A corpus based analysis of -kan and -i in Indonesian

Choi, Hannah Yun Jung

2019

Choi, H. Y. J. (2019). A corpus based analysis of -kan and -i in Indonesian. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/136955>

<https://doi.org/10.32657/10356/136955>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 07:21:48 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**A CORPUS BASED ANALYSIS
OF -KAN AND -I IN INDONESIAN**

**CHOI YUN JUNG
SCHOOL OF HUMANITIES
2019**

A Corpus Based Analysis of *-kan* and *-i* in Indonesian

Choi Yun Jung

School of Humanities

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Master of Arts

2019

Statement of Originality

I certify that all work submitted for this thesis is my original work. I declare that no other person's work has been used without due acknowledgement. Except where it is clearly stated that I have used some of this material elsewhere, this work has not been presented by me for assessment in any other institution or University. I certify that the data collected for this project are authentic and the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

November 14, 2019

Choi Yun Jung

Supervisor Declaration Statement

I have reviewed the content of this thesis and to the best of my knowledge, it does not contain plagiarised materials. The presentation style is also consistent with what is expected of the degree awarded. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accordance with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

A handwritten signature in black ink, reading "Francis Bond." The signature is written in a cursive, flowing style.

November 14, 2019

Assoc. Prof. Francis Bond

Authorship Attribution Statement

*(A) This thesis does not contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

November 14, 2019

Choi Yun Jung

Acknowledgement

I would first like to thank my supervisor Francis Bond. He consistently allowed this thesis to be my own work but steered me in the right the direction whenever I needed it and for that I am grateful. I'm also grateful to him for agreeing to supervise me halfway through my candidature. I also must thank Frantisek Kratochvil, my initial supervisor, who saw my passion in linguistics many years ago and took a chance on me despite my shortcomings. I am always grateful to you for helping me become the researcher I am today.

I would also like to thank the experts who were involved in advising me for this research project: Prof. Hiroki Nomoto, David Moeljadi, Luis Morgado da Costa and Le Tuan Anh. Without their passionate participation and input, this accomplishment would not have been possible. I would also like to thank the Computational Linguistics Lab and fellow graduate students during my time in NTU for making the journey a memorable one. A special thanks also goes to the staff members at the School of Humanities, especially Christina and Qiao En who are always patient, understanding and ready to help in any way they can.

To those I treasure who stuck with me to the end: Ian Crowder, Michael Yoshitaka Erlewine, Luis Morgado da Costa, Le Tuan Anh and Van, Ashley Marie Yeap, Irfan Silalahi, Arkadiusz Janz, Choi Hyun Jung and Mark Gottenborg Jensen. Every day at least one of you cheered me on from all over the world. Thank you for giving me a reality check every time I needed one, and for celebrating the little victories with me along the way. I could not have done it without you. Thank you also to those who have read and gave comments on the drafts and provided technical help: Luis Morgado da Costa, Le Tuan Anh, David Moeljadi and Arkadiusz Janz and Mark Gottenborg Jensen. Thank you for being my friends, and for reading my unpolished work without judgment. Each of you helped me in so many more ways than you could imagine.

Finally, I must express my profound gratitude to God, whose glory I do this for, my incredibly loving parents, my grandfather Manaek Sihombing, my siblings and to all my extended family for providing me with the unfailing support and continuous prayers.

Hannah Choi

Singapore, 2019

Contents

Statement of Originality	i
Supervisor Declaration Statement	ii
Authorship Attribution Statement	iii
Acknowledgement	iv
List of Abbreviations and Conventions	viii
List of Figures	ix
List of Tables	x
Abstract	xii
I Background	1
1 Introduction	2
1.1 Statement of research	5
1.2 Major Contributions of the thesis	5
1.3 Organization of the thesis	6
2 Literature Review	8
2.1 Indonesian Language	8
2.1.1 Available resources for Indonesian	9
2.1.2 Typology and Morphology of Indonesian	9
2.2 <i>Men-</i> Verbs	11
2.3 Suffix <i>-kan</i> on Verbs	13
2.4 Suffix <i>-i</i> on Verbs	16
2.5 Alternation Patterns of Verbs with <i>-kan</i> and <i>-i</i>	19
2.5.1 Displaced Theme/Locative Object Alternation	19
2.5.2 Dative Alternation	21

2.5.3	Causative/Applicative Alternation	22
2.6	Classifying Verbs in Indonesian	23
3	Research Method	25
3.1	Overview	25
3.2	Resources used in thesis	25
3.2.1	Leipzig Corpora Collection Indonesian Sub-corpus	25
3.2.2	MALINDO Morph Morphological Dictionary	27
3.3	Processing the corpus	28
3.3.1	Reading in the corpus	28
3.3.2	Identification of verbs form the corpus	28
3.3.3	Selection and Grouping of Verbs	29
3.4	Annotation of the Database	31
3.4.1	Creation of the Semantic Role List	34
3.5	Processing the Annotated Databse	36
II	Findings	38
4	Results	39
4.1	Overview	39
4.2	The Annotated Corpus	39
4.3	Syntactic Arguments Overview	40
4.4	Group Overview	41
4.5	Role Overview	45
5	Analysis	51
5.1	Overview	51
5.2	- <i>kan</i> and - <i>i</i> as Valency Changing Morphemes	51
5.3	Alternation Patterns	53
5.3.1	Displaced Theme/Locative Object Alternation	57
5.3.2	Causative/Locative Alternation	58
5.3.3	Dative Alternation	60
5.3.4	Benefactive/Applicative Suffix Alternation	62
5.3.5	Experiencer-Stimulus Alternation	63
5.3.6	Causative/Applicative Alternation	66
5.4	Comparing the results with KBBI	68

III	Summary	70
6	Conclusion	71
6.1	Overview	71
6.2	Summary	71
6.3	Newly available resources	72
6.4	Limitations and future work	73
IV	Bibliography and Appendices	74
	Bibliography	75
A	Source code	80
B	List of Verbs Annotated	81
C	Three Argument Valency Lists by Group	88

List of Abbreviations and Conventions

In this work I follow the Leipzig glossing rules presented in (<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>). The following abbreviations are used for syntactic classes, functions, and other categories:

-	affix boundary
=	clitic boundary
1	first person
2	second person
3	third person
ACT	active
CLF	classifier
COP	copula
DEF	definite
EXCL	exclusive
FUT	future
INCL	inclusive
LOC	locative
NEG	negation
PASS	passive
PL	plural
PRF	perfect
PROG	progressive
REDUP	reduplication
REL	relativizer
SG	singular

List of Figures

3.1	NTU-MC Annotation Interface	32
-----	---------------------------------------	----

List of Tables

2.1	Sound changes in inflectional rule with <i>meN-</i> from Moeljadi (2018)	11
2.2	Formation of <i>meN-</i> Verbs	12
2.3	Formation of <i>meN-</i> verbs adapted from (Arka et al., 2009)	17
2.4	Displaced Theme <i>-kan</i> and Location <i>-i</i> Object Alternation	20
2.5	Dative Alternation	21
2.6	Causative/Applicative Alternation	22
3.1	Indonesian data available on LCC	26
3.2	Data in MALINDOmorph for lemma <i>makan</i> “eat”	27
3.3	Groups and Features	30
3.4	Criteria for groups	30
3.5	No. of Roots Found per Group	30
3.6	Example SIDs from the database	31
3.7	Surface forms removed from the dataset	32
3.8	Syntactic argument role list	33
3.9	List of Affix Functions	33
3.10	Semantic Roles by Sheinfux et al. (2017)	35
3.11	Semantic Roles for the Annotation Task	37
4.1	Number of tagged sentences per group	39
4.2	Number of tagged sentences by type	40
4.3	Syntactic Arguments per Group (%)	40
4.4	Guide for Tables 4.6 – 4.10	42
4.5	All Roles by Group	43
4.6	Agent, Patient and Theme Roles by Group	45
4.7	Cause, Force, Experiencer and Stimulus Roles by Group	46
4.8	Location, Goal, Source, Recipient, Beneficiary Roles by Group	48
4.9	Instrument, Contant, Result, Attribute, Purpose Roles by Group	49
4.10	Degree, Duration, Extent, Manner Roles per Group	50
5.1	Syntactic Arguments per Group (%)	52

5.2	Location by Group	52
5.3	Most Frequent Valence for +kan, +i, Null Groups	55
5.4	Most Frequent Valence for All Groups	56
5.5	Displaced Theme <i>-kan</i> and Location <i>-i</i> Object Alternation	57
5.6	Location, Goal, Themes Role by Group	58
5.7	Causative-Locative Applicative Alternation	59
5.8	Dative Alternation	60
5.9	Benefactive/Applicative Suffix Alternation	62
5.10	Experiencer and Stimulus Roles by Group	63
5.11	Stimulus-Experiencer Alternation	64
5.12	Stimulus-Experiencer Valence	64
5.13	Causative/Applicative Alternation	66
5.14	Comparing roots with KBBI	69
B.1	List of Annotated Group 1 Verbs (NKI)	81
B.2	List of Annotated Group 2 Verbs (nKI)	83
B.3	List of Annotated Group 3 Verbs (NkI)	84
B.4	List of Annotated Group 4 Verbs (NKi)	85
B.5	List of Annotated Group 5 Verbs (nKi)	86
B.6	List of Annotated Group 6 Verbs (nkI)	86
B.7	List of Annotated Group 7 Verbs (Nki)	87
C.1	Verb Subcategorization Frames for Group NKI verbs	94
C.3	Subcategorization Frames for Group NkI verbs	96
C.3	Subcategorization Frames for Group nKI verbs	97
C.4	Subcategorization Frames for Group NKi verbs	101
C.5	Subcategorization Frames for Group nKi verbs	102
C.6	Subcategorization Frames for Group nkI verbs	102
C.7	Subcategorization Frames for Group Nki verbs	103

Abstract

The importance of capturing patterns of shared verb behavior through verb classes was called to attention by Fillmore (1970) in his seminal paper “The Grammar of Hitting and Breaking”. In his work, he recognized that verbs in English could be grouped into classes based on their semantic similarity as well as shared grammatical behavior and argument realization. Specifically, he showed how hit and break verbs are each members of larger classes of verbs whose members share comparable patterns of behavior such as participation in the causative alternation and interpretations available to their passive particles (1970:125). Other studies have since been done in English which confirmed and expanded on Fillmore’s findings (Dixon, 1992; Jackendoff, 1992; Levin & Hovav, 1991), most notably by Levin (1993) in her seminal book “English Verb Classes and Alternations”. Moving beyond English, the idea of semantically related verb classes having shared syntactic behaviors has also been identified and explored in other languages such as Lhasa Tibetan (DeLancey, 1995), Kimaragang Dusun (Kroeger, 2010) and Indonesian (Voskuil, 1996). Most recently, this idea has been implemented computationally in Hebrew by Sheinfx et al. (2017). Their study proposed an analysis that explained argument structure phenomena in Hebrew by distinguishing between semantic and syntactic selection and stating the constraints in each level separately.

Indonesian (ISO 639-3: ind), is the national language of the multilingual Indonesian archipelago. This Austronesian language is spoken by more than 22 million speakers as a first language (Lewis, 2009). As an agglutinating language, Indonesian employs several suffixes and prefixes with verbal roots (Sneddon et al., 2010). This thesis explores the use of these affixes together with argument role information as a basis of identifying verb sub-classes in Indonesian.

The data used in this thesis comes from the Indonesian sub-corpora found in the Leipzig Corpora Collection (LCC) (Quasthoff et al., 2006) which contains over 15 million sentences from news and web articles. I searched the corpus for verbs containing the prefix *meN-* and suffixes *-kan* and *-i*. I then grouped these verbs into one of seven distinct groups according to its morphological behavior. I selected 50 verb roots from each group and extracted a total of 4800 sentences for further analysis. I annotated these sentences with semantic roles and arguments based on a list adapted from Sheinfx et al. (2017).

I found that it is possible to use the morphological information of affixes to arrive at a coarse-grained sub-classification of verbs in Indonesian that confirms the findings from existing research. I also show that more fine-grained classification can be achieved using semantic

information from argument roles.

PART I:

BACKGROUND

Chapter 1

Introduction

This work explores the sub-classes of active *meN*- verbs in Standard Indonesian. I investigate this through the lens of verb morphology and verb valency information. Verb valency encodes information of the presence and number of syntactic arguments for a verb (Tesnière, 1959). Verbs with the same valence patterns can be grouped into sub-classes and classes. The idea of verb classes and sub-classes has been well studied in literature, and the importance of capturing patterns of shared verb behavior through verb classes was called to attention by Fillmore (1970) in his seminal paper “The Grammar of Hitting and Breaking”. He showed that beyond a verb’s meaning, there is lexical information encoded in the verb that motivates the arguments it can take. He compared the behaviors of the two verbs ‘hit’ and ‘break’ as seen in (1) and (2).

- (1) a. The stick *broke*.
- b. John *broke* the stick (with a rock).
- c. A rock *broke* the stick.
- d. *A rock *broke* the stick with a hammer.

- (2) a. *The tree *hit*.
- b. John *hit* the tree (with a hammer).
- c. A rock *hit* a tree.
- d. *A rock *hit* a tree with a stick.

He noted that the verb ‘break’ *requires* the affected NP and *permits* agent, instrument or both while the verb ‘hit’ *does not permit* the affected NP without either an agent or an instrument. Levin (1993, p.7) then built on and expanded on this by showing that these diathesis alternations can show the different meaning components of a verbs such that verb classes can form. She used alternations to classify the verbs in the English language and she further showed that the differences in the behavior of the verbs and classes can be explained with the alternations sensitive to particular meaning components of the verb in each class. She showed there

is a distinct pattern of behavior in the transitive verbs ‘break’, ‘hit’, ‘cut’ and ‘touch’ shown in (3) with respect to three relevant alternation patterns, namely the conative, body-part possessor ascension and middle alternations.

- (3) a. Margaret *cut* the bread.
b. Janet *broke* the vase.
c. Terry *touched* the cat.
d. Carla *hit* the door.

In particular, while the middle construction is available to ‘cut’ and ‘break’, it is not available to ‘hit’ and ‘touch’ as shown in (4):

- (4) a. The bread *cuts* easily.
b. Crystal vases *break* easily.
c. *Cats *touch* easily.
d. *Door frames *hit* easily.

The conative construction is available to ‘cut’ and ‘hit’ but not ‘break’ and ‘touch’ as shown in (5).

- (5) a. Margaret *cut* at the bread.
b. *Janet *broke* at the vase.
c. *Terry *touched* at the cat.
d. Carla *hit* at the door.

She also showed that other verbs corresponded to these verbs in showing the same pattern of behavior and forming classes (Levin, 1993, p.7). The members of each class share aspects of meaning and have syntactic and semantic properties in common within each group (Fillmore, 1970; Hale & Keyser, 1986).

Moving beyond English, verb classes are also found to be formed in the same way in other languages (A. Jones et al., 1996) and while these classes can differ from language to language, the meaning components found through these alternations are similar, suggesting some evidence that this is a cross-linguistic phenomenon.

Looking into Indonesian (ISO 639-3: ind), an agglutinative language, we see that what English shows through syntactic alternations here it is often marked using morphology. For example, the suffix *-kan* participates in the causative alternation in Indonesian as seen in (6a) and (6b).

- (6) a. *Cangkir=nya pecah.*
cup=3sg break
“The cup broke/is broken” (Cole et al., 2008, p. 123)
- b. *Janet mem-(p)ecah-*(kan) cangkir=nya.*
Janet meN.break-KAN cup=3sg
“Janet broke her cup (Janet caused the cup to break)” (Cole et al., 2008, p. 123)
- c. * *Janet mem-(p)ecah-i cangkir=nya.*
Janet meN.break-I cup=3sg
“Attempted: Janet broke her cup (Janet caused the cup to break)” (Cole et al., 2008, p. 123)

Standard Indonesian has several verbal prefixes, among which are *meN-*, *di-*, *ber-*, *per-* and *ter-*. It also has two verbal suffixes *-kan* and *-i*, with all affixes performing different functions. While most available reference grammars have described these suffixes in general (Sneddon et al., 2012; Macdonald, 1976; Alwi et al., 2014), only a handful have tried to explore their exact uses and the verbs associated with them. The most comprehensive study done on sub-categorizing *-kan* verbs was done by Cole et al. (2008) using four different constructions. Arka et al. (2009) has described the suffix *-i* in great detail from the LFG perspective. The question of which verbs or classes of verbs take which affixes, and more importantly, why a verb does or does not take an affix was raised by Vamarasi et al. (1999) and he attempted to answer it for the affixes *ke...an*. Beyond that, this question has not been addressed in great depth or breadth in literature. Technically, the information on a verb’s affinity to an affix is encoded in dictionaries since Indonesian dictionaries present each root with all known morphological forms as determined by a group of lexicographers, including all words derived from a root through the many morphological processes it can undergo. This naturally raises the question: can we use this robust morphology of Indonesian as a tool to inform our understanding of Indonesian verb classes? More specifically, can clustering verbs by their affixes help us capture the aspects of meaning that serve minimally to distinguish the verbs participating in different alternations?

To answer these questions I surveyed the biggest available corpus of written Indonesian to identify all instances of these affixes and grouped the verbs by their behaviors. I first extracted verbs in the corpus containing *meN-* verbs appearing with the suffixes *-kan* or *-i* from the Indonesian corpora of the Leipzig Corpora Collection (LCC) (Quasthoff et al., 2006) (see also Section 3.2.1). Then, I classified the verbs by their morphological behavior (i.e. their ability to take one, both or none of the affixes) into seven groups. I then created a database of sentences containing the target verbs and annotated their arguments with semantic roles using a tagging interface developed by Bond et al. (2015). The annotated database was analyzed group by group to identify verb subcategories that can be useful in characterizing the verb classes, and in the

process, detailing how these affixes function.

This information can be useful to inform our understanding of how verb classes in Indonesian work, and also provide a data point that can be incorporated into existing computational grammars of Indonesian and other natural language processing systems.

1.1 Statement of research

The aim of my research is to explore the classification of verbs affixed with *meN-* and *-kan/-i* on the basis of morphological behavior of the affixes and shared argument structures using corpus-based data. This work systematically explores verbs that take part in productive morphological processes in Indonesian to show how affixes could be used as a tool to classify verbs in a language that productively uses them.

Having outlined what this thesis attempts to do, I will briefly set out to describe what it does not attempt to do. This study is a preliminary and on-going work so it does not attempt to be exhaustive or definitive in its classification. This work only explored a small facet of verbs, i.e. its interaction with the prefix *meN-* and suffixes *-kan* and *-i* which are three out of several verbal affixes. The attempted classification system does not account for every property of every verb because a system accounting for every property of every verb would be liable to produce classes of verbs with only one member, which would not provide an insightful picture of how verbs in Indonesian work in general. The verbs in this study were chosen specifically because they had an affinity with the two suffixes *-kan* and *-i*. This design decision meant that many verbs were intentionally excluded from the study. I also did not factor in the inherent aspect of a verb, i.e. *aktionsart* in the initial classification of the verbs since they were purely classified only using its morphology. However, I am keenly aware that the relationship between *-kan* and *-i* and the lexical aspect of a verb plays an important role in verb behavior and it deserves to be further investigated in greater detail.

The corpus used to prepare the data for this study contains millions of words used by speakers of the Indonesian language across various domains and styles and, as such, is very likely to include many inaccuracies and inconsistencies. This reflects the challenges that one faces when working over large amounts of data. All attempts have been made to be as comprehensive and representative as possible in creating the subset of data used in argument role annotation.

Finally, this work doesn't attempt at generalizing a system for semantic representation of verbs but provides a starting point for discussion in research and methodology of verb subcategorization for languages with a rich affix system.

1.2 Major Contributions of the thesis

The main contributions of this thesis are as follows.

1. This work is, to the best of my knowledge, the first to use morphological information from affixes as a tool to classify Indonesian verbs into sub-classes.
2. This work produced an annotated corpus of sentences with argument role information that can be used as training data to automatically identify verb subcategorization patterns in Indonesian.
3. This work provides researchers with a database of *-kan* and *-i* sentences for further analysis and annotation. It provides a basis for a comprehensive database of sentences grouped by roots that can be used to systematically explore the functions of each affix with the ability to include other affixes such as *di-*, *ber-*, and *ter-* to expand the study.
4. This work systematically analyzed *-kan* and *-i* verbs to explain the differences between them using corpus data.

I hope that the work begun here will be a useful resource for researchers interested in the study of Indonesian and verb subcategorization and that the open source nature of my work will encourage collaboration among linguists and researchers in related fields.

1.3 Organization of the thesis

This thesis is divided into the following parts:

1. **Part I** motivates the research and provides background information on the Indonesian language, computational background, and the research method.
 - **Chapter 1** introduces the thesis and provides the statement of research as well as the contributions of the research.
 - **Chapter 2** presents the literature review. In section 2.1 I describe the Indonesian language and resources currently available on this language. I describe how Indonesian verbs work, and in section 2.6, I discuss the literature on verb sub-categorization in Indonesian.
 - **Chapter 3** describes the resources used as data in this thesis and the research method used to identify the verbs and sentences from the corpus to include in the database.
2. **Part II** presents the results from the morphological classification of verbs and analyzes the data collected from the annotated database and the verb subcategorization patterns that can be formed based on the annotated data.
 - **Chapter 4** presents the overview of the verb classes I identified on the basis of morphology. I discuss three other groupings: all verbs that can take *-kan*, all verbs

that can take *-i* and all verbs that do not take a suffix. The results from the annotated corpus for each group are presented.

- **Chapter 5** identifies the verb subcategorization patterns in the data that characterize each verb sub-class based on the argument roles annotated in the corpus. Here I analyze the functions of *-kan* and *-i* as seen through their alternation patterns.

3. **Part III** concludes the thesis with limitations and areas for further research, as well as presents the new resources made available for further studies.

- In **Chapter 6**, I first summarize the contents of this thesis.
- This is followed by section 6.4 where I discuss the limitations of this research and gives suggestions for what the annotated database can potentially be used for in future research.

Chapter 2

Literature Review

2.1 Indonesian Language

Indonesian (ISO 639-3: ind) is a Western Malayo-Polynesian language from the Austronesian language family. More than 43 million speakers use it as their first language, mainly in the Republic of Indonesia where it is the official language and a lingua franca for more than 100 ethnic groups (Suryadinata et al., 2003; Alwi et al., 2014, pp. 1-2). It is also spoken by more than 156 million people as a second language across the Malay Archipelago (Lewis, 2009). It is a close relative to Standard Malay (ISO693-3: zsm) and while some variations exist between the two languages, they are mutually intelligible, with researchers estimating that they share 80-90% similarity in their lexical inventory (Asmah, 2001; Lewis, 2009). Indonesian is diglossic, with two stable variants (Sneddon et al., 2012). The “high” variant, or Standard Indonesian, is used in the government, education, religion, and media and it is most often referred to as the written variety. It has been standardized and codified over decades through spelling reforms from 1947 (Abas, 1987), to as recent as 2015 (Ministry of Education and Culture, Institute of Language Development and Cultivation). The language is maintained by the government’s Ministry of Education and Culture through the Institute for Language Development and Cultivation (Ministry of Education and Culture). This institute publishes the republic’s official dictionary, the Great Dictionary of Indonesian or Kamus Besar Bahasa Indonesia (KBBI) and supports the development and standardization of the language. This is the variant that I have chosen to investigate in my thesis, owing to the greater abundance of data available for this variant. The spoken variant is often referred to as the “low” variant and it is not as standardized or well described as the “high” variant. According to Sneddon (2006), “Colloquial Jakartan Indonesian” spoken most widely in the capital of Jakarta is the variant acquiring the “low” variant status as it becomes more standardized through its use in social media. This variety will not be studied in this thesis. A detailed study on Colloquial Indonesian has been done by Ewing (2005).

