed sentences can be useful for relation extraction and comparison of coverage. Bridging rules allow any two signs to combine into a phrase. Bridging rules have two rules: a unary bridge rule which produces a "bridge head" and a binary rule which links any two bridge heads. These rules overgenerate parse trees for ungrammatical sentences but cannot generate sentences from those trees. In the implementation, these rules have been added to ERG. The result was that ERG could successfully parse the Cathedral and the Bazaar open source essay with modest increase in cost compared to no bridging rules.

2. ACE's YY-mode for unknown word handling (Song, 2015b). As mentioned in Section 2.5.2 and Section 4.1, dealing with unknown words is very important. ACE (Packard, 2013a) has YY-mode which can parse sentences with unknown words. In order to run INDRA with the YY-mode inputs, the Indonesian POS Tagger (Rashel et al., 2014) has to be installed and set up.

3. Transfer grammar for machine translation (Packard, 2013b). The ACE system (Packard, 2013a) also provides LOGON-style transfer grammars for machine translation. At present stage, ACE can be used to translate some simple sentences such as *anjing menggonggong* (see Example 1.1 c) and *Adi mengejar Budi* (see Example 4.3 a) using the `inen` transfer grammar. Figure 4.11 shows us that the `inen` transfers the underspecified tense in INDRA into present tense and the third person nouns underspecified for number into 3PL. It also transfers the lexicon. The MRS produced by `inen` is similar to the one in ERG. Afterwards, `inen` transfers all the information into ERG

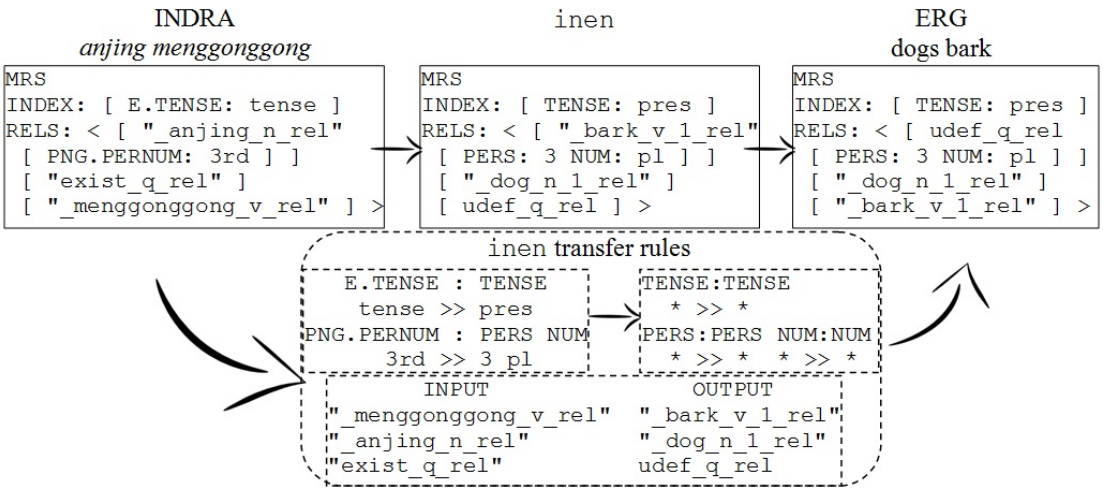


Figure 4.11: Translation process of *anjing menggonggong* using `inen`

and ERG generates the translation result.

4.3 Grammar evaluation

A test-suite containing representative set of sentences designed to show various semantic phenomena for Indonesian (MRS test-suite) was created based on the original set of 107 sentences in English (all positive test items).[6] The Indonesian MRS test-suite contains 172 sentences. The [incr tsdb()] tool (Oepen & Flickinger, 1998) is employed for grammar testing and profiling. From Table 4.9, we see that 55 out of 172 sentences can be parsed (overall coverage 32%).