The following sections describe the core features of Indonesian grammar, paying special attention how verbs are formed through the morphological process of affixation. This will pro-

vide the background to understanding the literature around the affix combinations that form the focus of this thesis, namely the interaction of the active voice *meN-* with the suffixes *-kan* and *-i*. I will also discuss verb classes before describing how I will answer the research questions outlined earlier in Section 1.1.

2.1.1 Available resources for Indonesian

Indonesian has been studied extensively with several reference grammars available both in Indonesian and English. The history of Old Malay from which Indonesian stems from is described by Adelaar & Prentice (1996). The first comprehensive reference grammars were written by Macdonald & Darjowidjojo (1967); Macdonald (1976). They discuss important morphological processes such as active and passive voice, affixation, reduplication and sentence patterns. Grammars and language guides were also written in by scholars such as Kridalaksana (1989) and Alwi et al. (2014) with in-depth analysis on verbs and how they can be classified based on their syntactic and morphological behaviors. Following that, Sneddon (1996); Sneddon et al. (2010, 2012) produced one of the most comprehensive and up-to-date resource grammars, which is still widely used today. These grammars provide detailed descriptions of affixes and word classes such as articles, demonstratives, common and proper nouns, pronouns, affixation processes and various sentence patterns. A full list of available resources on Indonesian written in various languages is available on Glottolog (Hammarström et al., 2013).

Besides resource grammars, digital resources for Indonesian are also available. A new resource for Indonesian has been developed by Moeljadi (2018) called INDRA, which is a computational grammar of Indonesian following the HPSG framework. It is able to parse a given Indonesian sentence and produce semantic analyses using Minimal Recursion Semantics (MRS) representations. Another resource is the republic's official dictionary, the Great Dictionary of Indonesian or KBBI (Amalia, 2016) which has a searchable interface (of Education et al., 2018) for all the lemmas and surface forms of words and their uses in Indonesian as determined by lexicographers. WordNet Bahasa (Bond et al., 2014) is a lexical database of Indonesian that links words together by their lexico-semantic relationships. It is used to semantically annotate corpora using synsets. The resources that I used in this thesis will be discussed in detail in Section 2.1.1.

In the following sections I will describe the basic structure of Indonesian and the process of verb formation.

2.1.2 Typology and Morphology of Indonesian

As mentioned, Indonesian is an Austronesian language and shares several language features with other Austronesian languages such as Tagalog (Cole et al., 2008).

Word order is highly constrained in Indonesian. Generally, the word order is Subject-Verb-Object with a Nominative-Accusative alignment pattern, similar to other Austronesian

languages. An example is shown in (7a). The voice system in Standard Indonesian, according to Cole et al. (2008), show some similarities to the Phillipine-type voice system. Each voice has been described by many researchers such as Macdonald (1976) and Sneddon et al. (2012). In the active voice, the prefix marking the verb is *meN-*. The active voice prefix *meN-* does not occur on imperative sentences as seen in (7b). In the passive, the voice marker is *di-* as seen in (7c), and the patient is the preverbal argument.

- (7) a. *dia me-makan kue itu*
 3SG ACT-eat cake that
 “He eats/is eating that cake.”
- b. *makan kue itu!*
 eat cake that
 “Eat that cake!”
- c. *kue itu di-makan=nya*
 cake that PASS-eat=3SG
 “That cake was eaten by him.”

Indonesian is also a head-initial language, where heads generally precede their dependants. The constituent order of the noun phrase shows the relation between a noun and the possessor or modifier. For example, in (8), the noun phrase *kue besar itu* ‘that big cake’ has the head *kue* preceding the adjective *besar* and demonstrative *itu*, and in *adiknya* ‘his brother’ the head *adik* precedes the possessive pronoun that modifies it.

- (8) *Andi mem-buat [kue besar itu] untuk [adik=nya]*
 Andi ACT-make cake big that for sibling=3SG.POSS
 “Andi made that big cake for his brother.”

Indonesian also uses prepositions to show the relation between noun phrases. In (8) there is a *untuk*-PP to indicate the benefactive argument for the predicate *membuat* ‘to make’.

Morphologically, Indonesian is a mildly synthetic language, with several morphemes per word. It is also agglutinative, with one grammatical category per morpheme (Larasati et al., 2011). It has a rich affix system with prefixes, suffixes, circumfixes, and to a smaller extent infixes (Sneddon et al., 2012, p. 29). Many morphological processes such as derivation and inflection make use of this affix system. When an affix combines with roots or stems, a number of morphophonemic alternations occur. The shape of a morpheme is phonologically conditioned by the initial phoneme of the root or stem and it is reflected in the orthography of the language as summarized by Moeljadi (2018) in Table 2.1 for the prefix *meN-*.

Allomorph of <i>meN-</i>	Initial orthography of the base		Example	
<i>mem-</i>	p	(lost)	<i>meN-</i> + <i>pakai</i>	→ <i>memakai</i> “use”
	pl, pr, ps, pt, b, bl, br, f, fl, fr, v	(retained)	<i>meN-</i> + <i>beli</i>	→ <i>membeli</i> “buy”
<i>men-</i>	t	(lost)	<i>meN-</i> + <i>tanam</i>	→ <i>menanam</i> “plant”
	tr, ts, d, dr, c, j, sl, sr, sy, sw, sp, st, sk, sm, sn, z	(retained)	<i>meN-</i> + <i>cari</i>	→ <i>mencari</i> “look for”
<i>meny-</i>	s	(lost)	<i>meN-</i> + <i>sewa</i>	→ <i>menyewa</i> “rent”
<i>meng-</i>	k	(lost)	<i>meN-</i> + <i>kirim</i>	→ <i>mengirim</i> “send”
	kh, kl, kr, g, gl, gr, h, q, a, i, u, e, o	(retained)	<i>meN-</i> + <i>ganti</i>	→ <i>mengganti</i> “replace”
<i>me-</i>	m, n, ny, ng, l, r, w, y	(retained)	<i>meN-</i> + <i>lempar</i>	→ <i>melempar</i> “throw”
<i>menge-</i>	(base with one syllable)		<i>meN-</i> + <i>cek</i>	→ <i>mengecek</i> “check”

Table 2.1: Sound changes in inflectional rule with *meN-* from Moeljadi (2018)

Verbs in Indonesian consist of a root with the possibility of one or more affixes with the template: (prefix)*-Root-(suffix)* (Arka et al., 2009). The voice-related prefix is on the outermost position while the innermost suffix is usually a transitivizer *-kan* or *-i*. Several suffixes and prefixes such as *meN-*, *di-*, *ber-*, *ter-*, *-i* and *-kan* are compatible with verbal roots and are often used as a basis of identifying verbs in Indonesian (Sneddon et al., 2010). However, very little is known beyond this initial identification. Unlike English where the derivational morphology is relatively limited, Indonesian has a rich morphology that can be used to inform us on the class of a verb.

For this thesis the affixes of interest are the verbal prefix *meN-* and suffixes *-kan* and *-i*. These are the most common verb forming affixes and have been well-described in literature. In the following sections, I explored each of these affixes at depth to establish our current understanding of how these affixes function.

2.2 *Men-* Verbs

According to Macdonald (1976), the prefix *meN-* forms verbs, which can either be transitive or intransitive. It is a marker for the active voice, which emphasizes the actor or subject of an action, in contrast to *di-* which emphasizes the undergoer of the action without needing to

mention the actor (i.e. passive voice). Some verbs require *meN-* to be affixed to base to produce a well-formed verb. Verbs can be formed by combining *meN-* with nominal, adjectival and verbal bases. When *meN-* is affixed to noun base, it forms verb with various meanings related to the noun base, they tend to be intransitive verbs that only take one argument, typically the actor. When *meN-* is affixed to adjectival bases, it forms verbs with the meaning 'to become [base]' and they tend to be intransitive, with the single argument being the undergoer of the action. When *meN-* is affixed to a verb base, the resulting verb can either be transitive or intransitive.

While this lies outside of the scope of the thesis, according to Sneddon (1996), *meN-* may frequently be dropped in colloquial use or in speech, and also in common verbs such as *drink* 'minum', *makan* 'eat' and *mohon* 'request', although the affixed forms exist and can be used.

Class	Base	gloss	Derived Verb	gloss
Noun	<i>tari</i>	'dance'	<i>menari</i>	'to dance'
	<i>darat</i>	'land'	<i>mendarat</i>	'to land'
	<i>jerit</i>	'shout'	<i>menjerit</i>	'to scream'
Adjective	<i>hangat</i>	'hot'	<i>menghangat</i>	'to (become) hot'
	<i>merah</i>	'red'	<i>memerah</i>	'to (become) red'
	<i>tipis</i>	'thin'	<i>menipis</i>	'to (become) thin'
	<i>gelap</i>	'dark'	<i>menggelap</i>	'to (become) dark'
	<i>turun</i>	'descend'	<i>menurun</i>	'to descend'
Verb (intransitive)	<i>datang</i>	'come'	<i>mendatang</i>	'to rise (up)'
	<i>jatuh</i>	'fall'	<i>menjatuh</i>	'to fall'
	<i>lihat</i>	'hot'	<i>melihat</i>	'to see'
Verb (transitive)	<i>buka</i>	'open'	<i>membuka</i>	'to open'
	<i>tunggu</i>	'wait'	<i>menunggu</i>	'to wait'
	<i>potong</i>	'cut'	<i>memotong</i>	'to cut'

Table 2.2: Formation of *meN-* Verbs

Reduplication can also occur on bases with *meN-*. There are two ways this affixation can occur. Firstly, the prefix can be attached to the first member of the reduplicated base e.g. *tari* 'dance', *menari-nari* 'to cavort' and *tingkat* 'level', *meningkat-ningkat* 'to increase (levels)'. The resulting verb suggests an iterative action. Secondly, the prefix can be attached to the second member of the reduplicated base with the resulting verb having a reciprocal meaning for some bases e.g. *kenal* 'know', *kenal-mengenal* 'to get acquainted (with each other)' and *tolong* 'help', *tolong-menolong* 'help (each other)' but not in others e.g. *bahu* 'shouler', *bahu-membahu* 'to work shoulder to shoulder' and *sedang* 'sufficient', *sedang-menyedang* 'barely sufficient'.

meN- verbs can combine with the verbal suffixes *-kan* and *-i*. In Section 2.3 I will explore how *-kan* interacts with *meN-* verbs, followed by the how *-i* interacts with *meN-* verbs in Section 2.4. I will then discuss the alternation patterns that these verb types are known to participate in.

2.3 Suffix *-kan* on Verbs

Numerous works have been written on the *-kan* suffix from different perspectives, from the simple broad descriptions found in grammars (Sneddon, 1996; Alwi et al., 2014; Macdonald, 1976) to in-depth analyses of the uses and morphosyntactic properties of these affixes (Cole et al., 2008; Voskuil, 1996). Most researchers generally agree on the basic functions of *-kan* although it is still a source of considerable confusion (Kroeger, 2010).

Kridalaksana (1989) described the function of *-kan* as a transitivizer in terms of derivational rules that turn roots into verbs. Firstly, *-kan* can be affixed to nouns, verbs, and adjectives to causativize it (i.e. "cause to be [base]"). Some examples can be seen in (9), (10) and (11) below.

(9) Verb → Verb

- a. *Pesawat itu terbang.*
airplane that fly

"That plane is flying."

- b. *Pilot itu men-(t)erbang-*(kan) pesawat model mutakhir buatan Amerika.*
Pilot that ACT-fly-KAN airplane model latest made.in America

"That pilot flew the newest American-made model airplane (Lit: The pilot caused the newest American-made model airplane to fly)." (Kridalaksana, 1989, p. 53)

(10) Noun → Verb

- a. *Dia adalah se-orang raja.*
3SG COP one-CLF king

"He is a king."

- b. *Penduduk primitif itu me-raja-kan dokter yang sedang berpraktek di daerah mereka.*
resident primitive that ACT-king-KAN doctor REL PROG practice LOC
area 3PL

"The primitive residents made the doctor who was practicing in the area a king." (Kridalaksana, 1989, p. 53)

Cole et al. (2008, pp.123) adds that *-kan* as a causativizer introduces a new syntactic argument. Adjectives such as *putih* "white" seen in (11) and inchoative verbs such as *pecah* "break" (6) and *tewas* "die" (13) are causativized by the addition of a causer. The types of base predicates that can be causativized with *-kan* are stative predicates such as *bersih* "clean" and *lebar*

“wide”, change-of-state predicates such as *jatuh* “fall” and psychological predicates such as *bosan* “bored” and *senang* “glad” (Sneddon, 1996).

(11) Adjective → Verb

- a. *Wajah=nya putih.*
face=3SG white

“His face is white.” (Cole et al., 2008, p. 123)

- b. *Ia mem-(p)utih-*(kan) wajah=nya.*
3SG ACT-white-KAN face=3SG

“He whitened his face.” (Cole et al., 2008, p. 123)

- (12) a. *Cangkir=nya pecah.*
cup=3SG break

“The cup broke/is broken.” (Cole et al., 2008, p. 123)

- b. *Janet mem-(p)ecah-*(kan) cangkir=nya.*
Janet ACT-break-KAN cup=3SG

“Janet broke her cup.” (Cole et al., 2008, p. 123)

- (13) a. *Banyak orang tewas.*
many people dead

“Many people died/are dead.” (Cole et al., 2008, p. 123)

- b. *Kecelakaan itu men-(t)ewas-*(kan) banyak orang.*
accident that ACT-dead-KAN many people

“The accident killed many people.” (Cole et al., 2008, p. 123)

Secondly, *-kan* introduces a new benefactive argument to transitive verb bases (i.e. “do [base] for someone”). According to Kridalaksana (1989), the roots that can be transitivized to introduce a benefactive argument are nominal and verbal. For instance, the beneficiary in (14a) that is expressed as an optional adjunct phrase can become a benefactive argument when *-kan* is affixed to the base as in (14b).

- (14) a. *Tika mem-(p)anggang roti itu (untuk Eric).*
 Tika ACT-bake bread that for Eric
 “Tika baked the bread (for Eric).” (Cole et al., 2008, p. 124)
- b. *Tika mem-(p)anggang-*(kan) Eric roti itu.*
 Tika ACT-bake-KAN Eric bread that
 “Tika baked Eric the bread.” (Cole et al., 2008, p. 124)

However it is not always the case that the suffix *-kan* increases the valency of the verb. In some cases, the suffix is optional when the sentence already has a goal-PP as seen in (15).

- (15) a. *Dia meng-ikat tali itu.*
 3SG ACT-tie rope that
 “He tied the rope.” (Sneddon, 1996)
- b. *Dia meng-ikat-(kan) tali itu ke anjing.*
 3SG ACT-tie-KAN rope that to dog
 “He tied the rope to the dog.” (Sneddon, 1996)

To add to the complication, some transitive verbs necessarily take the suffix *-kan* in order to form a well-formed verb. There is no benefactive or causative meaning added with the affix as seen in examples (16) and (17). In these examples the bare *meN-* verbs are ungrammatical without *-kan*. Such cases are often explained as a result of the preposition *akan* becoming part of the verb (Sneddon et al., 2010) i.e. *runding akan* ‘to discuss of/about’ → *rundingkan* ‘to discuss’.

- (16) a. *Ia me-runding-kan rencana baru.*
 3SG ACT-discuss-KAN plan new
 “He discussed a new plan.” (Cole & Son, 2004, p.345)
- b. **Ia me-runding rencana baru.*
 3SG ACT-discuss rope that
 “He discussed a new plan.” (Cole & Son, 2004, p.345)
- (17) a. *Dia tidak mem-(p)ikir-kan saya.*
 3SG NEG ACT-discuss-KAN plan
 “She does not think about me.” (Cole & Son, 2004, p.345)

- b. **Dia tidak mem-(p)ikir saya.*

3SG NEG ACT-discuss plan

“He does not think about me.” (Cole & Son, 2004, p.345)

The suffix *-kan* also plays a similar role for some verbs in ditransitive constructions where it licenses a direct object without a causative or benefactive meaning as seen in (18a). This is in contrast to the applicative reading for the same verb with *-i* seen in (18b).

- (18) a. *Hasan meng-(k)irim-kan uang kepada saya.*

3SG ACT-send-KAN money to 1SG

“Hasan sent money to me.” (Cole & Son, 2004, p.346)

- b. *Hasan meng-(k)irim-i saya uang.*

3SG ACT-discuss-APPL plan new

“Hasan sent me money.” (Cole & Son, 2004, p.347)

Cole & Son (2004) notes that *-kan* gives the appearance of being a homonymous form with multiple functions. It has a high degree of complexity in its distribution and it is widely used in the language. Linguistically it is an intriguing morpheme because it does not behave like a prototypical causative nor a prototypical applicative. What is clear is that *-kan* has the ability to change the argument structure of a verb.

2.4 Suffix *-i* on Verbs

The discussion of *-kan* almost always leads to a passing mention of *-i*. It is often discussed as an obligatory point of comparison because of their seemingly similar distribution and use. However, it has not been studied at the quite the same depth as *-kan*. Reference grammars such as Sneddon (1996) describe *-i* as having two major functions: to indicate that the object is the location of the action (i.e. locative *-i*) and to indicate repeated action (i.e. iterative *-i*). While this is true in the basic sense, *-i* is more complex, as noted by the most current most comprehensive study done by Arka et al. (2009) based on the Lexical Function Grammar (LFG) theory (Kaplan & Bresnan, 1982). Similar to *-kan*, *-i* can be affixed to nouns, adjectives and transitive and intransitive verbs as seen in Table 2.3.

When *-i* is affixed to a stem, syntactically it may or may not change the valency of the verb. With some stems there can be no change in valency. Instead there is a change in aspect such as intensity or iterativeness as seen in (19) and (20).

- (19) a. *Ia mem-(p)ukul saya.*
 3SG ACT-hit 1SG
 “S/he hit me.” (Arka et al., 2009)
- b. *Ia mem-(p)ukul-i saya.*
 3SG ACT-hit-I 1SG
 “S/he is hitting/was hitting me.” (Arka et al., 2009)
- (20) a. *Ia me-(p)egang pencuri itu.*
 3SG ACT-hold thief that
 “S/he held the thief.” (Arka et al., 2009)
- b. *Ia me-(p)egang-i pencuri itu.*
 3SG ACT-hold-I thief that
 “S/he is holding the thief tightly.” (Arka et al., 2009)

Class	Base	gloss	Derived Verb	gloss
Noun	<i>kulit</i>	‘skin’	<i>mengulit</i>	‘to skin’
	<i>gula</i>	‘sugar’	<i>menggulai</i>	‘to put sugar in’
	<i>ketua</i>	‘chair/head of an or- ganization’	<i>mengetuai</i>	‘to chair or head an organization’
Adjective	<i>panas</i>	‘hot’	<i>memanasi</i>	‘to heat’
	<i>basah</i>	‘wet’	<i>membasahi</i>	‘to dampen’
	<i>lengkap</i>	‘complete’	<i>melengkapi</i>	‘to complete’
Verb (intrans.)	<i>lompat</i>	‘jump’	<i>melompati</i>	‘to jump over’
	<i>tulis</i>	‘write’	<i>menulisi</i>	‘to write on’
	<i>duduk</i>	‘sit’	<i>menduduki</i>	‘to sit on’
Verb (trans.)	<i>kirim</i>	‘send’	<i>mengirim</i>	‘to send’
	<i>cium</i>	‘open’	<i>menciumi</i>	‘to kiss repeatedly’
	<i>siram</i>	‘wait’	<i>menyirami</i>	‘to spray with’

Table 2.3: Formation of meN- verbs adapted from (Arka et al., 2009)

If it does change the valency of the verb it is usually to increase the valence of the verb from an intransitive to a transitive and a transitive to a ditransitive by promoting an oblique argument to object as seen in (21) and (22) respectively. Generally, the argument that is incorporated as a direct object is a locative. This is usually referred to the applicative *-i* by Kroeger (2007); Arka et al. (2009); Sneddon et al. (2012)

- (21) Intransitive → Transitive

- a. *Ia duduk di kursi itu.*
3SG sit LOC chair that

“S/he is sitting/sat on the chair.” (Arka et al., 2009)

- b. *Ia men-duduk-i kursi itu.*
3SG ACT-sit-I chair that

“S/he is/was sitting on the chair.” (Arka et al., 2009)

(22) Transitive → Ditransitive

- a. *Ayah meng-(k)irim uang kepada dia.*
Father ACT-send money to 3SG

“Father sent money to him.” (Arka et al., 2009)

- b. *Ayah meng-(k)irim-i dia uang.*
Father ACT-send-I 3SG money

“Father sent him/her money.” (Arka et al., 2009)

Some stems can also produce a causative reading when combined with the suffix *-i* which is one point of confusion because this morpheme seemingly shares the functional load with *-kan* in this respect. For example, the root *takut* can have both the applicative form and causative form as seen in (23a) and (23b) with the contrasting causative *-kan* shown in (23c). However, some roots do not form causatives with *-i*, such as *kutuk* ‘curse’ → *mengutuki* ‘curse someone/something’, **mengutukan* ‘intended: cause someone to be cursed’.

- (23) a. *Sebut 1 hewan yang kamu takut-i.*
mention one animal REL 2SG afraid-I

“Name one animal that you are afraid of.” (Arka et al., 2009)

- b. *Jikalau tiada yang men-(t)akut-i mereka.*
if NEG REL ACT-afraid-I 3PL

“If there is no one thing that makes them afraid.” (Arka et al., 2009)

- c. *Dia meng-guna-kan tikus yang terperangkap dalam toples yang di-panas-kan untuk men-(t)akut-kan=nya.*
3SG use mouse REL trapped in jar REL PASS-heat-KAN
for ACT-afraid-KAN=3SG

“He used a mouse that was trapped in a jar that was heated to scare it.” [Wiki.2016.a]

Generally, the suffix *-i* is relatively more straightforward than *-kan*. *-i* can be affixed to nouns, adjectives and transitive and intransitive verbs. It has several functions, primarily as a valency changing applicative suffix that promotes an oblique argument to the core. The oblique that is often promoted is the Location. It can also change the valency of a root by functioning as a causative suffix. Lastly, if it does not change the valency of the verb, it adds an aspectual meaning to the verb, usually to indicate repetition or intensity of the event.

What kinds of roots take what affixes and why is not yet well described in the literature. Arka et al. (2009) mentions some patterns with roots that allow *-i*. Firstly, some roots only allow the applicative reading and not causative. In order to get the causative reading, the root needs to combine with *-kan*, for example *menjatuhi* ‘to drop’, not *‘make X drop’. Secondly, with nominal roots that are “conceptualized as displaced themes” tend not to have *-kan* forms attested. The object in these verbs typically takes the location as the object. Finally, some *-i* roots can appear as both a causative and an applicative, although *-kan* is preferred.

2.5 Alternation Patterns of Verbs with *-kan* and *-i*

As seen in the sections above, a root is able to take both/either *-kan* and *-i*, or in some cases none at all. What is evident is that in Indonesian, the combination of root + affix can alter the argument structure of verbs (i.e. an alternation pattern). Cross-linguistically, languages employ one or more of these alternation processes in different ways. Some examples of these alternation processes include the passive, antipassive, causative, dative, locative, conative, middle and applicative alternations (Baker, 1988).

Studies have explored these *-kan/-i* alternation patterns in Indonesian such as by Arka et al. (2009). In this section I list and summarize the most well-known alternation patterns for *-kan* and *-i*.

There are three major alternation patterns frequently mentioned in the literature: the displaced theme/locative alternation (Section 2.5.1), dative alternation (Section 2.5.2), and causative/applicative alternation (Section 2.5.3). These will be described in detail below.

2.5.1 Displaced Theme/Locative Object Alternation

In this alternation, both *-kan* and *-i* take a transitive verb and add a third argument to form a ditransitive verb. The preverbal argument is an Agent or doer of the action, but verbs affixed with *-kan* have what Kroeger (2007) calls displaced theme/patient as the direct object and Goal as the second object while verbs with *-i* have a Locative/Goal as the direct object and the Theme as the second object summarized in Table 2.4 (Arka et al., 2009). This is similar to the spray/load alternation in English (Levin, 1993).

Examples (24), (25) and (26) illustrate this alternation pattern. In each of these cases, the direct object of the *-kan* verbs are unmarked noun phrases indicating the Theme while the direct

	ARG1	ARG2	ARG3
<i>meN-ROOT-kan</i>	AGENT NP	THEME NP	GOAL <i>ke-PP</i>
<i>meN-ROOT-i</i>	AGENT NP	LOCATION/GOAL NP	THEME <i>dengan-PP</i>

Table 2.4: Displaced Theme *-kan* and Location *-i* Object Alternation

objects of the *-i* verbs are marked with prepositional phrases indicating where the action is directed towards (i.e. Goal).

- (24) a. *Dia meny-(s)iram bunga di halaman depan.*
3SG ACT-spray flower LOC yard front

“S/he is spraying the flowers in the front yard.”

- b. *Dia meny-(s)iram-kan air ke bunga.*
3SG ACT-spray-KAN water to flower

“S/he is spraying water to the flowers.”

- c. *Dia meny-(s)iram-i bunga dengan air.*
3SG ACT-spray-I flower with water

“S/he is spraying the flowers with water.”

- (25) a. *Anak-anak men-(t)empel poster di dalam kelas.*
child-REDUP ACT-stick poster LOC in class

“The children are sticking posters in the class.” (Arka et al., 2009)

- b. *Anak-anak men-(t)empel-kan poster ke tembok.*
child-REDUP ACT-stick-KAN poster to wall

“The children stuck a poster on the wall (the children caused the poster to stick to the wall).” (Arka et al., 2009)

- c. *Anak-anak men-(t)empel-i tembok dengan poster.*
child-REDUP ACT-stick-I wall with poster

“The children are sticking posters all over the walls.” (Arka et al., 2009)

- (26) a. *Air itu sedang meng-alir ke sawah.*
 water that PROG ACT-flow to rice.field
 “Water is flowing to the rice field.” ((Arka et al., 2009) from (Kroeger, 2007))
- b. *Dia meng-alir-kan air itu ke sawah=nya.*
 3SG ACT-flow-KAN water that to rice.field=3SG.POSS
 “S/he caused the water to flow to his/her rice field.” ((Arka et al., 2009) from (Kroeger, 2007))
- c. *Dia meng-alir-i sawah=nya dengan air itu.*
 3SG ACT-flow-I rice.field=3SG.POSS with water that
 “S/he flooded his/her rice field with water.” ((Arka et al., 2009) from (Kroeger, 2007))

This alternation is also characterized by the causative function of *-kan* contrasting with the applicative function of *-i*. In each of the examples, *-kan* functions as a causative, producing verbs meaning “cause to X” (i.e. *siram* ‘to spray’ → *menyiramkan* ‘cause to spray’, *tempel* ‘to stick’ → *menempelkan* ‘cause to stick’, *alir* ‘to flow’ → *mengalirkan* ‘to cause to flow’) while *-i* is an applicative, turning the oblique argument of the bare verb in (24a), (25a) and (26a) which is marked by preposition phrases into the direct object of the verb marked by a noun phrase.

2.5.2 Dative Alternation

The dative alternation has been discussed by Arka et al. (2009) and Nur Amirah Binte Khairul Anuar et al. (2017) and it contrasts *-kan* and *-i* verbs in a similar way to English. In this alternation, *-kan* verbs take the role Source, Recipient and Theme as its arguments while *-i* verbs take the roles Source, Theme and Recipient (Nur Amirah Binte Khairul Anuar et al., 2017) as summarized in Table 2.5.

	ARG1	ARG2	ARG3
<i>meN-ROOT-kan</i>	SOURCE NP	THEME NP	RECIPIENT <i>ke/kapada/pada-PP</i>
<i>meN-ROOT-i</i>	SOURCE NP	RECIPIENT NP	THEME NP

Table 2.5: Dative Alternation

An example of a verb that participates in this alternation is *meminjami/meminjamkan* ‘to lend’ as seen in (27). Here *-kan* is functioning as a causative marker in *meminjamkan* ‘cause to lend’ taking the Theme as the direct object and the Recipient of the lending marked by a prepositional phrase. On the other hand, *-i* is functioning as an applicative, marking both its Recipient and Theme arguments with an NP.

- (27) a. *Ia mem-(p)injam uang dari saya.*
 3SG ACT-borrow money from 1SG
 “S/he borrowed money from me.” (Arka et al., 2009)
- b. *Ia mem-(p)injam-kan uang kepada saya.*
 3SG ACT-borrow-KAN money to 1SG
 “S/he lent money to me.”
- c. *Ia mem-(p)injam-i saya uang.*
 3SG ACT-borrow-I 1SG money
 “S/he lent me money.” (Arka et al., 2009)

2.5.3 Causative/Applicative Alternation

In this alternation, both *-kan* and *-i* are valency increasing constructions. An intransitive root with one argument combines with *meN-* and *-i* to increase its valency by adding either the A or O argument. An intransitive root with one argument combines with *meN-* and *-kan* to increase its valency by adding an external Causer. The alternation is summarized in Table 2.6. Consider the root *nikah/kahwin* ‘marriage’ in (28). In (28a), the verb *menikah* ‘to marry’ takes one argument while in the *-i* form (28b) the verb takes two arguments, both participants of the event of marrying. Both of the arguments are marked by noun phrases. In (28c) affixing *-kan* adds an external causer of the event, in this case the head chief. The third argument is marked by a prepositional phrase headed by *dengan* ‘with’.

	ARG1	ARG2	ARG3
<i>meN</i> -ROOT- $\emptyset$	AGENT/THEME NP	- -	- -
<i>meN</i> -ROOT- <i>i</i>	AGENT/THEME NP	THEME/PATIENT NP	- -
<i>meN</i> -ROOT- <i>kan</i>	CAUSER NP	AGENT/THEME NP	THEME/PATIENT NP

Table 2.6: Causative/Applicative Alternation

- (28) a. *Siti telah me-nikah.*
 Siti PRF ACT-marry
 “Siti has married.”
- b. *Dia me-nikah-i janda kaya bernama Anne.*
 3SG ACT-marry-I widow rich named Anne
 “He married a rich widow called Anne.”

- c. *Dan Tok Ketua pun mem-nikah-kan Kudin dengan Tijah.*
and head chief also ACT-marry-KAN Kudin with Tijah

“And the head chief married Tijah off to Kudin.” (from ‘Jambatan Laut China Selatan’, Sastera 1999, (Nur Amirah Binte Khairul Anuar et al., 2017)

2.6 Classifying Verbs in Indonesian

Alternations like those mentioned in the previous section are important because they show contrasting behaviors of verbs. In his seminal paper “The Grammar of Hitting and Breaking” Fillmore (1970) recognized that verbs in English can be grouped into classes based on shared grammatical behavior such as these alternations. Specifically he showed how verbs like *hit* and *break* are members of a larger classes of verbs whose members share similar patterns of behavior. Levin (1993) based her work on Fillmore and others ((Jackendoff, 1992; Dixon, 1991; Levin & Hovav, 1991) and presented her comprehensive findings in her book “English Verb Classes and Alternations” where she identified verb classes in English based on the alternations that the group were sensitive to. Moving beyond English, this idea has been identified and explored in other languages such as Lhasa Tibetan (DeLancey, 1995), and Kimaragang Dusun (Kroeger, 2010).

Looking into Indonesian, the closest attempt at establishing verb classes in a way similar to Levin was made by Voskuil (2000) who compared Indonesian with Dutch and Tagalog to create a taxonomy of verbs morphologically. He contrasted the affixes *-kan* and *-i* with the *-an* affix in Tagalog and *be-* prefix in Dutch. Since Indonesian has a rich affixation system, we can use that to inform our understanding of how verbs in Indonesian behave. Some work has been done here by Alwi et al. (2014) and Kridalaksana (1989) who analyzed and provided examples of verbs based on their morphological and syntactic types. They also provide lists of verbs and the affixes that they take or don’t take as well as the number of arguments they take (i.e. if they are intransitive, semitransitive, monotransitive or ditransitive). This research seeks to find out if grouping verbs using morphology can lead us to establish meaningful verb classes that share semantic features. A class of verbs should have distinguishable semantic and syntactic features from other classes. Ideally this work would group verbs based on all affixes in the Indonesian language but this is too big of a task and out of the scope of an MA. It can also lead to many fine grained groups of singletons with very specific features that may not be as informative. Instead, I will focus on the most commonly used affix *meN-* and its interaction with *-kan* and *-i*.

Using a large corpus of written Indonesian I will identify all verbs with *meN-* and group them by whether or not they take the affix *-kan* and *-i*. I will create a database of sentences containing these verbs that I identified, and then annotate the sentences with their syntactic information (the number of arguments they take) and semantic roles. In the following chapter I will explain

in detail how this was done. Then I will look for semantic similarities within each group using the roles I annotated.

Chapter 3

Research Method

3.1 Overview

In order to study the affixes *-kan* and *-i*, I first created a corpus of morphologically classified verbs and their sentences. These sentences were then turned into a database for annotation. This chapter explains the method that I used to extract the relevant verbs from a large corpus and how I annotated the sentences with syntactic and semantic role information.

3.2 Resources used in thesis

3.2.1 Leipzig Corpora Collection Indonesian Sub-corpus

The data used in this thesis comes from the Indonesian sub-corpora of the Leipzig Corpora Collection (LCC) (Goldhahn et al., 2012). The corpora found in the LCC is freely available online at <http://corpora.uni-leipzig.de>. It features monolingual corpora of 136 languages with the same format and collected from comparable sources online such as news articles and web pages. The Indonesian sub-corpora were recently reclassified from four distinct categories that covered a range from Malay to Indonesian: *msa* (Malay), *ind* (Indonesian), *ind-id* (Indonesian-Indonesia), *ind-bn* (Indonesian-Brunei) to three categories by regional variety: *zsm* (Brunei Malay/Malaysian/Singapore), *ind* (Indonesian) and *msa* (indeterminate). The languages were identified using linguistic properties such as spelling as well as metadata information such as the domain country of the website (Nomoto et al., 2018a). The reclassified Indonesian sub-corpus contains 18,530,000 sentences from news, web crawls and Wikipedia pages. The tab separated files in Table 3.1 correspond to the biggest files available on the LCC website for their respective years. I have chosen only the files reclassified as *ind* and excluded the files reclassified as *msa* (indeterminate), which can be either Malay or Indonesian, to generalize the results over just Indonesian. My decision to use the Indonesian sub-corpus was due to its availability, size and genre as I explain further below. All examples from the LCC corpus will be referenced by the

Year	Genre	Size (K)	File name	Corpus ID
2008	News	300	ind_news_2008.7.300K.tsv	News.2008
2009	News	300	ind_news_2009.8.300K.tsv	News.2009
2010	News	300	ind_news_2010.9.300K.tsv	News.2010
2011	News	300	ind_news_2011.10.300K.tsv	News.2011
2011	Newscrawl	3 000	ind_newscrawl_2011.5.3M.tsv	Crawl.2011
2012	News	300	ind_news_2012.11.300K.tsv	News.2012
2012	Newscrawl	300	ind_newscrawl_2012.6.3M.tsv	Crawl.2012
2012	Web	3,000	ind_web_2012.13.3M.tsv	Web.2012
2012	Mixed	3,000	ind_news_2012.4.3M.tsv	Mixed.2012
2013	Web	1,000	ind_web_2013.2.1M.tsv	Web.2013
2015	Web	3,000	ind_web_2009.8.300K.tsv	Web.2015
2016	Wikipedia	1,000	ind_wikipedia_2016.14.1M.tsv	Wiki.2016a
2016	Wikipedia	30	ind_wikipedia_2016.16.30K.tsv	Wiki.2016v
	Total	18,530		

Table 3.1: Indonesian data available on LCC

Corpus ID following Table 3.1.

1. **Availability.** The LCC and all corpora in it are open data under the Creative Commons License CC BY (Leipzig Corpora Collection, 2019).
2. **Size and structure.** The Indonesian sub-corpora from the LCC are the biggest data source available for written Indonesian, as far as I am aware. This is useful as the large size will help to capture generalizations of how the language is represented in standard Indonesian. Also, since the corpora are made up of randomly selected sentences from different sources (Goldhahn et al., 2012), biases from any one writer or genre can be minimised or avoided. While the issues listed above cannot be completely controlled since the method of data collection is not primary, choosing a big, random corpus is the best way of ensuring that the results are as close to real world use as possible.
3. **Written Indonesian.** According to Sneddon et al. (2010), the spoken or informal varieties of Indonesian have fewer affixes compared to the formal or written forms of Indonesian. Since my goal is to study the distribution of affixes on verbs I chose written Indonesian so that there will be more instances of affixes to study. The data, however, does include instances of spoken language written down as in reported speech in news or fiction. I included those in my analysis because of its standard (i.e. written) medium.

3.2.2 MALINDO Morph Morphological Dictionary

The second source of data comes from the MALINDO Morph morphological dictionary¹ (Nomoto et al., 2018b). This dictionary is the first free online morphological dictionary of Indonesian and Malay with a total of 232,550 lines containing information on root, prefix, suffix, circumfix, reduplication of surface forms. It is the largest morphological dictionary for Malay and Indonesian currently available. The data from this dictionary is taken from two authoritative dictionaries from Indonesian and Malay, the *Kamus Besar Bahasa Indonesia* (KBBI) and *Kamus Dewan*. MALINDO_Morph is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The lexicon does not have part of speech marked. To avoid nouns and adjectives I used only roots where the *meN-* prefix was attested.

As seen in table 3.2, the lemma *makan* ‘eat’ has 15 morphological forms, not including the possible morphological combinations with clitics. To account for all the possible forms that could appear in the corpus I created all possible hypothetical surface forms of these lemmas into a new dictionary. Specifically, I created the *0-* and *di-* prefixed surface forms and combined them with *-kan -i* and *-0*. Then I created even more surface forms by combining them with the personal pronoun *dia* that have been attested to attach to the beginning of a verb much like the person clitics *ku=* and *kau=*. The clitics *=nya*, *=mu*, *=ku* and the emphatic marker *=lah* were also accounted for. The resulting hypothetical dictionary contained 334 296 entries.

ID	Lemma	Surface form	Prefix	Suffix	Circumfix	Reduplication
cc-48826	makan	dimakan	di-	0	0	0
cc-48827	makan	kemakanan	0	0	ke-an	0
cc-48828	makan	ketermakanan	ter-	0	ke-an	0
cc-48829	makan	makan	0	0	0	0
cc-48830	makan	makan-makan	0	0	0	R-penuh
cc-48831	makan	makan-makanan	0	-an	0	R-penuh
cc-48832	makan	makanan	0	-an	0	0
cc-48833	makan	memakan	meN-	0	0	0
cc-48834	makan	memakani	meN-	-i	0	0
cc-48835	makan	memakankan	meN-	-kan	0	0
cc-48836	makan	pemakan	peN-	0	0	0
cc-48837	makan	pemakanan	0	0	peN-an	0
cc-48838	makan	sepemakan	se-+pe-	0	0	0
cc-48839	makan	termakan	ter-	0	0	0
cc-48840	makan	termakankan	ter-	-kan	0	0

Table 3.2: Data in MALINDO Morph for lemma *makan* ‘eat’

¹The dictionary is freely available at https://github.com/matbahasa/MALINDO_Morph.

3.3 Processing the corpus

Before processing the data, I removed a subset from the data namely “ind_mixed_2012.4.3M.tsv” to use as future test data. The remaining data set contained 15,530,000 sentences. In the following sections I will detail how I identified the verbs to be studied from the subcorpus and the types of annotations made on the corpus.

3.3.1 Reading in the corpus

The tab separated .tsv files were read using a python script written by me. The sentences in each file were tokenized using the default NLTK tokenizer i.e. the treebank word tokenizer which tokenizes by white space (Bird et al., 2009). I then identified the verbs from the tokenized sentences using a combination of regular expressions and the morphological dictionary.

3.3.2 Identification of verbs form the corpus

There are several possible methods to identify a verb in a corpus. One method is to rely on Part-of-Speech (pos) tags. In this method each token in the corpus must first be tagged with its Part-of-Speech information, either manually or automatically. Well-resourced languages such as English have robust Part-of-Speech taggers. While some work has been done to develop Part-of-Speech taggers for Indonesian by Pisceldo et al. (2009) and Fam et al. (2014), the current automatic Part-of-Speech taggers have not reached a high enough level of accuracy for fine-grained disambiguation. On the other hand, manually tagging Part-of-Speech for every token in the sentence was also not a feasible option. Therefore neither option was used in this research. Instead, verbs were identified using a combination of regular expression rules and a morphological dictionary. While not a perfect solution, it provided a level of control over the selection that I otherwise would not have with a PoS tagger.

The first step to identifying potential verbs without a pos tag was to use a regular expression, or a special text string for describing a search pattern which runs through the entire body of text and identifies tokens that match the requirements specified in the expression. Since Indonesian is a language with a rich morphology with verbs almost always taking an affix, I used that to form the requirements of my regular expression search. The criteria used to identify verbs in Indonesian were:

1. Tokens must begin with the lowercase characters
 - a) ‘me’ and followed by any one of the characters ‘m,n,l,w,r or y’,
 - b) ‘di’, ‘ku’, or ‘kau’.
2. It is followed by zero to any number of characters except for line terminators.
3. It is followed by zero to one occurrence(s) of the characters ‘kan’ or ‘i’.

4. It is followed by zero to any number of the characters ‘nya’, ‘mu’, or ‘ku’.
5. It is followed by zero to any number of the characters ‘lah’ at the end of the token.

This is captured by the following regular expression

`r' ^ ((me[mnlwry] | di | ku | kau) . *(kan | i) ?) (nya | mu | ku) * (lah) * $`

This method understandably identified tokens which fit the string description but were not verbs so they were identified as exceptions to be removed through a separate list which can be found in the script. Notably pronouns such as the third person plural *mereka* ‘3PL’, and third person singular *dia*, *ia* ‘3SG’ and words such as *digit* ‘digit’, *menit* ‘minute’, *bukan* ‘not’ which may look like they have the affixes *meN-* or *di-* or *-kan* but are in fact, not verbs.

It is important to note that only lower case tokens were used in the search. This is because capitalized verbs (i.e. verbs that begin a sentence) have a tendency of being nominalized as seen in example (29) and were thus not suitable for this analysis.

- (29) *Mengangkat berarti mengagungkan, memberikan pembelaan dan menyebut*
 meN.lift means glorify, meN.give.KAN defense and meN.call
berarti juga mengenalkan dan memahami kebesaran dan keagungan
 means also meN.recognize and meN.understand greatness and glory
Allah swt.
 Allah swt

“Lifting means glorifying, defending (lit. giving defense) and mentioning means also introducing and understanding the greatness and majesty of Allah Almighty” (LCC:Web.2012)
 (LCC Source: <https://agusalwafier.wordpress.com/category/buletin-at-taqwa/>)

The tokens that were identified were added to a dictionary together with the sentence it was found in for further processing. Concurrently, I searched the corpus for all the hypothetical surface forms created from the MALINDO Morph dictionary and added the results of that search to the dictionary created in the first step. If a sentence contained two occurrences of the same verb they were added as separate entries in the dictionary. A total of 6,814 roots were identified from the corpus, with 125,885 surface forms and 80,319,040 tokens.

3.3.3 Selection and Grouping of Verbs

All verbs in this study are prefixed with *meN-* and have either the suffix *-kan* or *-i* or none at all $-\emptyset$. In total there are 8 hypothetical combinations of affixes seen in Table 3.3. The groups have been named with upper case and lower case letters N/n, K/k and I/i that stand for the three affixes $-\emptyset$, *-kan* and *-i* respectively. Upper case letters indicate attested forms while lower case letters indicate non-attested forms. For example, Group 1 (NKI) will have roots with all three

Group No.	1	2	3	4	5	6	7	8
Group Name	NKI	nKI	NkI	NKi	nKi	nkI	Nki	nki
$-\phi$	+	-	+	+	-	-	+	-
<i>-kan</i>	+	+	-	+	+	-	-	-
<i>-i</i>	+	+	+	-	-	+	-	-

Table 3.3: Groups and Features

possible surface forms while Group 2 (nKI) will have surface forms with *-kan* and *-i* but not $-\phi$ and so on. Group 8 (nki) was excluded because it is an empty set.

The number of sentences for every surface form was counted and surface forms with less than 60 unique tokens were excluded from further sorting, so that there would be at least 60 sentences for each surface form to be annotated. For example, the lemma *elak* ‘avoid’ has two *meN*- forms: *mengelakkan* which occurred 28 times and *mengelak* which occurred 84 times. This root was excluded from the grouping because both forms did not meet the minimum counts of 60. I also assumed that because of the nature of the corpus, some unexpected uses of a surface form, anomalies or errors are expected. For example, a surface form that is commonly not considered a proper form in the dictionary or by native speaker intuition may be used by one source or writer. I set the maximum count of these occurrences at 6. A summary of the conditions for each group is presented in Table 3.4.

Group	<i>men-ROOT-ϕ</i>	<i>men-ROOT-kan</i>	<i>men-ROOT-i</i>
NKI	≥ 60	≥ 60	≥ 60
nKI	≤ 6	≥ 60	≥ 60
NkI	≥ 60	≤ 6	≥ 60
NKi	≥ 60	≥ 60	≤ 6
nKi	≤ 6	≥ 60	≤ 6
nkI	≤ 6	≤ 6	≥ 60
nKi	≤ 6	≥ 60	≤ 6

Table 3.4: Criteria for groups

A .html file with a directory of all verbs and surface forms found as well as a random selection of example sentences for each surface form from the corpus can be accessed on the Github repository (<https://github.com/hannahchoiyj/meng-kan-i>). 1,775 verbs that satisfied the criteria listed above were grouped by their morphological types and the absolute counts are presented in Table 3.5.

Group Name	NKI	nKI	NkI	NKi	nKi	nkI	Nki	Total
	47	32	56	287	371	122	860	1,775

Table 3.5: No. of Roots Found per Group

The program selected all the roots from the smaller groups NKI, nKI and NkI and selected 50 random roots from the bigger groups NkI, NKi, nKi, and nkI. The selected roots along with the first 999 sentences for each surface form were extracted and made into a database. Every sentence was indexed with a unique sentence ID (SID). Some example SIDs can be seen in Table (3.6). Each SID contains 4 pieces of information: the first digit indicates group it belongs to, the next two digits indicates the index of the root, the next two digits can be 01, 02 or 03 to indicate the type of surface form it is (e.g. 01 → $-\phi$, 02 → *-kan*, 01 → *-i*) and the last 3 digits indicate the sentence number from 0 to 999.

SID	Group	Root	Form	Sent.	Lemma	Gloss
10101000	1	01	01	000	<i>mengenal</i>	‘to know’
10102000	1	01	02	000	<i>mengenalkan</i>	‘to introduce someone to something’
10103000	1	01	03	000	<i>mengenal</i>	‘to know something’
10201000	1	02	01	000	<i>menyebrang</i>	‘to cross’
10202000	1	02	02	000	<i>menyebrangkan</i>	‘to cause something to cross’
10203000	1	02	03	000	<i>menyebrangi</i>	‘to cross something’
20101000	2	01	01	000	<i>memaklumkan</i>	‘to cause something to be understood’
20202000	2	01	02	000	<i>menmaklumi</i>	‘to understand/know/admit something’

Table 3.6: Example SIDs from the database

On closer inspection after the annotation process was complete, I noticed that eight roots were incorrectly grouped because of an error in the code that lemmatized the verbs. The code did not properly account for the phonological changes brought about by the affixation process. As such, the roots fell outside the minimum boundaries set out at the beginning of this section. These roots were removed from the results from all the tables that will be presented in Chapter 4. The eight roots are listed in Table 3.7.

3.4 Annotation of the Database

The database containing the sentences was uploaded to an annotation interface modified from the NTU Multilingual Corpus (NTU-MC) developed by Bond et al. (2015) for annotation. 20 roots from each group and at least 20 sentences for every surface form were annotated. In total I aimed to annotate at least 4800 sentences. The annotation task was carried from June 2018 - November 2018. Each sentence was annotated twice. First the syntactic arguments were identified and then each syntactic argument was annotated with a semantic role. Figure 3.1 shows the annotation interface.