This 32% coverage was got after the lexical acquisition described in Section 4.1 on page 39. Table 4.10 shows the coverage before and after lexical acquisition. The lexical acquisition has proved that by acquiring more lexical items, the grammar's coverage can be improved. Future work is to do more lexical acquisition for verbs, nouns, adjectives and adverbs. At the same time, lexical types, rules and

[6] `http://moin.delph-in.net/MatrixMrsTestSuiteIndonesian`

Table 4.9: Coverage in MRS test-suite

'ind/mrs/15-05-12/ace' Coverage Profile					
Aggregate	**total items** ♯	**positive items** ♯	**word string** ϕ	**total results** ♯	**overall coverage** %
$5 \leq$ *i-length* < 10	69	69	5.75	14	20.3
$0 \leq$ *i-length* < 5	103	103	3.27	41	39.8
Total	**172**	**172**	**4.27**	**55**	**32.0**

(generated by [incr tsdb()] at 12-may-2015 (00:24 h))

Table 4.10: Comparison of coverage in MRS test-suite before and after lexical acquisition

	Total results / total items (Coverage)
Before	52 / 172 (30.2%)
After	55 / 172 (32.0%)

constraints for new lexical items should be added.

At the present stage, INDRA can parse 0.2% of the first 400 sentences in NTU-MC as shown in Table 4.11. From my observation, most of the sentences in NTU-MC are long and has many relativizer *yang*. There are 2,253 tokens of *yang* in 2,197 sentences in the corpus. In order to achieve more coverage, phenomena which often appear such as relative clauses need to be prioritized.

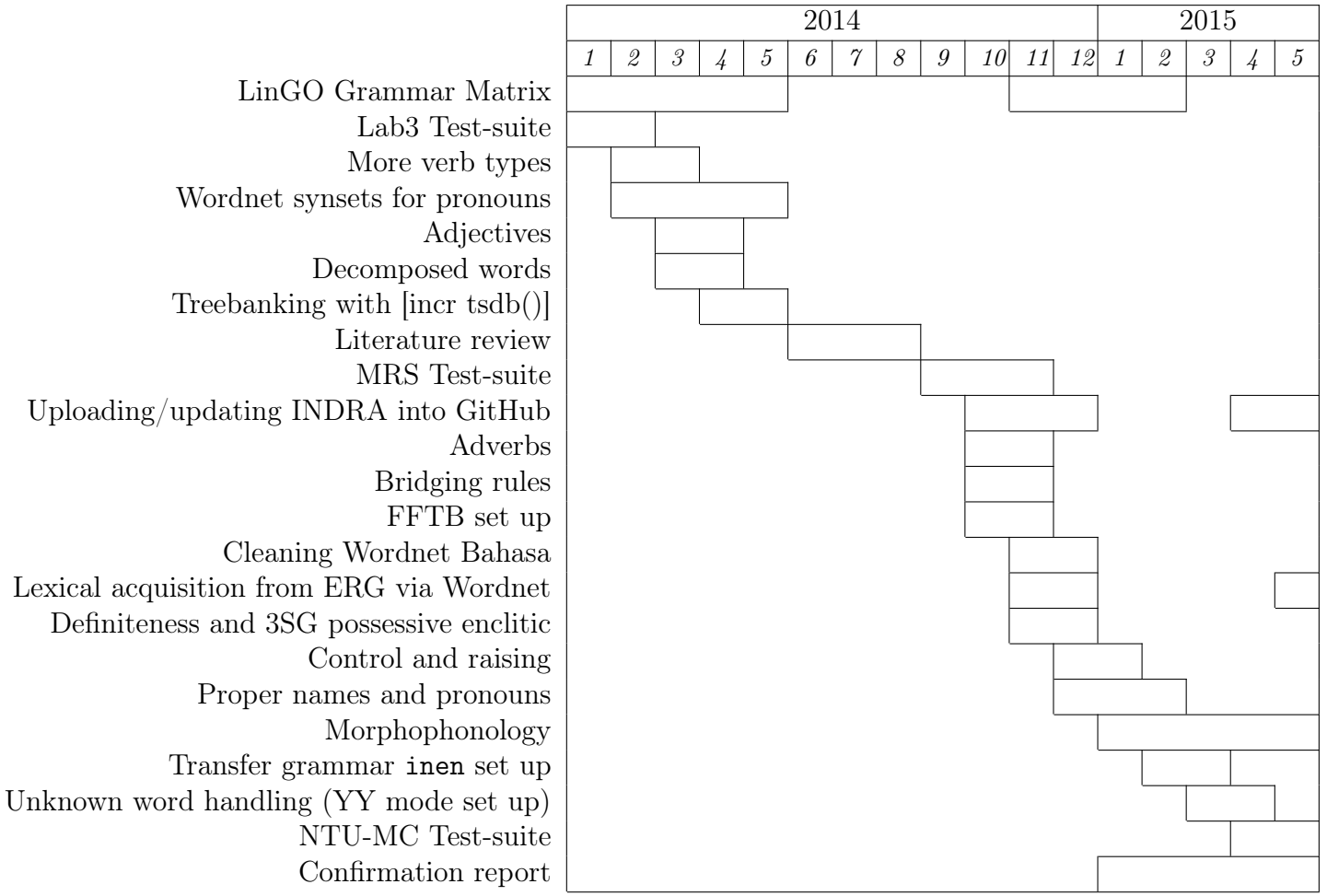
Table 4.11: Coverage of the first 400 sentences in NTU-MC