The annotation interface is able to create tags based on one or more lexical items in a sentence. The lexical items do not need to be adjacent to each other. For every sentence we first

Starting SID	Root	Surface Form	Gloss	Group
20201000	<i>tanya</i>	<i>menanyakan</i>	‘to ask something to someone’	nKI
20202000	<i>tanya</i>	<i>menanyai</i>	‘to ask someone’	nKI
21801000	<i>temu</i>	<i>menemukan</i>	‘find something’	nKI
21802000	<i>temu</i>	<i>menemui</i>	‘meet someone’	nKI
31101000	<i>pandang</i>	<i>memandang</i>	‘bahold’	NkI
31102000	<i>pandang</i>	<i>memandangi</i>	‘bahold (prog)’	NkI
51501000	<i>berita</i>	<i>memberitakan</i>	‘preach or tell’	nKi
60801000	<i>sudah</i>	<i>menyudahi</i>	‘finish or settle’	nkI
61101000	<i>cedera</i>	<i>mencederai</i>	‘injure’	nkI
70101000	<i>ubah</i>	<i>mengubah</i>	‘change’	Nki
70901000	<i>buang</i>	<i>membuang</i>	‘throw’	Nki

Table 3.7: Surface forms removed from the dataset

Learner Corpus' Tagger

CorpusDB: [Document: 1]

Sentence: "Saya masih harus mengajar bimbingan belajar dari Senin sampai Sabtu."

" Saya masih harus mengajar bimbingan belajar dari Senin sampai Sabtu .

mengajar	Default verb guess (DO NOT USE!)	<input type="text"/>	x
mengajar	ACT - Active	<input type="text"/>	x
Saya	ARG 1 - Subject (e.g. I ate the cake)	<input type="text"/>	x
Saya	AGENT - Volitional causer deliberately performing the event	<input type="text"/>	x
bimbingan belajar	ARG 2 - Object (e.g. I ate THE CAKE)	<input type="text"/>	x
bimbingan belajar	THEME - Participant affected by the event, no change of state	<input type="text"/>	x

Figure 3.1: NTU-MC Annotation Interface

identified the arguments of the predicate by clicking on the lexical items corresponding to an argument and creating a new tag. The new tag is labeled with the argument number. The process is repeated until all the arguments in the sentence have been identified. There are 5 possible argument labels listed in Table 3.8. I also decided not to make a distinction between core arguments and oblique arguments and instead annotated all the arguments as they appeared in the sentence. Sometimes a sentence had a missing argument e.g. the A argument of the sentence was dropped because it was previously mentioned outside of the tagged sentence. Those arguments were not identified or tagged.

The syntactic argument identification task was done by 3 annotators, myself and 2 other native speakers of Bahasa Indonesia. They were students and performed the task voluntarily. They were trained to identify arguments prior to the task and I personally oversaw them during the annotation task. The syntactic arguments identified by the 2 annotators were hand-checked by me for accuracy and consistency. A list of verb functions were also prepared for the annotation as seen in Table 3.9. The purpose of this list was to annotate the function that the affix was

Role	Description	Example
ARG1	Generally pre-verbal, it is usually the S/A argument of a transitive <i>meN</i> - verb	<i>saya memakan kue</i> “I eat cake”
ARG2	Generally post-verbal, it is usually the O argument of a transitive <i>meN</i> - verb	<i>saya memakan kue</i> “I eat cake ”
ARG3	Generally post-verbal after ARG2, it is usually the indirect object argument of a ditransitive <i>meN</i> - verb	<i>saya memasak kue itu untuk dia</i> “I eat cake ”
ARG4	Generally post verbal after ARG3, it is an extra argument found in some verbs that take more than 2 objects	I bet him \$50 he would lose
ARGX	Argument that usually takes the form of a clause or sentence	I gave him permission to eat

Table 3.8: Syntactic argument role list

performing.

However, after the annotation task was complete I discovered many errors and inconsistencies because of two main reasons: firstly, the identification of the function of the affix required a higher degree of linguistic knowledge than my volunteer annotators had and secondly, many verbs have more than two correct labels and while the annotation software allows multiple tags from a single category for each word, the resulting grouping will be too narrow for this initial attempt at classification. In addition I was unable to check and correct each tag made by the co-annotators because of constraints in time so I chose not to include this result in my analysis.

Label	Function
ACT	Active
PASS	Passive
CAUS	Causative
ITER	Iterative
BEN	Benefactive
APPL	Applicative
REFL	Reflexive

Table 3.9: List of Affix Functions

Once the syntactic annotation was complete, I hand-checked each identified argument for accuracy and annotated each syntactic argument with a semantic role from the list that can be found in 3.11. These are described in the next section.

3.4.1 Creation of the Semantic Role List

Semantic role annotation is the process of identifying arguments to a verb and their semantic roles. Fillmore (1970) used the term “case” for what is regularly referred to as “semantic roles” by identifying grammatical markers relating to semantic roles through morphological cases, prepositions, grammatical relations. However, such markers are not adequate indicators of noun phrases. If they were adequate, we would not even need semantic role labels because syntactic or morphological markers can be sufficient. For example, in English the preposition *with* is not only associated with instrument but also used to mark comitatives as seen in example (30):

(30) English, (Levin & Hovav, 2005, p.41)

- a. *Tracy washed the car with an enormous sponge.*
- b. *Tracy washed the car with Stacy.* (Levin, 2011)

The task of semantically annotating roles within a sentence is a complicated task with a host of problems and disagreements among researchers. Problems with semantic role annotation is described at depth by Levin & Hovav (2005). Most crucially, they note that “It is difficult to find reliable diagnostics for isolating precisely those arguments bearing a particular role” (p.38). There are no diagnostic tests that can be consistently applied to an argument to decide whether or not that argument bears that particular role (as compared to lexical and syntactic tests) in one language, and each language comes with idiosyncrasies that make creating a universal set of roles a very challenging task.

Another problem is identifying the correct granularity to use in defining semantic roles. Too coarse and it will not be informative enough to capture generalizations but too narrow and there will be too many semantic roles to make meaningful generalizations. A good-sized semantic role list should be able to minimally distinguish all identifiable semantic roles in the language without making unnecessary distinctions.

This has resulted in multiple semantic role lists being developed for different languages and using different theories (Fillmore et al., 2003; Schuler, 2005; Jaworski & Przepiórkowski, 2014). Sheinflux et al. (2017) created a semantic representation that captures the interplay between the syntax and semantic aspects of argument structure in Hebrew. Their role set was designed to be a syntactically driven semantic representation with guidelines outlined below:

1. The semantic roles need to be semantically contentful and clearly defined.
2. The semantic roles should be language-independent.
3. The semantic roles should be consistent across all predicates, that is, a semantic label should have the same denotation regardless of the predicate.
4. There should be one argument per semantic role.

With these guidelines in mind they derived a semantic role list by studying existing role lists and also by annotating a corpus. The list is presented in Table 3.10.

Arg1	The prototypical member of this category is an Actor. Also included are Perceiver and Causer.
Arg2	This category corresponds to what is often referred to as Theme. Members of this category are negatively characterized as not belonging to the other, more semantically coherent, categories.
Arg3	Members of this category are prototypically animate beings which are positively or negatively affected by the action or state. It includes arguments which can be labeled Affectee, Benefactive, Malfactive, or Recipient.
Arg4	This category includes semantic arguments which are Attributes, or in other words, predicative of another argument in the clause.
Arg5	This category includes semantic arguments which denote Source in both its physical and metaphorical meanings.
Arg6	This category includes semantic arguments which denote Goal in both its physical and metaphorical meanings.
Arg7	This category includes semantic arguments which denote Location in both its physical and metaphorical meanings.
Arg8	This category includes arguments which denote the Topic (of Communication).
Arg9	Members of this category denote Instrument arguments, usually an inanimate entity causally involved in the action.
Arg10	Members of this category denote Comitative arguments, usually an animate being who accompanies another participant of the event.

Table 3.10: Semantic Roles by Sheinflux et al. (2017)

This role list was used by Nur Amirah Binte Khairul Anuar et al. (2017) for an annotation task for Malay and Indonesian *-kan* and *-i* verbs but they encountered some problems while annotating their data set using the role list. Firstly, Arg1 (Actor) and Arg3 (Patient) were too broad to meaningfully distinguish the arguments in Indonesian, especially in the case of *-kan* verbs. As mentioned in Section 2.3, one of the main roles of *-kan* that contrasts with both $\emptyset$ and *-i* verbs is that *-kan* functions as a causative marker that introduces a Causer to the sentence. By merging all the roles related to Actor into Arg1 we will not be able to differentiate the Causer from the other arguments to study its function. Similarly, *-kan* is known to introduce a benefactive argument to the sentence. By subsuming all the roles under Patient we will not be able to differentiate the Benefactive participant from other arguments. Secondly, there was no Arg for Force and Experiencer or Stimulus in their role list.

In order to annotate my data, I curated my own semantic role specific to Indonesian. I did this for several reasons. Firstly, my goal was to understand the function(s) of the affixes *-kan* and *-i* and so I needed a list that was more semantically driven than Sheinflux et al. (2017). I

adopted the thematic role list from Payne (2006) who based it on the work by Berk (1999). This list was more fine-grained than the list by Sheinfx et al. (2017). I also added new roles that Indonesian could be sensitive to such as Semantic role list curated for this annotation task tried to be as close to the roles used by other researchers while at the same time narrow enough to capture the semantic roles. The roles that I decided on are listed in Table 3.11. The definition and examples in this table are taken from Payne (2006). I also included Indonesian examples.

As with the list of affix functions in 3.9, there is also the possibility of choosing more than one role in the semantic role list for a single argument. Take a sentence like “I sent John a letter”. Here “I” can either be an Agent and Source. While this was possible, I chose to follow the standard practice in semantic role labeling where annotators chose only one role per argument such as PropBank (Palmer et al., 2005).

3.5 Processing the Annotated Database

The annotated sentences are stored as an SQLite database (Hipp, 2010) following the schema provided by NTU-MC. The results of the annotation can be found in Chapter 4, and the database will be made available with the thesis.

Role	Description	Example
AGENT	Volitional causer deliberately performing the event	Bill ate his soup quietly, Tono memakan kue itu
STIMULUS	Entity that prompts sensory or emotional feeling - not deliberately	Bill ate his soup quietly, Tono memakan kue itu
FORCE	Non-volitional causer of an event	An avalanche destroyed the ancient temple. Hujan tercurah membasahi bumi
EXPERIENCER	The entity that receives sensory or emotional input of an event	Susan heard the song. Susi mendengar suara lonceng
THEME	The participant most directly affected by the event, no change of state	We believe in one God. I have two children. I put the book on the table. He gave the gun to the police officer. Toni memasak nasi untuk Budi
PATIENT	The participant that undergoes the action and changes its state	The falling rocks crushed the car. Hujan tercurah membasahi bumi
INSTRUMENT	Argument used to carry out the action	Jamie cut the ribbon with a pair of scissors. Ibu menulis surat dengan pena yang baru .
RECIPIENT	A special kind of goal associated with verbs expressing a change in ownership, possession.	I sent John the letter. He gave the book to her. Ibu memberi surat kepada saya
BENEFICIARY	The entity for whose benefit the action occurs	I baked Reggie a cake. He built a car for me. I fight for the king. Riko memasak nasi untuk/buat saya
CONTENT	The proposition or content of a propositional event	Mereka mengatakan bahwa dia sudah pulang
LOCATION	Where the action occurs	Johnny and Linda played carelessly in the park. Jono menyanyikan lagu itu di rumah
GOAL	Where the action is directed towards	The caravan continued on toward the distant oasis. He walked to school
SOURCE	Where the action originated	The rocket was launched from Central Command. She walked away from him
MANNER	The way in which an action is carried out	With great urgency, Tabitha phoned 911
PURPOSE	The reason for which an action is performed	Tabitha phoned 911 right away in order to get some help
CAUSE	What caused the action to occur in the first place; not for what, rather because of what	Because Clyde was hungry, he ate the cake
ADVERBIAL	Not a semantic role. The verb needs it. Only for some verbs behaving as adverbs	selamanya, biasanya

Table 3.11: Semantic Roles for the Annotation Task

PART II:

FINDINGS

Chapter 4

Results

4.1 Overview

This chapter presents the results from the annotation task and is divided into three parts. In Section 4.2, I present some statistics of the annotated corpus. In Section 4.4 I present the syntactic arguments and semantic roles associated with each group. Finally in Section 4.5 I discuss the distribution of each role across the groups.

4.2 The Annotated Corpus

As mentioned in the previous chapter, a total of 140 roots or 240 surface forms were annotated from seven groups. The list of verbs and surface forms along with their glosses can be found in Appendix B. At least 20 sentences were annotated for each surface form. As mentioned in Section 3.3.3, 8 roots were excluded from the data set after the annotation task because they were outside of the minimum boundaries and hence did not belong to the group. The revised corpus contains 4,671 sentences as seen in Table 4.1. The surface forms can also be divided into 3 large groups, with 77 surface forms affixed with *-kan*, 75 surface forms affixed with *-i* and 77 surface forms without affixes as seen in Table 4.2.

Groups	NKI	nKI	NkI	NKi	nKi	nkI	Nki	Total
No. of verbs	20	18	19	20	19	18	18	132
No. of surface forms	60	36	38	40	19	18	18	229
No. of sentences	1,210	720	793	831	382	378	357	4,671

N	<i>meN</i> -ROOT- $\emptyset$ attested	n	<i>meN</i> -ROOT- $\emptyset$ not attested
K	<i>meN</i> -ROOT- <i>kan</i> attested	k	<i>meN</i> -ROOT- <i>kan</i> not attested
I	<i>meN</i> -ROOT- <i>i</i> attested	i	<i>meN</i> -ROOT- <i>i</i> not attested

Table 4.1: Number of tagged sentences per group

Type	+kan	+i	Null
No. of surface forms	77	75	77
No. of sentences	1,576	1,511	1,584

Table 4.2: Number of tagged sentences by type

4.3 Syntactic Arguments Overview

Every argument in the corpus was hand-identified by annotators and then annotated twice, once for the syntactic role (i.e. ARG1, ARG2 etc), and then for its semantic role (Agent, Patient, Theme, etc). A total of 15,003 tags were recorded with 89 different (ARG, Role) pair combinations.

Label	NKI	nKI	NkI	NKi	nKi	nkI	Nki	+kan	+i	Null	All
ARG1	90.08	96.89	95.33	90.01	88.54	96.18	93.25	90.78	95.79	91.14	93.02
ARG2	85.45	92.41	89.7	87.97	92.12	98.33	92	97.84	97.46	78.88	90.23
ARG3	31.57	20.27	10.18	24.67	20.05	9.31	20.5	42.16	14.49	16.32	21.22
ARG4	0.91	0.25	0.12	0.72	0	0	0.5	0.72	0.37	0.49	0.45
ARGX	1.4	0.75	0	0	0	0	2.5	0.6	0.12	1.27	0.67

Table 4.3: Syntactic Arguments per Group (%)

Table 4.3 presents the distribution of the five syntactic argument tags per group. The values in Table 4.3 correspond to the relative frequency of the argument in relation to the number of sentences in the group. For example, In the group NKI (roots with attested *-kan*, *-i* and *-ø*), ARG1 was identified in 90.08% of the total number of sentences in that group. According to the data, ARG1 was the most common argument with 93% frequency of all sentences recorded. Considering that all the verbs are affixed with *meN-*, an active voice marker, one could expect a high percentage of ARG1 as the pre-verbal agent. However, there are instances where the agent (or ARG1) to the verb is not present in the sentence. One such example is seen in (31) for *merealisasikannya* ‘to realize/actualize’.

- (31) *Ia men-(t)ambah-kan, rencana masuk=nya BUMN ke Myanmar sudah*
 3SG ACT-add-KAN, plan enter-DEF BUMN into Myanmar PRF
di-bahas secara men-dalam tinggal ø me-realisasi-kan=nya saja.
 PASS-discuss in.a.manner ACT-deep left ø ACT-realize-KAN=3SG only

“He added, the plan to enter BUMN into Myanmar had been discussed in depth, just ø realizing it left”

In this case, the actor, or ARG1 of *merealisasikannya* ‘to realize/actualize’ is not present. One possible explanation could be because this argument is recoverable from the wider context of

the sentence. In Indonesian, it is common to join two sentences and omit the shared argument if the subject is shared in a complex sentence as seen in example (32).

- (32) *Menurut dia, salah satu cara untuk mengurangi jumlah orang-penghuni adalah*
 According 3sg wrong one way to decrease sum people-resident cop
dengan Ø me-mulang-kan mereka
 with Ø ACT-return-KAN 3PL

“According to him/her, one way to reduce the numbers of residents is to send them (back)”
 [SID: 50201005]

ARG2 was also found at high frequencies across the groups, with 90.23% overall frequency. This suggests that the verbs across all the groups generally take at least two arguments or are transitive. This is intuitive because it has been well-documented in literature that the affixes *-kan* and *-i* can be used to increase a verb’s valency. Surface forms without *-kan* or *-i* (group Null in the table) show a bigger difference between the frequencies of ARG1 (91.14%) and ARG2 (78.88%) which suggests that there is a higher chance that *meN*-ROOT-Ø verbs are intransitive. One example of an intransitive *meN*-ROOT-Ø verb is seen in example (33)

- (33) *Di tengah me-lesu-nya ekonomi global.*
 LOC middle ACT-become.slow-DEF, economy global

“In the midst of the global economy slowing down” [SID: 71601004]

Interestingly, for the group nkI, whose roots has the surface form containing the suffix *-i* but not *-kan* and *-Ø*, ARG2 is almost always present at 98.33%, even more than ARG1 in that group or in total. This suggests that *-i* may display some preference for an obligatory ARG2, which I will discuss further in chapter 5.

Clear differences between the effects of *-kan* and *-i* can be seen for ARG3. From the data we can see that the frequency of ARG3 more than doubles when the group takes *-kan* but not *-i* (e.g. group nkI vs nKi and group nKi vs nkI). This difference is also seen when we compare *-kan* only surface forms and *-i* only surface forms. This suggests that *-kan* is more commonly employed to increase a verb’s valency compared to *-i* and I will look into the possible reasons why in the following chapter.

4.4 Group Overview

Each syntactic argument was annotated with one of the 24 semantic roles from the list mentioned in Chapter 3. The results of the (arg, role) pairs for each group are presented in section 4.5. There are 5 colored bars for each semantic role in every group, and they represent absolute counts for that particular (ARG, Role) pair. Table 4.4 provides a guide to how the arguments are

presented in the table. The highest value from each role is recorded for every row. For example, the most common (ARG, Role) pair annotated was (ARG1, Agent) for the group +i with 1,108 occurrences. In the following sections I will outline the roles associated with each group.

168

The first black bar denotes the values for ARG1, the green bar denotes the values for ARG2. The red bar denotes the values for ARG3, the blue bar denotes the values for ARG4 and the final black bar represents ARGX. The value for the highest bar is indicated on the left of each row.

Table 4.4: Guide for Tables 4.6 – 4.10

Role	NKI	nKI	NkI	NKi	nKi	nkI	Nki	+kan	+i	Null
1108										
Agent										
353										
Patient										
813										
Theme										
111										
Cause										
100										
Force										
102										
Experiencer										
102										
Stimulus										
168										
Location										
157										
Goal										

49	
Source	
114	
Recipient	
43	
Beneficiary	
33	
Instrument	
83	
Content	
48	
Result	
24	
Attribute	
18	
Purpose	
1	
Degree	
4	
Duration	
8	
Extent	
31	
Manner	

Table 4.5: All Roles by Group

Group NKI has all three affix combinations *-kan*, *-i* and $-\emptyset$. In this group, the semantic roles associated with ARG1 is Agent, Theme, Force, Cause and Experiencer. ARG2 is most commonly associated with Patient and Theme but also includes Stimulus, Location and Goal, Source, Recipient, and Beneficiary. ARG3 is most commonly associated with Location, Goal and Source. Interestingly, this is the only group with similarly high numbers of arguments with the semantic role Goal for ARG2 and ARG3. This suggests an alternation pattern that will be further discussed in Chapter 5.

Group nKI has the affix combinations *-kan* and *-i* but not $-\emptyset$. In this group the semantic roles associated with ARG1 are Agent, Cause, Experiencer and Stimulus. An interesting pattern is

that Experiencer and Stimulus roles are shared between ARG1 and ARG2 at similar frequencies, which suggest an alternation. ARG2 is also associated with Recipient while ARG3 is associated with Result.

Group NkI has the affix combinations *-i* and $\emptyset$ but not *-kan*. In this group the semantic roles associated with ARG1 are Agent, Theme and Force. This group has one of the highest frequencies for the semantic role Force, with over 12% of sentences having Force as an ARG1. The verbs in this group do not have a *-kan* form and have of the lowest frequencies for the role Cause. Other groups that share this trend for Cause are nkI and Nki which suggests that *-kan* may play a role in the licensing of arguments with the semantic role Cause. There is a relatively low frequency of ARG3 arguments compared to groups with *-kan* verbs.

Group NKi has the affix combinations *-kan* and $\emptyset$ but not *-i*. In this group the semantic roles associated with ARG1 are Agent, Cause, Force and Instrument. ARG2 is associated with Patient, Theme, Content and Recipient. ARG3 is associated with Goal and Result.

Group nkI The verbs in this group only has one surface form with *-kan* and not *-i* or $\emptyset$. ARG1 is associated with Agent, Cause and Stimulus while ARG2 is associated with Patient, Theme, Experiencer, Recipient and Beneficiary. ARG3 is associated with Goal and Result.

Group nkI The verbs in this group have only one surface form with *-i* and not *-kan* or $\emptyset$. ARG1 is associated Agent, Cause and Experiencer while ARG2 is associated with Patient, Theme, Stimulus and Location. There is a low frequency of ARG3 arguments for this group and the last group Nki which suggests that most of the verbs in groups without a *-kan* form are not ditransitive.

Group Nki The verbs in this group have only one surface form with $\emptyset$ and not *-i* or *-kan*. These are verbs that do not take either *-kan* or *-i* forms. ARG1 is associated only with Agent, Theme and Force. ARG2 is associated with Patient, Theme and Experiencer. ARG3 is associated with Instrument, Result and Source.

In general, verbs with *-kan* (Group +kan) have the highest frequency of Agent as ARG1 compared to verbs with *-i* (Group +i) or verbs with $\emptyset$. *-kan* verbs also have a high frequency of ARG1 taking the semantic role Cause. Stimulus is also a role that is closely associated with ARG1 for *-kan* while Experiencer is closely associated with ARG2. This is in contrast to *-i* verbs which take the complimentary Experiencer role as ARG1 while taking Stimulus as ARG2. *-i* verbs also often take Goal and Location as ARG2 while having low frequencies of ARG3. This is in contrast to *-kan* verbs which have low frequencies for Goal and Location as ARG2 but instead have higher frequencies of these roles as ARG3. *-i* verbs take Recipient as ARG2 while *-kan* verbs are split between ARG2 and ARG3.

Verbs that neither take *-kan* or *-i* forms behave differently from the other two large groups. While the most frequent semantic role as ARG1 is Agent, ARG1 can also be Theme or Patient. ARG1 is also not associated with Cause, unlike *-kan* or *-i* verbs. ARG2 is most frequently Theme or Patient.

4.5 Role Overview

Table 4.6: Agent, Patient and Theme Roles by Group

The results have been broadly grouped by their similar behaviors. Table 4.6 shows the distribution of (ARG, Role) pairs across all the groups for the semantic roles Agent, Patient and Theme which are the at least three times more frequent than the other roles. Across all the groups Agent is the most frequent role for ARG1 while Patient and Theme are the most frequent roles for ARG2. Theme is associated with ARG3 across the groups at a lower frequency compared to the other two roles.

Table 4.7 presents the distribution of the next most frequent roles. From this table we can see that the semantic roles Cause and Force are predominantly ARG1. Cause is a highly frequent semantic role for groups that have *-kan* (NKI, nKI, NKI, nKi) while not for groups that

Table 4.7: Cause, Force, Experiencer and Stimulus Roles by Group

have *-i* (nKI, NkI). This suggests a reflection of *-kan*'s role as a causative marker as detailed in the literature by Sneddon et al. (2010) and Cole et al. (2008). However, it is also worth noticing that the frequency of Cause in the column *+i* shows that a considerable amount of *-i* verbs must contain Cause. This will be further discussed in Section 5. There is a high frequency of the semantic role Force for group NkI that is the result of three roots and 6 surface forms *menjangkit/menjangkiti* 'infect' *melanda* 'engulf', *melandai* 'slope' *merasuk* 'penetrate' *merasuki* 'enter,possess' which prototypically takes Force as the preverbal argument as in (34):

- (34) *Pada minggu kedua Februari, setidaknya banjir me-landa kawasan ini*
 On week second February, at.least flood av-engulf area this
sebanyak tiga kali.
 as.much.as three times

"In the second week of February, at least three floods struck this area." SID:31201010

Experiencer and Stimulus can either be ARG1 or ARG2 and usually co-occur within a sentence as seen in (35a). It is interesting to note that for these two roles there is a clear difference between the behavior of roots that take *-kan* and *-i*. Groups with only *-kan* verbs (i.e. Groups nKi, +kan) show that Experiencer is most frequently ARG2 while Stimulus is mostly ARG1. On the other hand, in groups with only *-i* (i.e. Groups nKI, +i), Experiencer is most frequently ARG1 while Stimulus is mostly ARG2. In the group that has both *-kan* and *-i* form such as Group nKI, there are both found in similar high frequencies. This trend will be further discussed as an alternation pattern in Chapter 5.

- (35) ARG1: Stimulus, ARG2: Experiencer

- a. *Tidak ada hal lain yang lebih meny-(s)enang-kan=ku.*
 NEG COP thing other REL more ACT-please-CAUS=1SG

"Nothing else pleases me more." SID 22001020

- (36) ARG1: Experiencer, ARG1: Stimulus

- a. *Anda selalu meny-(s)enang-i hal-hal baru.*
 2SG always ACT-please-I thing.REDUP new

"You always like new things." SID 22002020

Table 4.8 shows the distribution of semantic roles Instrument, Location, Goal, Source and Recipient. In contrast to Table 4.7 where the roles are associated with ARG1 or ARG2, the roles in this table are associated with ARG2 and ARG3. For ARG2, the roles with the highest frequency is Location and Recipient for verbs affixed with *-i*. For ARG3, Goal and Source are the most frequent roles for verbs affixed with *-kan*.

Table 4.8: Location, Goal, Source, Recipient, Beneficiary Roles by Group

For several semantic role lists, Recipient and Beneficiary are considered one role. For this annotation task they were separated to determine if the roles are used differently for *-kan* and *-i* verbs. According to the data there are different usage patterns for these two roles which justifies their distinction, at least for Standard Indonesian. Recipient and Beneficiary can either be ARG2 or ARG3 for *-kan* verbs (NKi, nKi, +kan) which suggests a possible alternation which will be explored in the next chapter. Recipient is also highly frequent in *-i* verbs while Beneficiary is frequent only for *-kan* verbs.

Table 4.9 show the distribution of Instrument, Content, Result, Attribute and Purpose roles. Although the counts are much smaller than the roles that have been discussed so far, interesting

Table 4.9: Instrument, Contant, Result, Attribute, Purpose Roles by Group

patterns emerge from these roles. Most notably, Instrument has a distribution pattern that is different from all the other roles. Firstly, it is a role used across all the groups relatively equally. Secondly, it is almost always either an ARG1 or ARG3, but not ARG2. Content, on the other hand, is a role that is generally shared between ARG2 and ARG3 across all the groups. It is also a role associated with ARGX.

Result, Attribute and Purpose are roles almost always associated with ARG3. However, these roles take up a very small percentage of all the arguments annotated.

Table 4.10 presents the distribution of semantic roles Degree, Duration, Extent and Manner. As seen from the table, the first three roles are very rarely used and as such may not be as relevant in a semantic role set for Indonesian. Manner is the only role that could be relevant in Indonesian. It is mostly an ARG3. This role is usually optional, and takes the preposition *dengan* ‘with’ as seen in example (37).

- (37) *jika kita meN-gaul-i=nya dengan mesra*
 if 3PL AV-embrace-APPL=3SG with intimate
 “if we embrace it intimately” [SID 60101000]

Role	NKI	nKI	NkI	NKi	nKi	nkI	Nki	+kan	+i	Null
1 ----- Degree
4 ----- Duration
8 ----- Extent
31 ----- Manner

Table 4.10: Degree, Duration, Extent, Manner Roles per Group

Chapter 5

Analysis

5.1 Overview

Most of the results corroborate what was discussed in the literature review in Chapter 2. In this chapter we will discuss valency changing morphemes in 5.2, the three most prominent alternation patterns discussed in the literature in Section 5.3 and comparisons of roots with the KBBI in 5.14.

5.2 *-kan* and *-i* as Valency Changing Morphemes

One of the main functions of *-kan* and *-i* listed in Section 2.3 was to increase the valency of verbs, from an intransitive verb to a transitive verb, and a transitive verb into a ditransitive verb. This can be seen in Table 5.1 which I have repeated from 4.3 in chapter 4. This table shows the percentage of each syntactic argument found for every group, as well as the percentage of syntactic arguments found for every sentence for a certain affix (i.e. column +*kan* for all sentences with *-kan*, +*i* for all sentences with *-i* and Null for all sentences with $-\emptyset$). As established in Section 2.2, ARG1 has a tendency to be dropped if it is recoverable from the context. The presence of an ARG2 does not imply that ARG1 is explicit in the sentence, but the presence of ARG3 heavily implies the presence of ARG2 and the presence of ARG4 implies the presence of ARG3. With that information in mind, from columns +*kan*, +*i* and Null we can see that *-kan* and *-i* verbs have a higher frequency of ARG2 (97.84% and 97.46% respectively) compared to ARG1 (90.78% and 95.79% respectively). This is a big contrast to $-\emptyset$ (Null) verbs with 78.88% of all $-\emptyset$ sentences having ARG2 but 91.14% of all $-\emptyset$ sentences having ARG1. This strongly suggests that verbs that take *-kan* and *-i* are at least more frequently transitive than $-\emptyset$ verbs that do not take *-kan* or *-i*. In addition 42.16% of all sentences with *-kan* (Group +*kan*) have an ARG3, compared to only 14.49% of all sentences with *-i* having an ARG3. The proportion of verbs having at least 3 roles is nearly three times more than *-i* verbs. This is consistent with the literature that describe *-kan* as increasing the valency of a transitive verb to a ditransitive verb

by adding a Causer.

Label	NKI	nKI	NkI	NKi	nKi	nkI	Nki	+kan	+i	Null	All
ARG1	90.08	96.89	95.33	90.01	88.54	96.18	93.25	90.78	95.79	91.14	93.02
ARG2	85.45	92.41	89.7	87.97	92.12	98.33	92	97.84	97.46	78.88	90.23
ARG3	31.57	20.27	10.18	24.67	20.05	9.31	20.5	42.16	14.49	16.32	21.22
ARG4	0.91	0.25	0.12	0.72	0	0	0.5	0.72	0.37	0.49	0.45
ARGX	1.4	0.75	0	0	0	0	2.5	0.6	0.12	1.27	0.67

Table 5.1: Syntactic Arguments per Group (%)

In Section 2.4 of the literature review, *-i* was mentioned to be an applicative morpheme that increases the valence of the verb by promoting an oblique argument to the object (ARG2). Generally this participant is a locative. This result is in line with the literature. The results for the role Location is reproduced in Table 5.2 from Table 4.8. From this table we can see that the frequency of Location as an ARG2 is highest for all sentences with *-i* (column +i) compared to *-kan* or *-i* sentences. In fact, the frequency for Location was higher for ARG3 in *-kan verbs*.

Table 5.2: Location by Group

Related to *-kan* and *-i* as valency changing morphemes, there are also other trends that are not previously discussed in literature. Comparing ARG1, we see that sentences with *-kan* have a tendency to drop their ARG1 more than sentences with *-i* or $\emptyset$. The percentage of sentences with ARG1 for the column +kan is lower than the average for all sentences. This is also evident in group nKi, which is the group that only has roots that have *-kan* surface forms, where only 88.54% of the sentences have an ARG1. The reasons for this is not immediately apparent. One possible reason could be that some roots in that group do not need to take ARG1 because that role is encoded in the lexicon. Another possible reason could be that *-kan* in general places less emphasis on having an explicit ARG1 in the sentence. Some examples for verbs that belong to this group without an overt ARG1 are: *melestarikan* ‘to conserve’, *merampingkan* ‘to streamline’, and *mengidentikkan* ‘to identify’. One example is shown in (38) below.

- (38) *Berbagai cara dan upaya bisa dilaksanakan untuk øme-lestari-kan khazanah*
various way and effort can be.implemented øto ACT-last-CAUS treasure
budaya yang beragam.
culture REL diverse

“Various ways and efforts can be carried out to preserve diverse cultural treasures” [SID 51601001]

5.3 Alternation Patterns

This section compares the annotation results against three prominent alternation patterns that were discussed in Section 2.5. First I will describe the alternation patterns and list some roots where this alternation patterns were attested along with some examples from the data. Wherever possible I will also discuss the patterns with respect to the general trends that were presented in Chapter 4. The roots involved in these alternations will be from groups NKI and nKI which have both *-kan* and *-i* forms attested. Included in Appendix C are subcategorization frames for the three argument sentences that were attested for the verbs in each group along with the frequency and a list of SIDs for that alternation pattern.

In order to discover the alternation patterns for the verbs I first extracted the semantic role patterns, which I will call **valences** since they inform us of the valency of the verb. A total of 262 valences were extracted from 4671 sentences. The top 5 valence with the highest relative frequency for *-kan* only sentences (column +kan), *-i* only sentences (column +i) and *-ø* only sentences (column Null) along with the combined relative frequencies for all sentences (column All) are presented in Table 5.3. Since the frames did not include ARG4 and ARGX roles, they have not been displayed in the tables. The values in the table e.g. 2.15 (n=26) refer to the relative frequency of the frame (in percentage) with the absolute counts in parenthesis. Values in **boldface** e.g. **9.7 (n=117)** indicate that it is one of the top 5 most frequent valence for that group.

From Table 5.3, we can see that the transitive Agent/Patient/None and Agent/Theme/None frames have the highest relative frequency frequent (appearing in 24% and 11% of all the sentences respectively) (see column All). However, the relative frequency of the other valences drop considerably after these two roles, with each role appearing in only 3% or less in all sentences. Generally this pattern also holds for *-kan* only sentences and *-i* only sentences having Agent/Patient/None and Agent/Theme/None as the most frequent at similar relative frequencies as the total. *-ø* only sentences also have Agent/Patient and Agent/Theme/None as the most frequent valences, but at a lower relative frequency compared to the total. The next highest frequencies in the top 5 valences for *-ø* only sentences are intransitive valences (Theme/None/None, Patient/None/None and Agent/None/None).

For *-kan* only sentences, the valences with the next highest frequencies are Agent/Theme/Goal, Stimulus/Experiencer/None and Agent/Content/None. On the other hand, *-i* only sentences on have Theme/Location/None, Experiencer/Stimulus/None and Agent/Recipient/None as the most frequent valences. The top 5 valence with the highest relative frequency for each group is presented in Table 5.4. Looking more closely into each group, we can see that Agent/Patient/None and Agent/Theme/None make up the highest percentage of sentences across all groups. In Group NKI, verbs that take *-kan* *-i* and *-ø* the most frequent valence are transitive or ditransitive.

ARG1	ARG2	ARG3	+kan	+i	Null	All
Agent	Beneficiary	None	1.33 (n=21)	0 (n=0)	0.06 (n=1)	0.47 (n=22)
Agent	Content	None	3.68 (n=58)	0.26 (n=4)	2.78 (n=44)	2.27 (n=106)
Agent	Experiencer	None	0.06 (n=1)	0 (n=0)	0.76 (n=12)	0.28 (n=13)
Agent	Location	None	0 (n=0)	2.45 (n=37)	0.82 (n=13)	1.07 (n=50)
Agent	None	None	0.82 (n=13)	0.86 (n=13)	3.91 (n=62)	1.88 (n=88)
Agent	Patient	None	11.29 (n=178)	13.77 (n=208)	9.34 (n=148)	11.43 (n=534)
Agent	Recipient	None	1.02 (n=16)	3.97 (n=60)	0.44 (n=7)	1.78 (n=83)
Agent	Theme	Goal	6.98 (n=110)	0.40 (n=6)	2.08 (n=33)	3.19 (n=149)
Agent	Theme	None	23.67 (n=373)	26.74 (n=404)	21.84 (n=346)	24.04 (n=1123)
Cause	Patient	None	3.68 (n=58)	1.26 (n=19)	0.19 (n=3)	1.71 (n=80)
Cause	Theme	None	1.46 (n=23)	3.71 (n=56)	0.38 (n=6)	1.82 (n=85)
Experiencer	Stimulus	None	0.63 (n=10)	6.22 (n=94)	1.83 (n=29)	2.84 (n=133)
Force	Patient	None	0.63 (n=10)	2.05 (n=31)	1.52 (n=24)	1.39 (n=65)
Patient	None	None	0 (n=0)	0.13 (n=2)	5.68 (n=90)	1.97 (n=92)
Stimulus	Experiencer	None	4.00 (n=63)	1.32 (n=20)	0 (n=0)	1.78 (n=83)
Theme	CoTheme	None	0.06 (n=1)	3.71 (n=56)	1.89 (n=30)	1.86 (n=87)
Theme	Location	None	0.13 (n=2)	7.48 (n=113)	2.08 (n=33)	3.17 (n=148)
Theme	None	None	2.86 (n=45)	0.99 (n=15)	7.13 (n=113)	3.70 (n=173)

Table 5.3: Most Frequent Valence for +kan, +i, Null Groups

ARG1	ARG2	ARG3	NKI	nKI	NkI	Nki	nKi	nkI	Nki
Agent	Beneficiary	None	0.25 (n=3)	0 (n=0)	0 (n=0)	0.12 (n=1)	4.71 (n=18)	0 (n=0)	0 (n=0)
Agent	Content	None	0.33 (n=4)	4.72 (n=34)	1.13 (n=9)	4.21 (n=35)	1.57 (n=6)	0 (n=0)	5.04 (n=18)
Agent	Experiencer	None	0 (n=0)	0 (n=0)	0 (n=0)	0 (n=0)	0.26 (n=1)	0 (n=0)	3.36 (n=12)
Agent	Location	None	0.33 (n=4)	0 (n=0)	3.28 (n=26)	0 (n=0)	0 (n=0)	5.03 (n=19)	0.28 (n=1)
Agent	None	None	2.06 (n=25)	1.53 (n=11)	3.53 (n=28)	1.32 (n=11)	0.52 (n=2)	0.79 (n=3)	2.24 (n=8)
Agent	Patient	None	9.67 (n=117)	14.58 (n=105)	5.67 (n=45)	11.07 (n=92)	10.47 (n=40)	21.96 (n=83)	14.57 (n=52)
Agent	Recipient	None	0 (n=0)	6.94 (n=50)	0 (n=0)	1.20 (n=10)	2.62 (n=10)	3.43 (n=13)	0 (n=0)
Agent	Theme	Goal	5.29 (n=64)	0.42 (n=3)	0 (n=0)	8.30 (n=69)	2.62 (n=10)	0.79 (n=3)	0 (n=0)
Agent	Theme	None	13.39 (n=162)	15.97 (n=115)	35.56 (n=282)	24.30 (n=202)	29.06 (n=111)	37.30 (n=141)	30.81 (n=110)
Cause	Patient	None	0 (n=0)	4.03 (n=29)	0 (n=0)	3.00 (n=25)	6.02 (n=23)	0.79 (n=3)	0 (n=0)
Cause	Theme	None	2.15 (n=26)	2.78 (n=20)	0.38 (n=3)	0.72 (n=6)	3.14 (n=12)	4.76 (n=18)	0 (n=0)
Experiencer	Stimulus	None	3.72 (n=45)	8.06 (n=58)	0 (n=0)	0.36 (n=3)	0 (n=0)	5.29 (n=20)	1.96 (n=7)
Force	Patient	None	0 (n=0)	0 (n=0)	6.81 (n=54)	1.20 (n=10)	0 (n=0)	0 (n=0)	0.28 (n=1)
Patient	None	None	1.32 (n=16)	0 (n=0)	0 (n=0)	6.98 (n=58)	0 (n=0)	0.53 (n=2)	4.48 (n=16)
Stimulus	Experiencer	None	0.17 (n=0)	7.64 (n=55)	0 (n=0)	0 (n=0)	6.81 (n=26)	0 (n=0)	0 (n=0)
Theme	CoTheme	None	0.083 (n=1)	0 (n=0)	9.46 (n=75)	0.12 (n=1)	0.26 (n=1)	2.38 (n=9)	0 (n=0)
Theme	Location	None	6.61 (n=80)	3.06 (n=22)	4.41 (n=35)	0.96 (n=8)	0 (n=0)	0.79 (n=3)	0 (n=0)
Theme	None	None	6.28 (n=76)	5.14 (n=37)	4.41 (n=35)	2.05 (n=17)	1.57 (n=6)	0.26 (n=1)	0.28 (n=1)

Table 5.4: Most Frequent Valence for All Groups

5.3.1 Displaced Theme/Locative Object Alternation

In this alternation (first described in Section 2.5.1) both *-i* and *-kan* verbs have three arguments. ARG1 is the Agent for both while in ARG2 the Theme for *-kan* verbs and Location for *-i* verbs. ARG3 in *-kan* verbs are marked by the Goal taking the preposition *ke* while for *-i* verbs are marked by the Theme taking the preposition *dengan*.

	ARG1	ARG2	ARG3
<i>meN-ROOT-kan</i>	AGENT NP	THEME NP	GOAL <i>ke</i> -PP
<i>meN-ROOT-i</i>	AGENT NP	LOCATION/GOAL NP	THEME <i>dengan</i> -PP

Table 5.5: Displaced Theme *-kan* and Location *-i* Object Alternation

The root *kirim* ‘send’ (39a) was the only example in the data that followed this pattern. In both sentences ARG1 is an Agent. ARG2 is a theme for the *-kan* sentence and a location for the *-i* sentence. ARG3 in both sentences are preposition phrases (PPs). In the *-kan* sentence ARG3 is a Goal headed by the preposition *ke* ‘to’ and in the *-i* sentence ARG3 is a Theme with the preposition *dengan* ‘with’.

- (39) a. *Jadi Ratu Duyung yang meng-(k)irim-kan=mu ke pulau ini?*
 so queen mermaid REL ACT-send-KAN=2SG to island this
 “So the mermaid queen sent you to this island?” SID 11302003
- b. *Para peneliti meng-(k)irim-i setiap lowongan kerja dengan dua lamaran.*
 DEF researcher ACT-send-I every opening job with two applications
 “The researchers sent each job opening with two applications.” SID 11303001

However, looking at the results shown in Table 5.6 reproduced here from Tables 4.8 and 4.6, we can see that for *-kan* verbs (column +kan), the most common argument for Theme was ARG2 at 813 instances out of a total of 910 instances of Theme in this group. The most common argument for Goal was ARG3 at 157 instances out of a total of 176 of Goal in this group. This trend supports the alternation pattern for *-kan* as shown in Table 5.5.

Looking at *-i* verbs (column +i), the alternation pattern is not as clear. The most common argument for Location was ARG2 at 168 instances out of a total of 180 instances of Location in this group. The most common argument for Theme was ARG2 at 605 instances out of a total of 910 instances of Theme in that group. Theme was an ARG3 on 63 instances. While this number is not as high as expected, it’s still in line with literature especially when considering that (ARG3, Theme) in *-i* verbs is almost twice as frequent as (ARG3, Theme) in *-kan* verbs (33).

Table 5.6: Location, Goal, Themes Role by Group

When we consider the characteristic function of *-kan* as a causativizer and *-i* as an applicative marker taking Location as the object (see Section 2.4), we discover that some verbs display a similar alternation which I will describe in the next section.

5.3.2 Causative/Locative Alternation

In this alternation, the valence for *-kan* is the same as the alternation in the previous section but the valence for *-i* is transitive, with Theme and Location as its arguments, summarized in Table 5.7 . Verbs in the data that pattern in this way are generally verbs of directed motion. Examples (40), (41) and (42) show this alternation for the verbs *menyebrangkan/menyebrangi* ‘to traverse across’ *melangkahkan/melangkahi* ‘to step over’ and *menerbangkan/menerbangi* ‘to fly’ respectively. Other verbs attested in the data include *menjejakkan/menjejaki* ‘to trek’ and *menerangkan/menerangi* ‘to shine/illuminate’.

	ARG1	ARG2	ARG3
<i>meN-ROOT-kan</i>	AGENT NP	THEME NP	GOAL <i>ke/kepada-PP</i>
<i>meN-ROOT-i</i>	THEME NP	LOCATION/GOAL NP	- -

Table 5.7: Causative-Locative Applicative Alternation

- (40) a. *Sambil menunggu Kapal Motor Cepat Kartini I yang akan meny-(s)eberang-kan*
while wait ship motor fast Kartini I REL FUT ACT-CROSS-KAN
kami ke pulau.
1PLEXCL to island
“While waiting for the Fast Ship Kartini I which will cross us across to the island”
SID 10202018
- b. *Satu orang tertangkap ketika sedang meny-(s)eberang-i sungai.*
one person caught when PROG ACT-CROSS-I river
“One person was caught while crossing the river.” SID 10203015
- (41) a. *ia hendak me-langkah-kan kaki=nya ke dalam.*
3SG want ACT-step-KAN foot=3SG to inside
“S/he wanted to step foot inside” SID 10302004
- b. *negara adidaya tunggal tersebut telah me-langkah-i batas keadilan yang*
country power first DEF PRF ACT-step-I limit fairness REL
dibuat=nya sendiri.
build=3SG REFL
“The single superpower has overstepped the limits of justice it had set for itself” SID
10303017
- (42) a. *para pembajak akan men-(t)erbang-kan=nya ke Gedung Putih.*
DEF hijacker FUT ACT-fly-KAN=3SG to House White
“The hijackers will fly it into the White House” SID 21401016
- b. *Sebanyak dua maskapai lain yang sudah men(t)erbang-i Australia ...*
as.many.as two airlines other REL PRF ACT-step-I Australia ...
“As as many as two other airlines who have flown Australia ...” SID 21402020

In these examples, *-i* functions as an applicative that adds a location as the ARG2 position with Theme being the ARG1. *-kan* on the other hand, functions as a causative that adds an Agent or Causer in ARG1 that acts on the Theme in ARG2 as it move towards the Goal in ARG3. This is an alternation pattern which is not widely discussed in the literature but was attested at a surprisingly high frequency in my dataset.

5.3.3 Dative Alternation

In this alternation (previously discussed in Section 2.5.2), both *-kan* and *-i* verbs take the same 3 arguments: Source, Recipient and Theme. For *-kan* ARG1 is the Source while ARG2 is a Theme and ARG3 is the Recipient marked with a prepositional phrase. In contrast, for *-i* verbs, ARG1 is a Source while ARG2 is the Recipient and ARG3 is a Theme, both marked by a noun phrase. This is summarized in Table 5.8 (taken from Section 2.5.2).

	ARG1	ARG2	ARG3
<i>meN-ROOT-kan</i>	SOURCE NP	THEME NP	RECIPIENT <i>ke/kapada/pada-PP</i>
<i>meN-ROOT-i</i>	SOURCE NP	RECIPIENT NP	THEME NP

Table 5.8: Dative Alternation

This alternation exists in the data and some of the roots that participate in this alternation are *menganugerahkan/menganugerahi* ‘to bestow’ and *menghadiahkan/menghadiahi* ‘to gift/present’ shown in (43) and (44).