'ind/ntumc400/15-05-11/ace' Coverage Profile					
Aggregate	**total items** ♯	**positive items** ♯	**word string** ϕ	**total results** ♯	**overall coverage** %
$70 \leq$ *i-length* < 75	1	1	72.00	0	0.0
$65 \leq$ *i-length* < 70	1	1	68.00	0	0.0
$55 \leq$ *i-length* < 60	3	3	55.33	0	0.0
$50 \leq$ *i-length* < 55	4	4	51.00	0	0.0
$45 \leq$ *i-length* < 50	4	4	47.25	0	0.0
$40 \leq$ *i-length* < 45	8	8	41.62	0	0.0
$35 \leq$ *i-length* < 40	26	26	36.58	0	0.0
$30 \leq$ *i-length* < 35	35	35	32.03	0	0.0
$25 \leq$ *i-length* < 30	55	55	26.71	0	0.0
$20 \leq$ *i-length* < 25	97	97	21.88	0	0.0
$15 \leq$ *i-length* < 20	104	104	17.25	0	0.0
$10 \leq$ *i-length* < 15	48	48	12.37	0	0.0
$5 \leq$ *i-length* < 10	13	13	7.15	1	7.7
$0 \leq$ *i-length* < 5	1	1	2.00	0	0.0
Total	**400**	**400**	**22.94**	**1**	**0.2**

(generated by [incr tsdb()] at 11-may-2015 (10:33 h))

4.4 Work progress

The work progress so far is presented in the following table.



Chapter 5

Research plan

> "[E]xploring the structure of language [...] is like working on a gigantic puzzle – one so large that it could occupy many lifetimes."
> Sag et al. (2003: 9)

This chapter summarizes my PhD work for the next two and a half years.

My main contributions are as follows:

1. To the best of my knowledge, this work is the first to analyze linguistic phenomena in Indonesian within the HPSG framework.
2. This work proposes to create a robust computational grammar based on the HPSG analysis which can parse at least 60% of the Indonesian text in NTU-MC.
3. The application will be used for Indonesian-English machine translation.

My goals are as follows:

1. Acquiring lexical items and analyzing several frequent phenomena in Indonesian within HPSG framework, such as:
   (a) Relative clauses
   (b) Inflection rules and passive constructions
   (c) Copula constructions

   (d) Noun modification for noun phrases

   (e) Topic-comment constructions

2. Increasing the coverage of INDRA in NTU-MC and treebanking continuously

3. Building up Indonesian-English machine translation

My plans are as follows:

1. More literature reviews, particularly Arka's LFG papers for analysis in HPSG

2. More test-suites, particularly a test-suite containing example sentences in Sneddon et al. (2010)

3. Treebank NTU-MC

4. Improve Indonesian-English transfer grammar for machine translation

5.1 Table of contents for the proposed thesis

I refer to "A Computational Grammar for Deep Linguistic Processing of Portuguese" or "LXGram" Branco & Costa (2008), "Bulgarian Resource Grammar" or "BURGER" Osenova (2010), and "Jacy: An Implemented Grammar of Japanese" Siegel et al. (forthcoming) for the table of contents of my proposed thesis.[1] The table below shows the comparison of contents in those three grammars. The order of the contents is mainly based on Jacy. The star symbol * in column INDRA indicates the contents which have been worked on so far and will be revised and extended if needed.

[1] A draft of "Danish in Head-Driven Phrase Structure Grammar" book written by Stefan Müller and Bjarne Ørsnes (`https://hpsg.fu-berlin.de/~stefan/Pub/danish.html`) will also be referred to.