- (43) a. *Universitas Bonn meng-anugerah-kan gelar “Doktor Honoris Causa”*
 Bonn University ACT-bestow-KAN title “Doctor Honorary Causa”
 kepada Pdt[sic].
 to pastor

“Bonn University bestowed the tile “Honorary Doctorate” to the Pastor. SID 20601010

- b. *Ia juga meng-anugerah-i Li gelar Pangeran Wuwei.*
 3SG also ACT-bestow-I Li title Prince Wuwei

“He also bestowed Li the title Prince Wuwei.” SID 20602019

- (44) a. *Perupa Arahmaiani meng-hadiahkan-kan perangkap tikus kepada Wakil*
 artist Arahmnaiani ACT-gift-KAN trap mouse to vice
 Ketua DPRD Provinsi .
 chairman DPRD province

“Arahmnaiani gifted a mousetrap to the vice chairman of the DPRD province.” 20901007

- b. Ford **meng-hadiah-i**=nya sebuah biola sebagai ungkapan terima kasih.
 Ford ACT-gift-I=3SG one-CLF violin as expression receive kindness

“Ford gifted him a violin as an expression of thanks.” SID 20902007

However, the dative alternation does not need to be between two morphological forms. The data also shows that it is possible to have a dative alternation for a single morphological form. In (46a) and (45a) we can see that *-kan* can also function like *-i* and vice versa. According to native speakers, while the sentences in (43) and (45) are all acceptable, there are certain preferred frames for each verb, especially in which argument directly proceeds the verb. Theme is the preferred argument immediately following a *-kan* while Recipient is the preferred argument immediately following an *-i* as in (43). This intuition is reflected in the corpus. When we consider the first 20 sentences with *menganugerahkan* in the corpus, five were monotransitive, eight sentences had the alternation with Recipient in ARG2 and Theme in ARG3 while seven sentences had the other alternation with Theme in ARG2 and Recipient in ARG3. In contrast, in the first 20 sentences with *menganugerahi*, 2 sentences were monotransitive, 16 sentences had the alternation with Recipient in ARG2 and Theme in ARG3 while only 2 sentences had the other alternation with Theme in ARG2 and Recipient in ARG3. This pattern also repeats for the root *hadiah* ‘gift’. In the first 20 sentences with *menghadiahkan* in the corpus, seven were monotransitive, four sentences had the alternation with Recipient in ARG2 and Theme in ARG3 while 10 sentences had the other alternation with Theme in ARG2 and Recipient in ARG3. In contrast, in the first 20 sentences with *menghadiahi*, three sentences were monotransitive, 14 sentences had the alternation with Recipient in ARG2 and Theme in ARG3 while only 2 sentences had the other alternation with Theme in ARG2 and Recipient in ARG3.

- (45) a. Huang Taiji **meng-anugerah-kan**=nya gelar Pangeran Gongshun.
 Huang Taiji ACT-bestow-KAN=3SG title Prince Gongshun

“Huang Taiji bestowed him the title Prince Gongshun. SID 20601019

- b. Pemerintah Pusat telah memberikan keputusan untuk **meng-anugerah-i**
 Government Central PRF give decision to ACT-bestow-I
status Kota sepenuhnya kepada Lembaga ini.
 status town full to institution this

“The central government has decided to bestow a full town status to this institution.”
 SID 20602011

- (46) a. Mama memilih untuk **meng-hadiahkan-kan**=nya kalung ini.
 mother choose to ACT-gift-KAN=3SG necklace this

“Mother chose to gift him/her this necklace” SID 20901004

- b. *Wasit I Nyoman Merta meng-hadiah-i kartu merah pada Nur Fatoni dan Agus Wiryo.*
 Wasit I Nyoman Merta ACT-gift-I red card to Nur Fatoni and
 Agus Wiryo

“Wasit I Nyoman Merta awarded a red card to Nur Fatoni and Agus Wiryo.” SID 20902011

5.3.4 Benefactive/Applicative Suffix Alternation

Another valence pattern was attested at a high frequency in my data and is shown in (47) and (48). In these two examples, there are no differences in the syntax or argument roles for both the *-kan* and *-i* verbs. The choice of the affix between *-kan* and *-i* determine the resulting meaning of the verb. Both *-kan* and *-i* take three arguments, the first being the Agent, the second the Recipient and the last argument the Theme all syntactically NPs. Both morphological forms share the same argument structure and semantic roles. The only alternation is in the choice of suffix.

According to native speaker intuition, the *-kan* forms have an implied meaning that the action is performed to the Recipient for their benefit (i.e. a Benefactive function) while the *-i* forms are applicative, incorporating the otherwise dative Recipient for the *meN-ø* verb as an ARG2. Other verbs in my data that follow this pattern are: *mengajari/mengajarkan* ‘to teach’, *menamai/menamakan* ‘to name’, and *menasehati/menasehatkan* ‘to advise’.

	ARG1	ARG2	ARG3
<i>meN-ROOT-kan</i>	AGENT NP	BENEFICIARY NP	THEME NP
<i>meN-ROOT-i</i>	AGENT NP	GOAL NP	THEME -

Table 5.9: Benefactive/Applicative Suffix Alternation

- (47) a. *Dia itu adalah wanita yang pertama kali meng-(k)irim-kan=ku SMS doa yang begitu indah.*
 3SG thet COP woman REL first time ACT-send-KAN=1SG SMS prayer
 REL SO beautiful

“She is the first woman to send me a prayer sms that is so beautiful” 11302005

- b. *Eunhyuk kemudian meng-(k)irim-i=mu SMS.*
 Eunhyuk then ACT-send-I=1SG SMS,
 “Eunhyuk then sent you an SMS.” SID 11303004

- (48) a. *Li langsung meng-(t)awar-kan=nya jabatan sebagai anggota staff=nya.*
 Li immediately ACT-offer-KAN=3SG position as member staff=3SG
 “Li immediately offered him a position as his staff member.” 10502020
- b. *SBY meng-(t)awar-i Irul posisi menteri ekonomi.*
 SBY ACT-offer-I Irul position minister economy
 “SBY offered Irul the Minister of Economy position.” SID 10503020

It could be argued that the ARG2 Recipient =*ku* in (47a) and *ku* in (47b) are metaphorical Goals for *sms* and so can be subsumed under the *-i* alternation pattern: Agent/Goal/Theme outlined in Section 2.5.1 as its variant. However, syntactically the ARG3 Theme in the alternation pattern discussed in Table 5.5 are PPs taking the preposition *dengan* while in (47a) and (47b) both ARG3s are NPs. It could be that two valences are available for the root and lexically encoded. Ultimately more sentences need to be annotated to fully explore this. The root can also be semantically annotated to determine if the each pattern is associated with a different sense of the root.

5.3.5 Experiencer-Stimulus Alternation

One alternation that was attested at high frequencies in my data that has not been widely described in literature is the Experiencer-Stimulus alternation. Looking at the results presented here (from Table 4.7 in Chapter 4), there is a clear tendency for *-kan* verbs to have Stimulus as ARG1 while *-i* verbs to have Experiencer as ARG1 as summarized in Table 5.11. *kagum* ‘admire’ (49) and *senang* ‘pleasure’ (50) are two roots that participate in this alternation.

Table 5.10: Experiencer and Stimulus Roles by Group

Surface Form	gloss	ARG1	ARG2
<i>mengagumkan</i>	‘to fascinate’	STIMULUS	EXPERIENCER
<i>mengagumi</i>	‘to admire’	EXPERIENCER	STIMULUS
<i>menyenangkan</i>	‘to please’	STIMULUS	EXPERIENCER
<i>menyenangi</i>	‘to enjoy’	EXPERIENCER	STIMULUS

Table 5.11: Stimulus-Experiencer Alternation

- (49) a. *Upacara mereka amat meng-(k)agum-kan=ku.*
 Ceremony 3PL very ACT-amaze-KAN=1SG
 “Their ceremony amazed me.” SID 20501002
- b. *Said Nursi juga sangat meng-(k)agum-i Maulana Rumi.*
 Said Nursi also very ACT-amaze-I Maulana Rumi
 “Said Nursi also admires Maulana Rumi very much.” SID 20502001
- (50) a. *Tidak ada hal lain yang lebih meny-(s)enang-kan=ku.*
 NEG COP thing other REL more ACT-please-CAUS=1SG
 “Nothing else pleases me more.” SID 22001020
- b. *Anda selalu meny-(s)enang-i hal-hal baru.*
 2SG always ACT-please-I thing.REDUP new
 “You always like new things.” SID 22002020

There are also roots where the *-kan* and *-i* forms do not alternate and instead follow one pattern or the other. Some examples have been listed in Table 5.12. The root *celaka* ‘woe’ form *-kan* and *-i* verbs with ARG1 and ARG2 as Stimulus and Experiencer respectively following the *-kan* alternation pattern in Table 5.11 and seen in (51). The roots *tentram* ‘calm’ and *bingung* ‘confused’ both do not have attested *-i* forms and so follow the *-kan* alternation pattern without and *-i* counterpart ((52), (53))

Surface Form	gloss	ARG1	ARG2
<i>mencelakakan</i>	‘to harm’	STIMULUS	EXPERIENCER
<i>mencelakai</i>	‘to harm’	STIMULUS	EXPERIENCER
<i>menentramkan</i>	‘to calm’	STIMULUS	EXPERIENCER
<i>membingungkan</i>	‘to confuse’	STIMULUS	EXPERIENCER
<i>menyesalkan</i>	‘to regret’	EXPERIENCER	STIMULUS
<i>menyesali</i>	‘to regret’	EXPERIENCER	STIMULUS

Table 5.12: Stimulus-Experiencer Valence

- (51) a. *Yang paling parah, Martono bisa **men-celaka-kan**=ku.*
REL most severe, Martono can ACT-harm-CAUS=1SG
‘‘At worst, Martono can harm me.’’ SID 21601005
- b. *Ia lalu pergi ke dukun untuk **men-celaka-i** Rusnan dan Indri.*
3SG then go to healer for ACT-harm-I Rusnan and Indri
‘‘He then went to the healer to harm Rusnan and Indri.’’ SID 21602006

(52) Note: *-i* form is not attested in the data or KBBI

- a. *Kalimat Thayyibah **men-(t)entram-kan**=nya*
Sentence Thayyibah ACT-calm-CAUS=1SG
‘‘Thayyibah’s words calmed him/her’’ SID 50301005

(53) Note: *-i* form is not attested in the data or KBBI

- a. *Lita itu adalah gadis yang membanggakan sekaligus **mem-bingung-kan**=nya*
Lita that COP girl REL cause.to.be.proud simultaneously ACT-confuse-CAUS-3SG
‘‘Lita is a girl that simultaneously makes him/her proud and confuses him/her’’ SID 51801013

On the other hand, the root *sesal* ‘regret’ form *-kan* and *-i* verbs with ARG1 and ARG2 as Stimulus and Experiencer respectively following the *-i* alternation pattern as seen in (54).

- (54) a. *Kami sangat **meny-(s)esal-kan** sikap pihak maskapai karena mereka baru*
1PLINCL very ACT-regret-CAUS attitude party airlines because 3PL just
mengumumkan penerbangan ditunda setelah beberapa jam kami
announce flight was.delayed after several hour 1PLINCL
menunggu di bandara.
wait LOC airport
‘‘We deeply regret the attitude of the airline because they just announced the flight was delayed after several hours of waiting at the airport.’’ SID 10802007
- b. *Namun, membiarkan masalah atau orang ketiga masuk dalam hubungan*
However, leave problem or person third enter into relationship
*maka Anda akan **meny-(s)esali**=nya.*
then 2SG FUT ACT-regret-I=3SG
‘‘However, allowing a problem or third person to enter the relationship will make you regret it.’’ SID 10803020

The roots that participate in these alternations have a common semantic characteristic of being psych verbs that express some psychological state such as *confuse*, *calm* and *please* and take an Experiencer as one of their arguments (Primus, 2004, p.377).

5.3.6 Causative/Applicative Alternation

This alternation previously outlined in Section 2.5.3 and summarized in Table 5.13 (from Table 2.6) was also attested in my data. A prototypical verb that undergoes this alternation is *menikahkan/menikahi* ‘to marry’. In this alternation, *-kan* introduces an external Causer as the ARG1 while *-i* verbs may or may not introduce a new argument. If it does increase the valency the intransitive verb becomes a transitive verb. If the valency is not increased, the addition of *-i* usually changes the aspectual meaning of the verb. Some examples of verbs that participate in the same alternation in my data set include: *mengenalkan/mengenali* ‘to know’ *mencampurkan/mencampuri* ‘to mix’ *mendasarkan/mendasari* ‘to base (something)’ *menggenapkan/menggenapi* ‘to fulfil’.

	ARG1	ARG2	ARG3
<i>meN-ROOT-ø</i>	AGENT/THEME NP	- -	- -
<i>meN-ROOT-i</i>	AGENT/THEME NP	THEME/PATIENT NP	- -
<i>meN-ROOT-kan</i>	CAUSER NP	AGENT NP	THEME/PATIENT NP

Table 5.13: Causative/Applicative Alternation

In the examples (55),(56) and (57) below, verbs derived with *-i* have two arguments while verbs derived with *-kan* have three arguments, with the external Causer as ARG1. The ARG3 arguments are also marked by a PP (e.g. *pada/kepada*-PP)

- (55) a. *Saya sangat meng-(k)enal=nya.*
1SG very ACT-know=3SG
“I know him/her very well”. SID 10101009
- b. *kita akan segera meng-(k)enal-i=nya*
1PLINCL FUT immediately ACT-know-I=3SG
“We will know him/her/it immediately”. SID 10103007
- c. *Dia meng-(k)enal-kan=ku kepada seorang teman=nya*
3SG ACT-know-KAN=1SG to one.CLF friend=3SG
“S/he introduced me to a friend of his/hers”. SID 10102001

- (56) a. *Harga monitor itu hampir **men-dekat** harga LCD monitor yang digunakan*
 price monitor DEF almost ACT-near-KAN price LCD monitor REL used.by
televisi studio.
 television studio

“The price of the monitor almost nears the price of the LCD monitor used by the television studio”. SID 12201015

- b. *Warta Kota dilarang **men-dekat-i** kantor KAB*
 News City prohibited ACT-near-I office KAB

“City News is prohibited from nearing the KAB office”. SID 12203008

- c. *Facebook **men-dekat-kan=nya** pada teman dan keluarga*
 Facebook ACT-near-KAN=3SG to friend and family

“Facebook brings him near to his friends and family”. SID 12202016

- (57) a. *Walikota, menurut Avian, harus memikirkan hal-hal yang sangat*
 mayor, according.to Avian, must think things.REL very
men-dasar.
 ACT-base

“The mayor, according to Avian must think of things that are very basic”. SID 11101008

- b. *Integritas adalah suatu elemen karakter yang **men-dasar-i** timbulnya*
 integrity COP one element character REL ACT-base-I emergence
pengakuan profesional.
 recognition professional

“Integrity is the one character element underlies the emergence of professional recognition.” SID 11103005

- c. *Ia **men-dasar-kan=nya** pada hadits yang cukup populer*
 3SG ACT-base-KAN=3SG to hadith REL enough popular

“S/he based it on hadiths which were quite popular”. SID 11102006

The annotation task designed in Chapter 3 was adequate enough to identify some of the roots that follow these alternations. Annotating more sentences for these roots that show this alternation will allow us to explore it in more detail. In particular we can use it to statistically determine just how productive these alternations are, and if any one surface form show this alternation to a greater extent.

5.4 Comparing the results with KBBI

As mentioned in 2.1.1, the *Kamus Besar Bahasa Indonesia* (KBBI) (Amalia, 2016) lists roots and all surface forms associated with that root in their entries. These surface forms were determined by lexicographers and there are some examples included to illustrate their use. I compared the 132 roots annotated from the corpus against the dictionary to identify the percentage of overlap between the surface forms found in the dictionary. While most of the surface forms described in the dictionary were attested in the data, 27 out of 140 roots (20%) from the corpus had surface forms that were different from what the dictionary described. There were three different types of discrepancies as listed below:

1. **KBBI only:** A surface form is listed in the KBBI entry for the root but there were no instances of the surface form in the data.
2. **Corpus only:** A surface form is attested at least 60 times in the data but it was not listed in the KBBI entry for the root.
3. **Not verb:** A surface form is attested at least 60 times in the data but it not listed as a verb in KBBI.

Table 5.14 lists these roots and the type of discrepancy they have with the KBBI.

The findings from this comparison can be useful in informing ourselves and the KBBI of how these verbs in Indonesian are used differently in the corpus from the dictionary. In particular, the surface forms which have been attested in the corpus but not in KBBI can be added to the KBBI to further enrich this resource. This information can also be useful in building Indonesian lemmatizers in future work. Also, while KBBI does not list frequency, if they wanted to they could note the forms that have not been attested in the corpus as rare or infrequent.

No.	Root	gloss	Database Group	KBBI Group	Discrepancy Type
1	<i>jilat</i>	‘lick’	NkI	Nki	Corpus only
2	<i>hukum</i>	‘punish’	NkI	Nki	Corpus only
3	<i>pungut</i>	‘pick’	NkI	NKI	KBBI only
4	<i>lawat</i>	‘visit’	NkI	Nki	Corpus only
5	<i>pegang</i>	‘hold’	NkI	Nki	Corpus only
6	<i>tunggang</i>	‘mount’	NkI	Nki	Corpus only
7	<i>wawancara</i>	‘interview’	NkI	nKI	Corpus only
8	<i>jangkit</i>	‘infect’	NkI	nkI	KBBI only
9	<i>dalang</i>	‘puppeteer’	NkI	NKI	KBBI only
10	<i>tanggap</i>	‘percieve’	NkI	NKI	KBBI only
11	<i>organisasi</i>	‘organize’	NKi	Nki	Corpus only
12	<i>hitam</i>	‘black’	NKi	NKI	Corpus only
13	<i>realisasi</i>	‘realization’	NKi	Nki	Corpus only
14	<i>singkir</i>	‘set aside’	NKi	NKI	KBBI only
15	<i>uap</i>	‘steam’	NKi	NKI	Corpus only
16	<i>pulang</i>	‘return’	nKi	nKI	KBBI only
17	<i>gencar</i>	‘intensify’	nKi	NKi	KBBI only
18	<i>kokoh</i>	‘strong’	nKi	nKI	KBBI only
19	<i>gaul</i>	‘mix’	nkI	NKI	KBBI only
20	<i>akal</i>	‘wit, cunning’	nkI	nKI	KBBI only
21	<i>khianat</i>	‘betray’	nkI	NkI	KBBI only
22	<i>selimut</i>	‘blanket’	nkI	nKI	KBBI only
23	<i>gandrung</i>	‘love’	nkI	nKI	KBBI only
24	<i>cabul</i>	‘lewd’	nkI	NkI	KBBI only
25	<i>tekun</i>	‘diligent’	nkI	nKI	KBBI only
26	<i>miris</i>	‘devastating’	nkI	Adjective	Not verb
27	<i>lesu</i>	‘sluggish’	Nki	Adjective	Not verb

Table 5.14: Comparing roots with KBBI

PART III:

SUMMARY

Chapter 6

Conclusion

6.1 Overview

This chapter concludes the thesis by summarizing the general findings of the study in Section 6.2 and outlining the new resources available in Section 6.3. I will also discuss some of the limitations and ways to overcome them in future work in Section 6.4.

6.2 Summary

This study is a corpus based analysis of *-kan* and *-i* morphemes in Indonesian. *-kan* and *-i* have been often discussed together in literature as verbal suffixes that can attach to a stem to change its argument structure by changing the valency of the resulting verb. They seem to function similarly in some verbs but produce completely different meanings in other verbs. This suggests that the information on which verbs or classes of verbs take which affixes, and more importantly, why a verb does or does not take an affix are stored in the lexicon. This suggests that natural classes form based on the morphological behavior of verbs and their affixes. However, this has not been addressed systematically in the literature. To address this, I surveyed the biggest available corpus of standard written Indonesian to identify all verbs associated with *-kan* and *-i*.

The LCC Corpus containing 15M sentences of written Standard Indonesian from various genres provided the source of data for this study. A morphological dictionary called MALINDO Morph morphological dictionary was used as a tool together with a python script I made to identify all verbs and the surface forms in the corpus. I chose to study a subset of the verbs namely verbs in the active voice prefixed by *meN-*. The verbs were grouped according to their morphological profile for the feature *-kan* *-i* and $\emptyset$ using the selection criteria of minimum frequency. The verbs were sorted into 7 mutually exclusive groups. The same python program randomly selected 20 roots for each group in the corpus. The verbs that were grouped along with the sentences were annotated for their syntactic arguments (i.e. ARG1, ARG2, ARG3, ...etc) and semantic roles (i.e. Agent, Patient, Theme, ...etc) using a role list curated by me based on

the literature and influenced by Sheinflux et al. (2017). 140 roots were selected for annotation. These roots and their sentences were exported as an annotatable SQLite database using the NTU-MC Tagging interface (Bond et al., 2015). The annotated database was processed using another python script and the frequencies of the arguments and roles were calculated.

The results of the annotation process generally corroborate what we know from the literature. Both *-kan* and *-i* function as a valency increasing suffix and the data shows that indeed when we contrast the presence of ARG2 (O argument) in *-kan* and *-i* verbs with $\emptyset$ verbs, we can see that the percentage of sentences having an ARG2 is much higher than ARG1 in the former groups than the latter. In addition, the proportion of verbs having at least 3 argument roles for *-kan* verbs is nearly three times more than *-i* verbs. This is consistent with what we know *-kan* functions as, which is a causative, adding a Causer role.

Looking into the semantic annotation, the most prominent roles for all the verbs with *-kan* *-i* and $\emptyset$ were Agent in ARG1, and Patient and Theme in ARG2. Other major roles associated with *-kan* were Cause in ARG1 and Goal in ARG3 and Experiencer in ARG2 while Stimulus in ARG1. The major roles associated with *-i* verbs were Location and Recipient in ARG2. These findings are also in line with what is expected in the literature.

I also discussed some alternation patterns that were mentioned in the literature and presented verb subcategorization frames for groups of verbs with both *-kan* and *-i* that were found in the corpus.

6.3 Newly available resources

This work produced new resources that are available to view or use in research. Firstly, the open-source code used to read the corpus and database are available for download and use from GitHub, a web-based hosting service using a version control system for tracking changes in computer files. The repository can be found here <https://github.com/hannahchoiyj/meng-kan-i>. A list of the files available can be found in Appendix A.

The LCC Corpora was used to create a database of sentences with *meN-* verbs that take *-kan* or *-i* as well as $\emptyset$ - bare verbs with *-kan* or *-i*. A total of 4,902 sentences of the database have been annotated. The list of annotated verbs and their meanings can be found in Appendix B

The results of the annotation is also provided in the GitHub repository. They contain frequencies of the syntactic arguments and the semantic roles that were annotated for each group. In addition to that, Appendix C contains the list of subcategorization frames for three argument sentences in each group along with their SIDs. This will be a valuable resource for researchers interested in looking specifically at alternation patterns for *-kan* and *-i* with corpus evidence and examples.

6.4 Limitations and future work

Throughout this preliminary study I have identified some of the limitations in Chapters 3,4 and 5. Here I discuss some ways to improve the study for the future.

One of the most limiting aspects of the study was the sampling method and size. Randomly sampling 140 roots from thousands of roots found with minimal selection criteria led to some important roots well discussed in the literature that were missed out. For example, the Benefactive alternation for roots such as *waris* ‘inherit/bequath’ and *pinjam* ‘borrow/lend’ were not annotated and I was unable to explore them. In future work I will focus the study to known semantic clusters in most languages, such as *break* verbs or *borrow* verbs.