	LXGram	BURGER	Jacy	INDRA
Theoretical framework	✓		✓	✓*
Grammar engineering:				
- Development environment	✓	✓	✓	✓*
- Development cycle			✓	✓*
- Treebanking			✓	✓*
- Online documentation			✓	
Current state of the grammar	✓	✓	✓	✓*
Application of the grammar			✓	✓*
Morphology	✓			
Basic structures:				
- Core phrase structures	✓		✓	✓*
- Headedness			✓	
- Word order	✓		✓	✓*
Verbs and Adjectives:				
- Verb subcategorization		✓	✓	✓*
- Adjective subcategorization			✓	✓
- Inflectional rules	✓		✓	✓*
- Auxiliary constructions	✓	✓	✓	✓*
- Passive constructions		✓	✓	✓
- Causative constructions			✓	
- Copula constructions		✓		✓
- Imperative constructions				✓
- Topic-comment constructions				✓
Nominal structures:				
- Ordinary nouns	✓	✓	✓	✓*
- Determiners	✓	✓	✓	✓*
- Proper names	✓		✓	✓*
- Pronouns	✓	✓	✓	✓*
- Nominalizations		✓	✓	
- Temporal nouns			✓	
- Noun modification	✓		✓	✓
- Relative clause constructions	✓	✓	✓	✓
- Numeral classifiers			✓	✓
- Numbers and quantifiers	✓	✓		✓

	LXGram	BURGER	Jacy	INDRA
- Possessives	✓	✓		
- Gerunds	✓	✓		
- Noun ellipsis	✓			
- Inflectional rules	✓			
Particles		✓	✓	✓
Adverbs:				
- Adverb subcategorization			✓	✓*
- Interrogatives			✓	✓*
Demonstratives	✓		✓	✓*
Honorifics	✓		✓	
Variation between dialects	✓			
Pragmatics	✓			
Idioms	✓			

The other contents such as online documentation, headedness, honorifics and idioms may also be included. Negation will be included in each section for constructions. Phenomena which appear frequently in the NTU-MC will be analyzed in the first place.

A tentative table of contents for the proposed thesis is as follows.

1. Introduction
   (a) Theoretical framework
   (b) Grammar engineering
   (c) Current state of the grammar
2. Literature review
3. Methodology
4. Basic structures
   (a) Core phrase structures
   (b) Word order
5. Verbs

   (a) Verb subcategorization
   (b) Inflectional rules
   (c) Auxiliary constructions
   (d) Copula constructions
   (e) Topic-comment constructions
   (f) Passive constructions
   (g) Imperative constructions

6. Nouns

   (a) Common nouns and proper names
   (b) Pronouns
   (c) Determiners
   (d) Classifiers
   (e) Numbers and quantifiers
   (f) Relative clause constructions