Closely related to this, the verbs that were annotated for syntactic and semantic roles were not annotated for meaning. This posed a problem in dealing with multiple senses or homophonous roots which could have different alternation patterns. In future work I will semantically annotate the meaning of the root by adding Wordnet synsets, which will provide a rich source of lexical information that can better inform my analysis. Aside from annotating the meaning of the verb, I can also annotate whether or not a verb is stative, durative, progressive, etc because that can provide another layer of information to why verbs behave the way that they do.

Finally, I would like to increase the coverage of the morphemes being studied. In particular I would like to contrast the behavior of *-kan* and *-i* morphemes as it interacts with the passive marker *di-* or other verbal markers that are often discussed in literature together with *meN-* such as *ber-*, *ter-*, *memper-* and *diper-*. The availability of a verb to be passivized is important as it changes the argument structure of the verb and it should be considered when exploring the functions of *-kan* and *-i*. In it will be interesting to explore the morphological verb classes that emerge when we consider all of these affixes at the same time.

Other future work will also include refining the semantic role list by collapsing some roles that are not as relevant for annotating arguments in Indonesian sentences such as Degree, Duration, Extent and Purpose.

Despite the limitations, this study is the first quantitative study of *-kan* and *-i* semantic patterns. Further, it provides the ground work for future studies. The findings of this large-scale corpus-based study also present a different angle of looking into the study of *-kan* and *-i*.

PART IV:

BIBLIOGRAPHY AND APPENDICES

Bibliography

- A. Jones, Douglas, Robert Berwick, Franklin Cho, Zeeshan Khan, Karen T. Kohl, Naoyuki Nomura, Ulrich Sauerl & Brian Ulicny. 1996. Verb classes and alternations in bangla, german, english, and korean .
- Abas, Husen. 1987. *Indonesian as a unifying language of wider communication: a historical and sociolinguistic perspective* (Pacific Linguistics Series D 73). Canberra: The Department of Linguistics, Research School of Pacific Studies.
- Adelaar, K Alexander & David John Prentice. 1996. Malay: its history, role and spread. *Atlas of languages of intercultural communication in the Pacific, Asia and the Americas* 2(1).
- Alwi, Hasan, Soenjono Dardjowidjojo, Hans Lapoliwa & Anton M. Moeliono. 2014. *Tata Bahasa Baku Bahasa Indonesia*. Jakarta: Balai Pustaka 3rd edn.
- Amalia, Dora (ed.). 2016. *Kamus Besar Bahasa Indonesia*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa 5th edn.
- Arka, I Wayan, Mary Dalrymple, Meladel Mistica, Surriel Mofu, Avery Andrews, Jane Simpson, Miriam Butt & Tracy Holloway King. 2009. A linguistic and computational morphosyntactic analysis for the applicative-i in indonesian. In *Proceedings of the lfg09 conference*, 85–105.
- Asmah, O. 2001. The malay language in malaysia and indonesia: From lingua franca to national language. *The Asianists ASIA* 2.
- Baker, Mark C. 1988. Incorporation: A theory of grammatical function changing .
- Berk, Lynn M. 1999. *English syntax: from word to discourse*. Oxford University Press Oxford.
- Bird, S, E Klein & E Loper. 2009. Nltk book.
- 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.
- Bond, Francis, Luís Morgado da Costa & Tuấn Anh Lê. 2015. IMI — a multilingual semantic annotation environment. In *ACL-2015 system demonstrations*, .
- Cole, Peter, Gabriella Hermon & Yanti. 2008. Voice in Malay/Indonesian. *Lingua* 118(10). 1500–1553.
- Cole, Peter & Min-Jeong Son. 2004. The argument structure of verbs with the suffix-kan in indonesian. *Oceanic Linguistics* 43(2). 339–364.
- DeLancey, Scott. 1995. Verbal case frames in english and tibetan. *unpublished ms., Department of Linguistics, University of Oregon, Eugene, OR* .

- Dixon, Robert MW. 1991. *A new approach to english grammar, on semantic principles*. Oxford University Press.
- Dixon, Robert MW. 1992. *A new approach to english grammar, on semantic principles*. Oxford University Press.
- of Education, Ministry, Institute of Language Development Culture & Cultivation. 2018. Kamus besar bahasa indonesia daring. URL <https://kbbi.kemdikbud.go.id/>.
- Ewing, Michael C. 2005. Colloquial indonesian. *The Austronesian languages of Asia and Madagascar* 7. 227.
- Fam, Rashel, Andry Luthfi, Arawinda Dinakaramani & Ruli Manurung. 2014. Building an indonesian rule-based part-of-speech tagger. In *2014 international conference on asian language processing (ialp)*, 70–73. doi: 10.1109/IALP.2014.6973521.
- Fillmore, Charles. 1970. The grammar of hitting and breaking. *Readings in English transformational grammar* 120–133.
- Fillmore, Charles J, Christopher R Johnson & Miriam RL Petruck. 2003. Background to framenet. *International journal of lexicography* 16(3). 235–250.
- Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Lrec*, vol. 29, 31–43.
- Hale, Ken & Jay Keyser. 1986. Some transitivity alternations in english. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"* 20(3). 605–638.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank (eds.). 2013. *Glottolog 2.4*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL <http://glottolog.org/>.
- Hipp, Kennedy D. Mistachkin J., D. R. 2010. Hadoop. URL <https://www.sqlite.org/src/info/2ef4f3a5b1d1d0c4>.
- Jackendoff, Ray. 1992. *Semantic structures*, vol. 18. MIT press.
- Jaworski, Wojciech & Adam Przepiórkowski. 2014. Syntactic approximation of semantic roles. In *International conference on natural language processing*, 193–201. Springer.
- 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.
- Kridalaksana, Harimurti. 1989. *Pembentukan kata dalam bahasa indonesia*. Gramedia.

- Kroeger, Paul. 2007. Morphosyntactic vs. morphosemantic functions of indonesian-kan. In *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan* 229–251.
- Kroeger, Paul R. 2010. The grammar of hitting, breaking, and cutting in kimaragang Dusun. *Oceanic Linguistics* 1–20.
- 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.
- Leipzig Corpora Collection. 2019. Terms of usage. URL <http://wortschatz.uni-leipzig.de/en/usage>.
- Levin, B. & M.R. Hovav. 2005. *Argument realization* Argument Realization. Cambridge University Press. URL <https://books.google.com.sg/books?id=msi9a50gHVYC>.
- Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.
- Levin, Beth. 2011. Verb classes within and across languages. *Handout, Leipzig, April*.
- Levin, Beth & Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *cognition* 41(1-3). 123–151.
- Lewis, M. Paul. 2009. *Ethnologue: Languages of the World*. Dallas, Texas: SIL International 16th edn. URL <http://www.ethnologue.com>.
- Macdonald, R. Ross. 1976. *Indonesian Reference Grammar*. Washington: Georgetown University Press 2nd edn.
- Macdonald, Roderick Ross & Soenjono Darjowidjojo. 1967. *A student's reference grammar of modern formal indonesian*. Georgetown University Press.
- Ministry of Education and Culture. 2019. Institute of language development and cultivation. URL <http://badanbahasa.kemdikbud.go.id/lamanbahasa/sejarah>.
- Ministry of Education and Culture, Institute of Language Development and Cultivation. 2015. Indonesian general spelling guidelines. URL badanbahasa.kemdikbud.go.id/lamanbahasa/sites/default/files/PUEBI.pdf.
- Moeljadi, David. 2018. *An indonesian resource grammar (indra) and its application to a tree-bank (jati)*. Singapore: Nanyang Technological University PhD dissertation.
- Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018a. Reclassification of the leipzig corpora collection for malay and indonesian. *NUSA: Linguistic studies of languages in and around Indonesia* (65). 47–66.

- Nomoto, Hiroki, Hannah Choi & David Moeljadi. 2018b. Malindo morph: Morphological dictionary and analyser for malay/indonesian. In *The 13th workshop on asian language resources*, 36–43. Miyazaki.
- Nur Amirah Binte Khairul Anuar, Hannah Choi & František Kratochvíl. 2017. Verb subcategorization: -kan and -i suffixing verbs in Malay and Indonesian. Presented at The Twenty-First International Symposium On Malay/Indonesian Linguistics (ISMIL 21).
- Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics* 31(1). 71–106.
- Payne, Thomas. 2006. *Exploring language structure: A student's guide*. Cambridge University Press.
- Pisceldo, Femphy, Mirna Adriani & Ruli Manurung. 2009. Probabilistic part of speech tagging for bahasa indonesia .
- Primus, Beatrice. 2004. Protorollen und verbttyp: Kasusvariaton bei psychischen verben. *Semantische Rollen* 377–401.
- Quasthoff, Uwe, Matthias Richter & Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on language resources and evaluation*, vol. 17991802, 21.
- Schuler, Karin Kipper. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon .
- Sheinflux, Livnat Herzig, Nurit Melnik & Shuly Wintner. 2017. Representing argument structure 1. *Journal of Linguistics* 53(4). 701–750.
- Sneddon, James Neil. 1996. *Indonesian: A comprehensive grammar*. London/New York: Routledge.
- 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.
- Sneddon, James Neil, K Alexander Adelaar, Dwi N Djenar & Michael Ewing. 2012. *Indonesian: A comprehensive grammar*. Routledge.
- Suryadinata, L., E.N. Arifin & A. Ananta. 2003. *Indonesia's population: Ethnicity and religion in a changing political landscape* Indonesia's population series. Institute of Southeast Asian Studies. URL <https://books.google.ca/books?id=nFckUneBbRIC>.
- Tesnière, Lucien. 1959. *Eléments de syntaxe structurale* .

- Vamarasi, Marit Kana et al. 1999. *Grammatical relations in bahasa indonesia*. Department of Linguistics, Research School of Pacific Studies, The
- Voskuil, Jan E. 2000. Indonesian voice and A-bar movement. In *Formal issues in austronesian linguistics*, 195–213. Springer.
- Voskuil, Jan Eckhard. 1996. *Comparative morphology: verb taxonomy in indonesian, tagalog and dutch*, vol. 21. Holland Academic Graphics.

Appendix A

Source code

The open-source code used to read the corpus and database are available for download and use from GitHub, a web-based hosting service using a version control system for tracking changes in computer files: <https://github.com/hannahchoiyj/meng-kan-i>

The files available in this repository are:

1. `corpus_crawl.py` is code to search the LCC to identify verbal surface forms, their underlying roots and attested sentences. It then sorts the surface forms by morphological type (i.e. the number and type of affixes they take) and chooses a customizable number of random roots that meet the selection criteria. Produces a `master.tsv` file that contains the roots, surface forms and sentences.
2. `corpus_checks.py` is code to read annotated LCC corpus of meN- verbs with -kan and -i suffixes in Indonesian. It takes a SQLite database file and produces the result of the annotation (i.e. the frequencies of the roles and arguments annotated) into tables that can be viewed on a browser. It also produces LaTeX compatible versions of the tables.
3. `results.html` Tables containing the frequencies for the Arguments and Roles annotated that can be viewed on a browser.
4. `results_sub.html` Tables containing the frequencies for the Roles annotated that can be viewed on a browser

The annotated database and text file containing the sentences in the database will be made available on <http://compling.hss.ntu.edu.sg/who/hannah/> for download

Appendix B

List of Verbs Annotated

Table B.1: List of Annotated Group 1 Verbs (NKI)

Starting SID	Root	Surface form	Gloss
10101000	<i>kenal</i>	<i>mengenal</i>	‘know’
10102000	<i>kenal</i>	<i>mengenalkan</i>	‘introduce’
10103000	<i>kenal</i>	<i>mengenali</i>	‘recognize’
10201000	<i>seberang</i>	<i>menyebrang</i>	‘cross’
10202000	<i>seberang</i>	<i>menyebrangkan</i>	‘make someone cross’
10203000	<i>seberang</i>	<i>menyebrangi</i>	‘cross’
10301000	<i>langkah</i>	<i>melangkah</i>	‘step’
10302000	<i>langkah</i>	<i>melangkahkan (kaki)</i>	‘step’
10303000	<i>langkah</i>	<i>melangkahi</i>	‘step over’
10401000	<i>ajar</i>	<i>mengajar</i>	‘teach’
10402000	<i>ajar</i>	<i>mengajarkan</i>	‘teach something’
10403000	<i>ajar</i>	<i>mengajari</i>	‘teach someone’
10501000	<i>tawar</i>	<i>menawar</i>	‘offer’
10502000	<i>tawar</i>	<i>menawarkan</i>	‘offer’
10503000	<i>tawar</i>	<i>menawari</i>	‘offer’
10601000	<i>ikut</i>	<i>mengikut</i>	‘follow’
10602000	<i>ikut</i>	<i>mengikutkan</i>	‘cause someone to join’
10603000	<i>ikut</i>	<i>mengikuti</i>	‘follow someone’
10701000	<i>jauh</i>	<i>menjauh</i>	‘distance’
10702000	<i>jauh</i>	<i>menjauhkan</i>	‘distance something from somewhere’
10703000	<i>jauh</i>	<i>menjauhi</i>	‘distance somewhere’
10801000	<i>sesal</i>	<i>menyesal</i>	‘regret’
10802000	<i>sesal</i>	<i>menyesalkan</i>	‘regret something/someone’
10803000	<i>sesal</i>	<i>menyesali</i>	‘regret something’
10901000	<i>siram</i>	<i>menyiram</i>	‘pour’

10902000	<i>siram</i>	<i>menyiramkan</i>	‘pour something ’
10903000	<i>siram</i>	<i>menyirami</i>	‘pour something ’
11001000	<i>kena</i>	<i>mengena</i>	‘achieve’
11002000	<i>kena</i>	<i>mengenakan</i>	‘cause to acieve’
11003000	<i>kena</i>	<i>mengenai</i>	‘hit’
11101000	<i>dasar</i>	<i>mendasar</i>	‘based on’
11102000	<i>dasar</i>	<i>mendasarkan</i>	‘based on’
11103000	<i>dasar</i>	<i>mendasari</i>	‘based on’
11201000	<i>tular</i>	<i>menular</i>	‘contagious’
11202000	<i>tular</i>	<i>menularkan</i>	‘infect something on someone’
11203000	<i>tular</i>	<i>menulari</i>	‘infect someone with something’
11301000	<i>kirim</i>	<i>mengirim</i>	‘send’
11302000	<i>kirim</i>	<i>mengirimkan</i>	‘send for someone’
11303000	<i>kirim</i>	<i>mengirimi</i>	‘send to someone’
11601000	<i>turun</i>	<i>menurun</i>	‘descend’
11602000	<i>turun</i>	<i>menurunkan</i>	‘lower something down’
11603000	<i>turun</i>	<i>menuruni</i>	‘lower something down’
11701000	<i>campur</i>	<i>mencampur</i>	‘mix’
11702000	<i>campur</i>	<i>mencampurkan</i>	‘mix something with/for something/one’
11703000	<i>campur</i>	<i>mencampuri</i>	‘mix something with something’
11801000	<i>padat</i>	<i>mamadat</i>	‘condense’
11802000	<i>padat</i>	<i>memadatkan</i>	‘cause to be compact’
11803000	<i>padat</i>	<i>mamadati</i>	‘pack close together’
11901000	<i>urus</i>	<i>mengurus</i>	‘take care of’
11902000	<i>urus</i>	<i>menguruskan</i>	‘take care of’
11903000	<i>urus</i>	<i>mengurusi</i>	‘take care of’
12001000	<i>nikah</i>	<i>menikah</i>	‘marry’
12002000	<i>nikah</i>	<i>menikahkan</i>	‘marry someone to someone’
12003000	<i>nikah</i>	<i>menikahi</i>	‘marry someone’
12101000	<i>jejak</i>	<i>menjejak</i>	‘step’
12102000	<i>jejak</i>	<i>menejejakkan</i>	‘step ’
12103000	<i>jejak</i>	<i>menjejaki</i>	‘step’
12201000	<i>dekat</i>	<i>mendekat</i>	‘draw near to’
12202000	<i>dekat</i>	<i>mendekatkan</i>	‘cause to become close’
12203000	<i>dekat</i>	<i>mendekati</i>	‘approach’

Table B.2: List of Annotated Group 2 Verbs (nKI)

Starting SID	Root	Surface form	Gloss
20101000	<i>maklum</i>	<i>memaklumkan</i>	‘to cause something to be understood’
20102000	<i>maklum</i>	<i>memaklumi</i>	‘to understand/know/admit something’
20201000	<i>tanya</i>	<i>menanyakan</i>	‘to ask something to someone’
20202000	<i>tanya</i>	<i>menanyai</i>	‘to ask someone’
20301000	<i>cemar</i>	<i>mencemarkan</i>	‘to cause someone’s something to be ruined’
20302000	<i>cemar</i>	<i>mencemari</i>	‘to ruin someone’
20401000	<i>kerja</i>	<i>mengerjakan</i>	‘to do something’
20402000	<i>kerja</i>	<i>mengerjai</i>	‘to do someone’
20501000	<i>kagum</i>	<i>mengagumkan</i>	‘to cause someone to admire something’
20502000	<i>kagum</i>	<i>mengagumi</i>	‘to admire someone’
20601000	<i>anugerah</i>	<i>menganugerahkan</i>	‘to honor something to someone’
20602000	<i>anugerah</i>	<i>menganugerahi</i>	‘to honor someone’
20701000	<i>tidur</i>	<i>menidurkan</i>	‘to lay someone on something’
20702000	<i>tidur</i>	<i>meniduri</i>	‘to lay with someone (biblical sense)’
20801000	<i>nafkah</i>	<i>menafkahkan</i>	‘to cause something to provide for someone’
20802000	<i>nafkah</i>	<i>menafkahi</i>	‘to provide for someone’
20901000	<i>hadiah</i>	<i>menghadiahkan</i>	‘to gift something to someone’
20902000	<i>hadiah</i>	<i>menghadiahi</i>	‘to gift someone something’
21001000	<i>nama</i>	<i>menamakan</i>	‘to name something as something’
21002000	<i>nama</i>	<i>menamai</i>	‘to name something’
21101000	<i>amin</i>	<i>mengaminkan</i>	‘to say ”amen” to something’
21102000	<i>amin</i>	<i>mengamini</i>	‘to say ”amen” to something’
21201000	<i>genap</i>	<i>menggenapkan</i>	‘round out’
21202000	<i>genap</i>	<i>menggenapi</i>	‘fulfil’
21301000	<i>terang</i>	<i>menerangkan</i>	‘explain’
21302000	<i>terang</i>	<i>menerangi</i>	‘illuminate’
21401000	<i>terbang</i>	<i>menerbangkan</i>	‘fly something’
21402000	<i>terbang</i>	<i>menerbangi</i>	‘fly over something’
21501000	<i>curiga</i>	<i>mencurigakan</i>	‘suspicious’
21502000	<i>curiga</i>	<i>mencurigai</i>	‘be suspicious of something’
21601000	<i>celaka</i>	<i>mencelakakan</i>	‘cause harm’
21602000	<i>celaka</i>	<i>mencelakai</i>	‘cause harm’
21701000	<i>teriak</i>	<i>meneriakkan</i>	‘shout something’
21702000	<i>teriak</i>	<i>meneriaki</i>	‘shout at’
21801000	<i>temu</i>	<i>menemukan</i>	‘find something’
21802000	<i>temu</i>	<i>menemui</i>	‘meet someone’
21901000	<i>nasihat</i>	<i>menasihatkan</i>	‘counsel or advise’
21902000	<i>nasihat</i>	<i>menasihati</i>	‘counsel or advise’
22001000	<i>senang</i>	<i>menyenangkan</i>	‘cause to please’
22002000	<i>senang</i>	<i>menyenangi</i>	‘enjoy or like’

Table B.3: List of Annotated Group 3 Verbs (NkI)

Starting SID	Root	Surface form	Gloss
30101000	<i>jilat</i>	<i>menjilat</i>	‘lick’
30102000	<i>jilat</i>	<i>menjilati</i>	‘lick (iter)’
30201000	<i>hukum</i>	<i>menghukumi</i>	‘punish’
30202000	<i>hukum</i>	<i>menghukumi</i>	‘punish (iter)’
30301000	<i>pungut</i>	<i>memungut</i>	‘pick up’
30302000	<i>pungut</i>	<i>memunguti</i>	‘pick up (iter)’
30401000	<i>tebang</i>	<i>menebang</i>	‘cut down’
30402000	<i>tebang</i>	<i>menebangi</i>	‘cut down (iter)’
30501000	<i>lawat</i>	<i>melawat</i>	‘pass’
30502000	<i>lawat</i>	<i>melawati</i>	‘pass (prog)’
30601000	<i>pegang</i>	<i>memegang</i>	‘hold’
30602000	<i>pegang</i>	<i>memegangi</i>	‘hold (prog)’
30701000	<i>kutuk</i>	<i>mengutuk</i>	‘curse’
30702000	<i>kutuk</i>	<i>mengutuki</i>	‘curse (prog)’
30801000	<i>lintas</i>	<i>melintas</i>	‘cross’
30802000	<i>lintas</i>	<i>melintasi</i>	‘cross (prog)’
30901000	<i>tunggang</i>	<i>menunggang</i>	‘mount’
30902000	<i>tunggang</i>	<i>menunggangi</i>	‘mount (prog)’
31001000	<i>wawancara</i>	<i>mewawancara</i>	‘interview’
31002000	<i>wawancara</i>	<i>mewawancarai</i>	‘interview (iter)’
31101000	<i>pandang</i>	<i>memandang</i>	‘behold’
31102000	<i>pandang</i>	<i>memandangi</i>	‘behold (prog)’
31201000	<i>landa</i>	<i>melanda</i>	‘fall on’
31202000	<i>landa</i>	<i>melandai</i>	‘fall gradually’
31301000	<i>lampau</i>	<i>melampau</i>	‘exceed’
31302000	<i>lampau</i>	<i>melampaui</i>	‘exceed’
31401000	<i>sisir</i>	<i>menyisir</i>	‘comb’
31402000	<i>sisir</i>	<i>menyisiri</i>	‘comb (iter)’
31501000	<i>jangkit</i>	<i>menjangkit</i>	‘contract’
31502000	<i>jangkit</i>	<i>menjangkiti</i>	‘infect’
31601000	<i>rasuk</i>	<i>merasuk</i>	‘infiltrate’
31602000	<i>rasuk</i>	<i>merasuki</i>	‘enter into the body’
31701000	<i>kandung</i>	<i>mengandung</i>	‘contain’
31702000	<i>kandung</i>	<i>mengandungi</i>	‘contain’
31801000	<i>dalang</i>	<i>mendalang</i>	‘play with puppet’
31802000	<i>dalang</i>	<i>mendalangi</i>	‘orchestrate (a plot)’
31901000	<i>sahut</i>	<i>menyahut</i>	‘answer a call’
31902000	<i>sahut</i>	<i>menyahuti</i>	‘answer a call’
32001000	<i>tanggap</i>	<i>menanggapi</i>	‘perceive (with senses)’
32002000	<i>tanggap</i>	<i>menanggapi</i>	‘respond (to a comment)’

Table B.4: List of Annotated Group 4 Verbs (NKi)