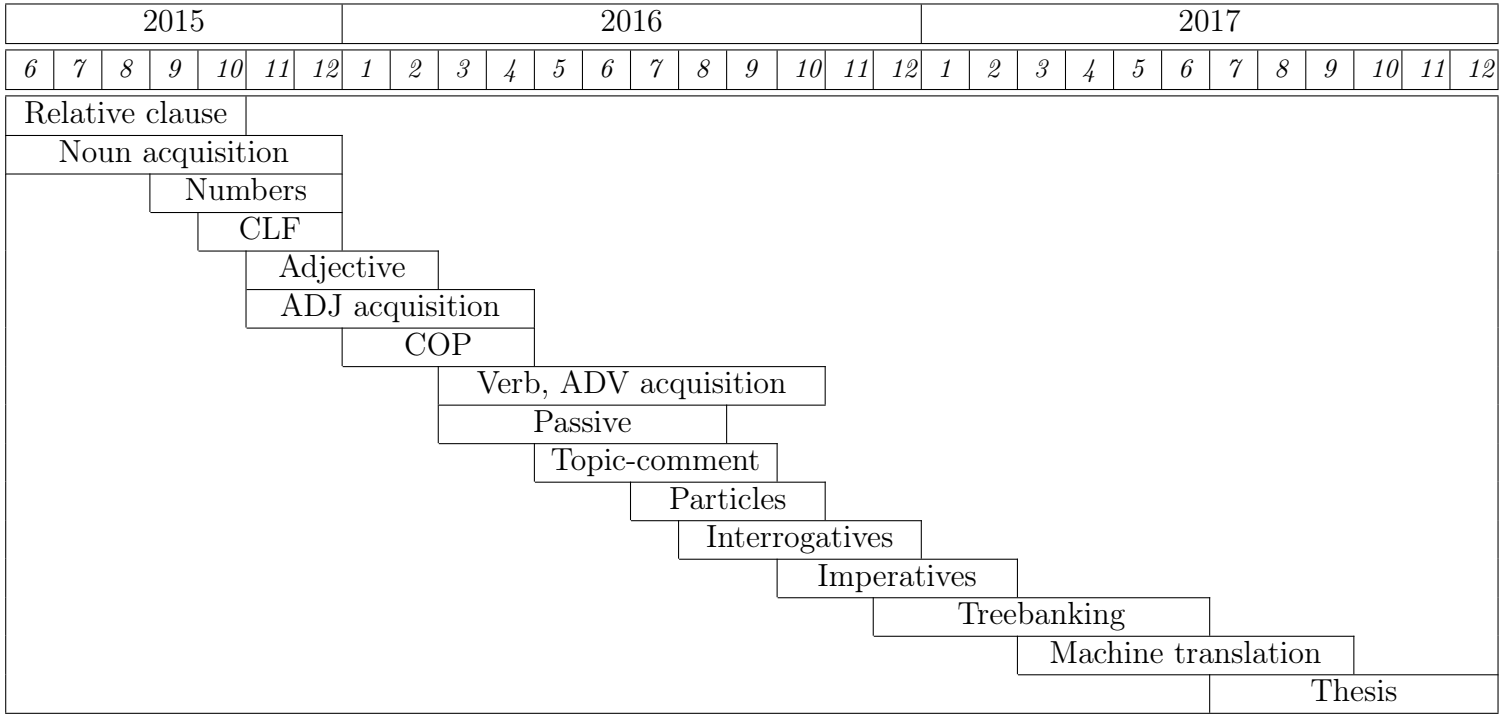
7. Adjectives and adverbs

   (a) Adjective subcategorization
   (b) Adverb subcategorization
   (c) Interrogatives

8. Treebank
9. Machine translation
10. Conclusion

5.2 Timetable

Based on the order of phenomena which will be worked on, a tentative timetable for the proposed research is as follows.



The timetable can be mainly divided into six phases:

1. June 2015 - December 2015: Nouns
2. November 2015 - April 2016: Adjectives
3. January 2016 - February 2017: Verbs
4. December 2016 - June 2017: Treebanking
5. March 2017 - September 2017: Machine translation
6. July 2017 - December 2017: Thesis writing

At the end of each phase, a draft of the corresponding chapter of the proposed thesis will be submitted.

Following the methodology in Chapter 3, for each phenomenon in nouns, adjectives and verbs:

1. thorough literature review will be conducted
2. a new test-suite particular to the phenomenon will be created
3. an analysis and the HPSG model will be made
4. the analysis will be implemented in TDL files
5. the grammar will be compiled and tested using test-suites
6. the grammar will be debugged if some problems are found
7. the sentences in test-suites will be treebanked and evaluated
8. a write-up will be submitted
9. the grammar will be updated in GitHub

Since human language is a complex system, in which its parts are intertwined, some changes may be made to the previously analyzed phenomena when problems are found.

INDRA Meta-information

(as of 17 May 2015)

Maintainer	David Moeljadi
Contributors	David Moeljadi, Sanghoun Song, Francis Bond, Michael Goodman, Dan Flickinger
Contact email	davidmoeljadi@gmail.com
Grammar name	Indonesian Resource Grammar
Short grammar name	INDRA
Language name	Indonesian
ISO code	ind
Repository	`github.com/davidmoeljadi/INDRA`
Latest revision	17 May 2015
License	MIT (`svn.delph-in.net/erg/trunk/LICENSE`)
Grammar type	Experimental grammar
Required external resources	Indonesian POS Tagger (for unknown word handling)
Lexical items	1,235 items, 939 of which were extracted from ERG via Wordnet Bahasa
Lexical rules	6
Grammar rules	20
Features	135
Types	1,596 types
Associated resources	bridging rules, transfer grammars for machine translation, unknown word handling with YY mode
Test-suites	
`matrix` test-suite	2 positive items and 1 negative item 100% coverage
`lab3` test-suite	66 positive items and 40 negative items 78.8% coverage

(as of 17 May 2015)

`mrs` test-suite	172 positive items
	32.0% coverage
`controlraising` test-suite	8 positive items and 2 negative items
	37.5% coverage
`ntumc400` test-suite	400 positive items
	0.2% coverage
Phenomena coverage	Basic word order
	Noun subcategorization
	Pronouns
	Proper names
	Definiteness
	Possessive enclitics
	Decomposed words
	Attributive adjectives
	Adverbs
	Verb subcategorization
	Tense
	Aspect
	Auxiliary construction
	Subject/object control and raising
	Sentential negation
	Coordination
	Yes/no question
	Morphophonology

List of Publications and Presentations

The following is a list of publications produced and presentations made during probationary candidature, from January 2014 to May 2015.

1. Moeljadi, David. 2014. How can we improve the Wordnet Bahasa? Presented at *Workshop/Hackathon for the Wordnet Bahasa*, Nanyang Technological University, Singapore, 26 October 2014. `http://compling.hss.ntu.edu.sg/events/2014-ws-wn-bahasa/pdf/How%20can%20we%20improve%20WNBahasa.pdf`.

2. Moeljadi, David. 2014. Usage of Indonesian Possessive Verbal Predicates: A Statistical Analysis Based on Storytelling Survey. *Tokyo University Linguistic Papers* 35: 155-176 (URI: `http://repository.dl.itc.u-tokyo.ac.jp/dspace/handle/2261/56385`).

3. Olsson, Bruno & David Moeljadi. 2014. A Parallel-Corpus Approach to sudah. Presented at *the 18th International Symposium on Malay/Indonesian Linguistics (ISMIL 18)*, Procida, Naples, Italy, 14 June 2014.

Bibliography

Abas, Husen. 1987. *Indonesian as a unifying language of wider communication: a historical and sociolinguistic perspective* (Pacific Linguistics Series D 73). Canberra.

Adelaar, Alexander. 2010. Structural Diversity in The Malayic Subgroup. In *The Austronesian languages of Asia and Madagascar*, 202–226. London and New York: Routledge Language Family Series.

Alwi, Hasan, Soenjono Dardjowidjojo, Hans Lapoliwa & Anton M. Moeliono. 2014. *Tata Bahasa Baku Bahasa Indonesia*. Jakarta: Balai Pustaka 3rd edn.

Alwi, Hasan, Dendy Sugono & Sri Sukesi Adiwimarta. 2008. *Kamus Besar Bahasa Indonesia Dalam Jaringan (KBBI Daring)*. 3rd edn. http://badanbahasa.kemdikbud.go.id/kbbi/.

Arka, I Wayan. 2000. Control theory and argument structure: explaining control into subject in Indonesian, Jakarta: The 4th International Malay and Indonesian Symposium.

Arka, I Wayan. 2010a. Categorial multifunctionality in Indonesian: linguistic and implementation issues, University of Konstanz: International ParGram Research Group Meeting.

Arka, I Wayan. 2010b. Dynamic and stative passives in Indonesian and their computational implementation, Jakarta: The 4th International MALINDO Workshop.

Arka, I Wayan. 2011. On modality and finiteness in Indonesian: complexities in =nya nominalisation. In *Proceedings of the International Workshop on TAM and Evidentiality in Indonesian Languages*, 73–89.

Arka, I Wayan. 2013a. Nonverbal predicates in Austronesian and Papuan languages: an LFG perspective, 30–46. Unud, Bali.

Arka, I Wayan. 2013b. On the typology and syntax of TAM in Indonesian. *NUSA* 55. 23–40.

Arka, I Wayan. 2014. Computational implementation of crossed control structures in Indonesian. In *Cahaya Bahasa: a festschrift in honour of Prof. Sutjaja*, 21–27. Denpasar: Swasta Nulus.

Arka, I Wayan, Mary Dalrymple, Mistica Meladel, Suriel Mofu & J. Simpson. 2009. A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In *Proceedings of the LFG 09 Conference*, Cambridge.

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

Arka, I Wayan & Mistica Meladel. 2011. Aspects of negation in Indonesian, Mountain View, California: International ParGram Meeting.

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

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The grammar matrix: an open-source starter-kit for the rapid development of cross-linguistically consistent broad- coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, 8–14. Taipei.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2011. Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis. In *Language from a Cognitive Perspective: Grammar, Usage and Processing*, 5–29. Stanford: CSLI Publications.

Bender, Emily M. & Antske Sibelle Fokkens. 2010. The LinGO Grammar Matrix: Rapid Grammar Development for Hypothesis Testing, Paris: Tutorial at HPSG 2010.

Bird, Steven, Edward Loper & Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc. http://www.nltk.org/book/.

Bond, Francis, Zhenzhen Fan & Lauren Gawne. 2015. Grammar Writing as Software Development, To be presented at Grammar Engineering Across Framework (GEAF) 2015, July 2015, Beijing, China.

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual wordnet, 1352–1362. Sofia.

Bond, Francis, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka & Shigeaki Amano. 2004. The Hinoki Treebank: A Treebank for Text Understanding. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, 158–167. Springer Verlag Lecture Notes in Computer Science.

Bond, Francis, Lian Tze Lim, Enya Kong Tang & Hammam Riza. 2014. The combined wordnet bahasa. *NUSA: Linguistic studies of languages in and around Indonesia* 57. 83–100. http://repository.tufs.ac.jp/handle/10108/79286.

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the sixth Global Wordnet Conference (GWC 2012)*, 64–71. Matsue.

Bond, Francis, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok & Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, 149–158. Sofia.

Branco, António & Francisco Costa. 2008. A Computational Grammar for Deep Linguistic Processing of Portuguese: LX Gram, version A.4.1. Tech. rep. Department of Informatics, University of Lisbon Lisbon.

Callmeier, Ulrich. 2000. PET - a platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering* 6(1). 99–107.

Copestake, Ann. 2000. Appendix: Definitions of Typed Feature Structures. *Natural Language Engineering* 6. 109–112.

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

Copestake, Ann & 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.

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

Dalrymple, Mary, Ronald M. Kaplan & Tracy Holloway King. 2004. Linguistic generalizations over descriptions. In *On-Line Proceedings of the LFG2004 Conference*, CSLI Publications. http://csli-publications.stanford.edu/LFG/9/lfg04.html.

Fellbaum, Christiane. 1998. *WordNet: an electronic lexical database*. Cambridge: MIT Press. http://wordnet.princeton.edu/man/wninput.5WN.html.

Fellbaum, Christiane. 2005. WordNet and wordnets. In *Encyclopedia of language and linguistics*, 665–670. Oxford: Elsevier 2nd edn. http://wordnet.princeton.edu.

Ferguson, Charles A. 1959. Diglossia. *Word* 15. 325–340.

Flickinger, Dan. 2000. On Building a More Efficient Grammar by Exploiting Types. *Natural Language Engineering* 6(1). 15–28.

Flickinger, Dan. 2014. Robust Parsing, Nanyang Technological University, Singapore: Grammar Engineering.

Flickinger, Dan, Yi Zhang & Valia Kordoni. 2010. Grammar Engineering for Deep Linguistic Processing: Project Seminar 2010. http://www.coli.uni-saarland.de/ yzhang/ge-ss10/lecture-01.pdf.

Fokkens, Antske, Emily M. Bender & Varya Gracheva. 2012. LinGO Grammar Matrix Customization System Documentation.

Fokkens, Antske Sibelle. 2014. *Enhancing Empirical Research for Linguistically Motivated Precision Grammars*. Saarbrücken: Department of Computational Linguistics, Universität des Saarlandes PhD dissertation.

Frermann, Lea & Francis Bond. 2012. Cross-lingual Parse Disambiguation based on Semantic Correspondence. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 125–129. Jeju Island, Korea: Association for Computational Linguistics. http://www.aclweb.org/anthology/P12-2025.

Fujita, Sanae & Francis Bond. 2008. A method of creating new valency entries. *Machine Translation* 21(1). 1–28.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank (eds.). 2013. *Glottolog 2.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org/.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and Language Processing*. New Jersey: Pearson Education, Inc. 2nd edn.

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

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

Lewis, M. Paul. 2009. *Ethnologue: Languages of the World*. Dallas, Texas: SIL International 16th edn. http://www.ethnologue.com.

Liaw, Yock Fang. 2004. *Indonesian Grammar Made Easy*. Singapore: Times Editions.

Macdonald, R. Ross. 1976. *Indonesian Reference Grammar*. Washington: Georgetown University Press 2nd edn.

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

Mintz, Malcolm W. 2002. *An Indonesian and Malay Grammar for Students*. Perth: Indonesian / Malay Texts and Resources 2nd edn.

Moeljadi, David. 2014a. How can we improve the Wordnet Bahasa?, Nanyang Technological University, Singapore: Workshop/Hackathon for the Wordnet Bahasa.

Moeljadi, David. 2014b. Usage of Indonesian Possessive Verbal Predicates: A Statistical Analysis Based on Storytelling Survey. *Tokyo University Linguistic Papers* 35. 155–176. http://repository.dl.itc.u-tokyo.ac.jp/dspace/handle/2261/56385.

Montolalu, Lucy R. & Leo Suryadinata. 2007. National Language and Nation-Building: The Case of Bahasa Indonesia. In *Language, Nation and Development in Southeast Asia*, 39–50. Singapore: Institute of Southeast Asian Studies.

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

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

Oepen, Stephan, Dan Flickinger, Kristina Toutanova & Christoper D. Manning. 2002. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria.

Oepen, Stephan & Daniel Flickinger. 1998. Towards systematic grammar profiling: Test suite technology ten years after. *Journal of Computer Speech and Language* 12(4). 411–436.

Oepen, Stephan, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén & 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*, 144–153. Skövde, Sweden.

Osenova, Petya. 2010. BUlgarian Resource Grammar – Efficient and Realistic (BURGER) Version 0.2. Tech. rep. Lingo Lab, CSLI Stanford.

Paauw, Scott H. 2009. *The Malay contact varieties of Eastern Indonesia: A typological comparison*: State University of New York at Buffalo PhD dissertation.

Packard, Woodley. 2013a. ACE, the Answer Constraint Engine. http://sweaglesw.org/linguistics/ace/.

Packard, Woodley. 2013b. AceTransfer. http://moin.delph-in.net/AceTransfer.

Packard, Woodley. 2014. FFTB: the full forest treebanker. http://moin.delph-in.net/FftbTop.

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

Riza, Hammam, Budiono & Chairil Hakim. 2010. Collaborative work on Indonesian WordNet through Asian WordNet (AWN). In *Proceedings of the 8th Workshop on Asian Language Resources*, 9–13. Beijing, China: Asian Federation for Natural Language Processing.

Sag, Ivan A., Thomas Wasow & Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. Stanford: CSLI Publications 2nd edn.

Seah, Yu Jie & Francis Bond. 2014. Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese, Reykjavik: 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation.

Siegel, Melanie & 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.

Siegel, Melanie, Emily M. Bender & Francis Bond. 2015. *Jacy: An Implemented Grammar of Japanese*. Stanford: CSLI Publications. Forthcoming.

Slayden, Glenn. 2012. *Array TFS Storage for Unification Grammars*: University of Washington Masters Thesis.

Sneddon, James Neil. 2006. *Colloquial Jakartan Indonesian*. Canberra: Pacific Linguistics.

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

Soderberg, Craig D. & Kenneth S. Olson. 2008. Indonesian. *Journal of the International Phonetic Association* 38(2). 209–213.

Song, Sanghoun. 2015a. JacyYYMode. http://moin.delph-in.net/JacyYYMode.

Song, Sanghoun. 2015b. ZhongYYMode. http://moin.delph-in.net/ZhongYYMode.

Sproat, Richard, Christer Samuelsson, Jennifer Chu-Carroll & Bob Carpenter. 2004. Computational Linguistics. In *The Handbook of Linguistics* Blackwell Handbooks in Linguistics, 608–636. Oxford: Blackwell Publishing.

Sugono, Dendy, Sugiyono & Meity Taqdir Qodratillah (eds.). 2014. *Kamus Besar Bahasa Indonesia Pusat Bahasa*. Jakarta: PT Gramedia Pustaka Utama 4th edn.

Tan, Liling & Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing* 22(4). 161–174.

Wasow, Thomas. 2004. Generative Grammar. In *The Handbook of Linguistics* Blackwell Handbooks in Linguistics, 295–318. Oxford: Blackwell Publishing.