Starting SID	Root	Surface form	Gloss
40101000	<i>kental</i>	<i>mengental</i>	‘thicken’
40102000	<i>kental</i>	<i>mengentalkan</i>	‘cause to thicken’
40201000	<i>injak</i>	<i>menginjak</i>	‘step on’
40202000	<i>injak</i>	<i>menginjakkan</i>	‘cause something to step on’
40301000	<i>organisasi</i>	<i>mengorganisasi</i>	‘organize’
40302000	<i>organisasi</i>	<i>mengorganisasikan</i>	‘cause to be organized’
40401000	<i>singkap</i>	<i>menyingkap</i>	‘uncover’
40402000	<i>singkap</i>	<i>menyingkapkan</i>	‘cause to be uncovered’
40501000	<i>hitam</i>	<i>menghitam</i>	‘blacken’
40502000	<i>hitam</i>	<i>menghitamkan</i>	‘cause to blacken’
40601000	<i>klasifikasi</i>	<i>mengklasifikasi</i>	‘classify’
40602000	<i>klasifikasi</i>	<i>mengklasifikasikan</i>	‘classify’
40701000	<i>tusuk</i>	<i>menusuk</i>	‘stab’
40702000	<i>tusuk</i>	<i>menusukkan</i>	‘stab something’
40801000	<i>bawa</i>	<i>membawa</i>	‘bring’
40802000	<i>bawa</i>	<i>membawakan</i>	‘bring’
40901000	<i>realisasi</i>	<i>merealisasi</i>	‘realize (achieve)’
40902000	<i>realisasi</i>	<i>merealisasikan</i>	‘cause to realize’
41001000	<i>singkir</i>	<i>menyingkir</i>	‘push aside’
41002000	<i>singkir</i>	<i>menyingkirkan</i>	‘cause something to be pushed aside’
41101000	<i>transformasi</i>	<i>mentransformasi</i>	‘transform’
41102000	<i>transformasi</i>	<i>mentransformasikan</i>	‘transform’
41201000	<i>hafal</i>	<i>mengafal</i>	‘memorize’
41202000	<i>hafal</i>	<i>menghafalkan</i>	‘memorize’
41301000	<i>suntik</i>	<i>menyuntik</i>	‘inject’
41302000	<i>suntik</i>	<i>menyuntikkan</i>	‘cause to inject’
41401000	<i>lukis</i>	<i>melukis</i>	‘paint’
41402000	<i>lukis</i>	<i>melukiskan</i>	‘cause to paint’
41501000	<i>uap</i>	<i>menguap</i>	‘evaporate or disappear’
41502000	<i>uap</i>	<i>menguapkan</i>	‘cause something to disappear’
41601000	<i>setor</i>	<i>menyetor</i>	‘pay store or surrender’
41602000	<i>setor</i>	<i>menyetorkan</i>	‘pay store or surrender’
41701000	<i>adu</i>	<i>mengadu</i>	‘petition’
41702000	<i>adu</i>	<i>mengadukan</i>	‘cause something to be petitioned’
41801000	<i>pecah</i>	<i>memecah</i>	‘break’
41802000	<i>pecah</i>	<i>memecahkan</i>	‘cause something to break’
41901000	<i>goyang</i>	<i>menggoyang</i>	‘sway’
41902000	<i>goyang</i>	<i>menggoyangkan</i>	‘cause something to sway’
42001000	<i>seleweng</i>	<i>menyeleweng</i>	‘swerve to the side’
42002000	<i>seleweng</i>	<i>menyelewengkan</i>	‘cause something to swerve’

Table B.5: List of Annotated Group 5 Verbs (nKi)

Starting SID	Root	Surface form	Gloss
50101000	<i>telor</i>	<i>menelorkan</i>	‘produce’
50201000	<i>pulang</i>	<i>memulangkan</i>	‘bring back’
50301000	<i>tenteram</i>	<i>menenteramkan</i>	‘calm down’
50401000	<i>bumihangus</i>	<i>membumihanguskan</i>	‘decimate’
50501000	<i>tewas</i>	<i>menewaskan</i>	‘destroy’
50601000	<i>janji</i>	<i>menjanjikan</i>	‘promise’
50701000	<i>identik</i>	<i>mengidentikkan</i>	‘equate’
50801000	<i>miris</i>	<i>memiriskan</i>	‘devastating’
50901000	<i>raya</i>	<i>merayakan</i>	‘celebrate’
51001000	<i>kampanye</i>	<i>mengkampanyekan</i>	‘campaign’
51101000	<i>gencar</i>	<i>menggencarkan</i>	‘intensify’
51201000	<i>ramping</i>	<i>merampingkan</i>	‘streamline’
51301000	<i>sekolah</i>	<i>menyekolahkan</i>	‘send to school’
51401000	<i>resep</i>	<i>meresepkan</i>	‘prescribe’
51501000	<i>berita</i>	<i>memberitakan</i>	‘preach or tell’
51601000	<i>lestari</i>	<i>melestarikan</i>	‘conserve’
51701000	<i>asuransi</i>	<i>mengasuransikan</i>	‘insure ensure assure’
51801000	<i>bingung</i>	<i>membingungkan</i>	‘confuse’
51901000	<i>kokoh</i>	<i>mengokohkan</i>	‘strengthen’
52001000	<i>suai</i>	<i>menyesuaikan</i>	‘adjust or adapt’

Table B.6: List of Annotated Group 6 Verbs (nkI)

Starting SID	Root	Surface form	Gloss
60101000	<i>gaul</i>	<i>menggauli</i>	‘biblical lay’
60201000	<i>akal</i>	<i>mengakali</i>	‘outsmart’
60301000	<i>khianat</i>	<i>mengkhianati</i>	‘betray’
60401000	<i>dana</i>	<i>mendanai</i>	‘fund’
60501000	<i>baru</i>	<i>membarui</i>	‘renew’
60601000	<i>awak</i>	<i>mengawaki</i>	‘man or manage’
60701000	<i>kelabu</i>	<i>mengelabui</i>	‘deceive’
60801000	<i>sudah</i>	<i>menyudahi</i>	‘finish or settle’
60901000	<i>karung</i>	<i>mengarungi</i>	‘wade through’
61001000	<i>selusur</i>	<i>menyelusuri</i>	‘go through carefully’
61101000	<i>cedera</i>	<i>mencederai</i>	‘injure’
61201000	<i>ziarah</i>	<i>menziarahi</i>	‘visit’
61301000	<i>selimut</i>	<i>menyelimuti</i>	‘envelop’
61401000	<i>gandrung</i>	<i>menggandrungi</i>	‘love’
61501000	<i>teman</i>	<i>menemani</i>	‘accompany’
61601000	<i>pelopor</i>	<i>mempelopori</i>	‘pioneer’
61701000	<i>jembatan</i>	<i>menjembatani</i>	‘bridge’
61801000	<i>jahil</i>	<i>menjahili</i>	‘bully’
61901000	<i>cabul</i>	<i>mencabuli</i>	‘violate’
62001000	<i>tekun</i>	<i>menekuni</i>	‘do something diligently’

Table B.7: List of Annotated Group 7 Verbs (Nki)

Starting SID	Root	Surface form	Gloss
70101000	<i>ubah</i>	<i>merubah</i>	‘transform’
70201000	<i>induk</i>	<i>menginduk</i>	‘find its source from’
70301000	<i>hardik</i>	<i>menghardik</i>	‘rebuke’
70401000	<i>sogok</i>	<i>menyogok</i>	‘bribe’
70501000	<i>korok</i>	<i>mengorok</i>	‘burrow’
70601000	<i>rekap</i>	<i>merekap</i>	‘recapitulate’
70701000	<i>kuras</i>	<i>menguras</i>	‘drain’
70801000	<i>songsong</i>	<i>menyongsong</i>	‘greet or carry out’
70901000	<i>buang</i>	<i>membuang</i>	‘throw’
71001000	<i>tebus</i>	<i>menebus</i>	‘redeem or atone’
71101000	<i>sinyalir</i>	<i>mensinyalir</i>	‘signal’
71201000	<i>kecam</i>	<i>mengecam</i>	‘investigate or criticize’
71301000	<i>santet</i>	<i>menyantet</i>	‘perform magic’
71501000	<i>bentak</i>	<i>membentak</i>	‘snap at’
71601000	<i>lesu</i>	<i>melesu</i>	‘slow down’
71701000	<i>puja</i>	<i>memuja</i>	‘praise’
71801000	<i>penggal</i>	<i>memenggal</i>	‘decapitate’
71901000	<i>reguk</i>	<i>mereguk</i>	‘enjoy’
72001000	<i>latih</i>	<i>melatih</i>	‘train’
72101000	<i>ratifikasi</i>	<i>meratifikasi</i>	‘ratify’

Appendix C

Three Argument Valency Lists by Group

ARG1	ARG2	ARG3	Count	List of SIDs
EXPERIENCER	STIMULUS	MANNER	3	10101017, 10103002, 10103004
EXPERIENCER	STIMULUS	INSTRUMENT	2	10103000, 10103003
AGENT	EXPERIENCER	STIMULUS	7	10102000, 10102001, 10102004, 10102005, 10102008, 10102011, 10102015
AGENT	STIMULUS	EXPERIENCER	4	10102017, 10102010, 10102003, 10102013
AGENT	PATIENT	MANNER	1	11701020
AGENT	PATIENT	INSTRUMENT	2	11203001, 10101020
AGENT	PATIENT	THEME	20	10501006, 11002002, 10503000, 10503003, 10503004, 10503006, 10503007, 10503008, 10503010, 10503012, 10503013, 10503014, 10503015, 10503016, 10503017, 10503019, 10401004, 10503020, 10401007, 10401011
AGENT	PATIENT	GOAL	5	10602007, 11702003, 11702004, 11702007, 11702008
AGENT	PATIENT	SOURCE	1	11602003

AGENT	PATIENT	PURPOSE	1	12003012
AGENT	PATIENT	CONTENT	19	10501008, 10403000, 10403001, 10403002, 10403003, 10403004, 10403005, 10403006, 10403007, 10403008, 10403009, 10403012, 10403013, 10403014, 10403015, 10403017, 10403020, 10403021, 10503002
AGENT	PATIENT	COPATIENT	31	11702016, 11701000, 11701001, 11701002, 11701004, 11701005, 11701008, 11701009, 11701010, 11701011, 11701012, 11701014, 11701015, 11701016, 11701018, 12002001, 12002002, 12002005, 12002006, 12002007, 12002008, 12002009, 12002012, 12002013, 12002015, 12002017, 11702001, 11702002, 11702009, 11702011, 11702013
AGENT	THEME	MANNER	9	11301000, 10603017, 11901007, 12101012, 11002007, 10603000, 10401017, 10202010, 11302011
AGENT	THEME	INSTRUMENT	2	10202016, 11902007

AGENT	THEME	LOCATION	24	12102016, 12102018, 12102019, 12102020, 10202004, 10602016, 10601001, 10302011, 11902011, 11901000, 11202000, 11901010, 12103001, 11202010, 11202011, 12102001, 12102002, 12102004, 12102008, 12102009, 12102010, 12102012, 12102014, 12102015
-------	-------	----------	----	---

AGENT	THEME	GOAL	66	10902016, 10603009, 10902018, 10902019, 10902017, 11302019, 11302020, 10502018, 11301001, 11301002, 11301003, 10603020, 10901007, 11301007, 11301009, 10602002, 10602003, 10602004, 10602005, 12202005, 11301012, 11301016, 10602009, 11301019, 12202014, 10602014, 10602015, 10202018, 10902008, 10302001, 10302004, 10302005, 10302006, 10902011, 10302010, 10302012, 10302013, 10302014, 10302017, 10902013, 10302019, 11102020, 11202001, 11202002, 11202003, 11202005, 11202013, 11202017, 11202018, 11202020, 10903018, 11302000, 10902000, 10902001, 11302003, 11302004, 10902002, 10902003, 11302007, 10902005, 11302009, 11302010, 10902006, 11302013, 10902014, 10902015
-------	-------	------	----	--

AGENT	THEME	SOURCE	20	10702000, 11102001, 11102002, 11102004, 11102006, 11102007, 11102008, 11102010, 11102011, 10702011, 11102012, 11102015, 10702016, 10702017, 10702018, 11102016, 11102018, 11102019, 10702015, 10603006
AGENT	THEME	PURPOSE	8	10202017, 11303015, 11501001, 10901009, 12102003, 12102006, 11301015, 11302014
AGENT	THEME	BENEFICIARY	6	11903018, 11301004, 10502007, 11302008, 11202009, 10502010
AGENT	THEME	DEGREE	1	12102005
AGENT	THEME	COTHEME	1	12103019
AGENT	THEME	ATTRIBUTE	4	11902009, 11902019, 11301006, 12102007
AGENT	THEME	DURATION	1	12102011
AGENT	LOCATION	MANNER	1	12101020
AGENT	GOAL	MANNER	1	10903001
AGENT	GOAL	THEME	12	10903009, 10901001, 10901002, 10901003, 10903017, 10901005, 10901006, 10901011, 10901018, 10901020, 10903005, 10903006
AGENT	CONTENT	MANNER	1	10402017
AGENT	CONTENT	BENEFICIARY	1	10402014
AGENT	BENEFICIARY	THEME	10	10502019, 10502020, 10402000, 10502001, 10502000, 10502003, 10502004, 10502005, 10502006, 10502014

AGENT	BENEFICIARY	CONTENT	5	10402016, 10402019, 10402004, 10402005, 10402015
AGENT	RECIPIENT	THEME	25	11301005, 11301013, 11301014, 11303000, 11303001, 11303002, 11303003, 11303004, 11303005, 11303006, 11303007, 11303008, 11303009, 11303010, 11303011, 11303013, 11303014, 11303016, 11303017, 11303018, 11303019, 11303020, 11303021, 11302001, 11302005
INSTRUMENT	THEME	GOAL	1	10202012
INSTRUMENT	THEME	SOURCE	1	10202009
INSTRUMENT	GOAL	LOCATION	1	10903000
THEME	MANNER	LOCATION	1	10201020
THEME	THEME	MANNER	1	11202012
THEME	LOCATION	MANNER	4	12101001, 10203002, 12101004, 12101007
THEME	LOCATION	GOAL	1	10203011
THEME	GOAL	MANNER	1	10201015
THEME	DEGREE	SOURCE	1	10701012
THEME	EXTENT	SOURCE	1	12201018
CAUSE	EXPERIENCER	STIMULUS	1	10102018
CAUSE	STIMULUS	EXPERIENCER	1	10102019
CAUSE	PATIENT	COPATIENT	1	12002019
CAUSE	THEME	GOAL	15	12202016, 12202017, 12202018, 12202019, 12202020, 12202000, 12202001, 12202002, 12202003, 12202004, 12202006, 12202007, 12202008, 12202012, 12202015
FORCE	PATIENT	THEME	1	10401006

FORCE	THEME	MANNER	1	11301011
FORCE	THEME	GOAL	1	11301018
FORCE	THEME	SOURCE	12	10702001, 10702002, 10202002, 10702004, 10702005, 10702006, 10702008, 10702009, 10702010, 10702012, 10702013, 10702014
FORCE	BENEFICIARY	CONTENT	1	10402003
CONTENT	AGENT	PATIENT	1	10403010

Table C.1: Verb Subcategorization Frames for Group NKI verbs

ARG1	ARG2	ARG3	Count	List of SIDs
EXPERIENCER	STIMULUS	ATTRIBUTE	1	20102013
AGENT	PATIENT	LOCATION	2	20701003, 20701020
AGENT	PATIENT	PURPOSE	2	20801017, 20801020
AGENT	PATIENT	ATTRIBUTE	2	20101008, 20101006
AGENT	THEME	MANNER	8	21201001, 21201002, 21202002, 21301012, 21002004, 21301014, 21301016, 21301018
AGENT	THEME	INSTRUMENT	3	21401017, 21202019, 21202005
AGENT	THEME	LOCATION	4	21301000, 21401003, 21401005, 20901013
AGENT	THEME	GOAL	3	21401000, 20101017, 21401016
AGENT	THEME	SOURCE	3	21401001, 21401019, 21401004

AGENT	THEME	RESULT	45	21002000, 21002001, 21002002, 21002003, 21002006, 21002008, 21002010, 21002011, 21002012, 21002013, 21002014, 21002015, 21002016, 21002017, 21002018, 21002019, 21002020, 21001000, 21001001, 21001002, 21001003, 21001004, 21001005, 21001006, 21001007, 21001008, 21001009, 21001010, 21001011, 21001012, 21001013, 21001014, 21001015, 21001016, 21001017, 21001018, 21001019, 21001020, 21201000, 21201009, 21201011, 21201012, 21201017, 21201018, 21201020
AGENT	THEME	RECIPIENT	11	20901001, 20901002, 20901019, 20901007, 20901008, 20801009, 21301008, 20901012, 20902011, 20601020, 20902015
AGENT	THEME	ATTRIBUTE	6	21502017, 21102018, 21502019, 20101003, 20101004, 20101016
AGENT	GOAL	CONTENT	2	20101013, 20101015
AGENT	CONTENT	RECIPIENT	3	21701001, 21701012, 21701015
AGENT	RECIPIENT	MANNER	1	20802004
AGENT	RECIPIENT	INSTRUMENT	1	21902008

AGENT	RECIPIENT	THEME	26	20902016, 20902017, 20902018, 20902020, 20901000, 20901004, 20901010, 20901015, 20901018, 20901020, 20602017, 20602018, 20602019, 20602020, 20601019, 20902001, 20902002, 20902003, 20902004, 20902005, 20902007, 20902008, 20902009, 20902010, 20902012, 20902014
AGENT	RECIPIENT	CONTENT	16	21702016, 21902020, 21901001, 21901002, 21901003, 21901005, 21901007, 21901011, 21702004, 21702005, 21901013, 21902007, 21702008, 21702009, 21901019, 21901020
CAUSE	THEME	RESULT	1	21201015
FORCE	THEME	GOAL	1	21401020

Table C.3: Subcategorization Frames for Group NkI verbs

ARG1	ARG2	ARG3	Count	List of SIDs
AGENT	PATIENT	THEME	1	30101020
AGENT	PATIENT	PURPOSE	1	31602016
AGENT	PATIENT	RESULT	1	30701020
AGENT	THEME	MANNER	15	30102017, 30202017, 32002017, 30601000, 30202000, 30602002, 32002002, 32002007, 30202005, 32002005, 30102007, 30202007, 32001018, 30202014, 30202015
AGENT	THEME	AGENT	1	32001009

AGENT	THEME	INSTRUMENT	4	30102009, 32002010, 30202010, 30101007
AGENT	THEME	LOCATION	2	30302009, 30402020
AGENT	THEME	PURPOSE	2	30301003, 30902004
AGENT	THEME	CONTENT	5	31902017, 31901001, 31901002, 31901004, 31902010
AGENT	THEME	EXTENT	2	30102000, 30102001
AGENT	THEME	ATTRIBUTE	21	30202003, 30202004, 30202009, 30202016, 30202020, 32002004, 32002016, 32001000, 32001001, 32001003, 32001004, 32001006, 32001007, 32001008, 32001011, 32001012, 32001013, 32001015, 32001016, 32001017, 32001020
AGENT	THEME	DURATION	1	30601003
AGENT	LOCATION	INSTRUMENT	1	31402018
AGENT	LOCATION	PURPOSE	5	31401006, 31401016, 31402008, 31401019, 31402015
AGENT	ATTRIBUTE	THEME	1	32001019
THEME	GOAL	INSTRUMENT	1	30501005
THEME	GOAL	PURPOSE	1	30501001
THEME	COTHEME	DURATION	1	31701001
FORCE	PATIENT	INSTRUMENT	1	31501000
FORCE	THEME	RESULT	1	30202013
FORCE	THEME	ATTRIBUTE	1	30202019

Table C.3: Subcategorization Frames for Group nKI verbs

ARG1	ARG2	ARG3	Count	List of SIDs
AGENT	PATIENT	MANNER	2	41102020, 41101005

AGENT	PATIENT	INSTRUMENT	8	40502021, 41401002, 40701003, 40701004, 40701007, 40201010, 41101014, 40701019
AGENT	PATIENT	LOCATION	2	40702008, 40701010
AGENT	PATIENT	GOAL	1	41502007
AGENT	PATIENT	SOURCE	4	41101002, 41101020, 41102012, 41102007
AGENT	PATIENT	BENEFICIARY	1	41102009
AGENT	PATIENT	RESULT	12	41101002, 41101015, 41801005, 41101007, 41401009, 41102002, 41102003, 41801009, 41801011, 41801012, 41101010, 41101012
AGENT	THEME	MANNER	3	41902017, 41901009, 40602013
AGENT	THEME	PATIENT	3	40702000, 41302012, 40702005
AGENT	THEME	INSTRUMENT	2	41602001, 41402003
AGENT	THEME	LOCATION	15	40202016, 40202017, 40202020, 40202001, 40202002, 40202003, 40202004, 40202005, 40202006, 40802002, 40202008, 40202010, 40202011, 40202014, 40202015

AGENT	THEME	GOAL	69	41302017, 41302018, 41302019, 41302020, 41301003, 40602000, 40602001, 40602003, 40602004, 41601015, 40602008, 40602010, 40602011, 40602012, 40602018, 41302010, 40602020, 40601000, 40601004, 40702001, 40702002, 40601010, 40702004, 40601013, 40702006, 40702007, 40702010, 40702012, 40601021, 40702014, 40702017, 40702018, 40702019, 40702020, 41602000, 41602002, 41602003, 41602004, 41602008, 41602010, 41602011, 41602014, 41602016, 41602017, 41602018, 41602020, 40801000, 40801001, 40801002, 40801003, 40801004, 41601001, 40801006, 41601002, 40801008, 41601003, 41601008, 41601009, 41302002, 41302003, 41601014, 40801015, 41302006, 41601017, 40801018, 40801019, 40801020, 41302013, 41302014
AGENT	THEME	SOURCE	5	41002018, 41002001, 41002006, 41002010, 41002015
AGENT	THEME	CONTENT	1	40801007

AGENT	THEME	RESULT	1	41901015
AGENT	THEME	COTHEME	1	40602016
AGENT	THEME	RECIPIENT	2	40402002, 40402010
AGENT	THEME	ATTRIBUTE	3	41402016, 41402018, 41402007
AGENT	CONTENT	GOAL	1	41701003
AGENT	CONTENT	CONTENT	1	41702000
AGENT	CONTENT	RECIPIENT	16	41702016, 41702017, 41702018, 41702019, 41702020, 41702001, 41702002, 41702004, 41702005, 41702006, 41702008, 41702010, 41701019, 41702012, 41702014, 41702015
AGENT	CONTENT	COAGENT	1	41701002
AGENT	RECIPIENT	THEME	12	41301001, 41301002, 40802000, 40802001, 40802005, 41301014, 41301015, 40802008, 40802009, 40802012, 40802013, 41301017
AGENT	RECIPIENT	LOCATION	1	41301008
AGENT	RECIPIENT	SOURCE	1	41301012
AGENT	RECIPIENT	CONTENT	2	41701016, 41701017
AGENT	RECIPIENT	RECIPIENT	1	40802004
INSTRUMENT	PATIENT	RESULT	1	41502020
INSTRUMENT	THEME	GOAL	4	40801016, 40801010, 40801012, 40602015
INSTRUMENT	THEME	RECIPIENT	1	41301005
CAUSE	PATIENT	GOAL	1	41502009
CAUSE	PATIENT	SOURCE	1	41101017
CAUSE	PATIENT	RESULT	2	41101001, 41102005
CAUSE	THEME	MANNER	1	40602005
CAUSE	THEME	GOAL	3	41102017, 40801017, 40801014
CAUSE	THEME	SOURCE	1	41002007
LOCATION	THEME	LOCATION	1	40202012
FORCE	PATIENT	RESULT	1	41101000

FORCE	THEME	SOURCE	1	41002020
-------	-------	--------	---	----------

Table C.4: Subcategorization Frames for Group NKi verbs

ARG1	ARG2	ARG3	Count	List of SIDs
AGENT	PATIENT	INSTRUMENT	2	51901001, 51901003
AGENT	PATIENT	GOAL	8	52001002, 52001004, 52001008, 52001012, 52001013, 52001015, 52001017, 52001019
AGENT	THEME	INSTRUMENT	1	50901012
AGENT	THEME	LOCATION	3	51001011, 51101013, 51101014
AGENT	THEME	GOAL	10	50201001, 51101003, 51101004, 50201008, 50201009, 51101016, 51001017, 50201018, 50201019, 51101020
AGENT	THEME	PURPOSE	2	51401011, 51401015
AGENT	THEME	BENEFICIARY	5	51401012, 51401013, 51401014, 51401017, 51401018
AGENT	THEME	RESULT	2	51001020, 51001007
AGENT	THEME	COTHEME	14	50701000, 50701003, 50701004, 50701006, 50701007, 50701009, 50701010, 50701011, 50701012, 50701013, 50701014, 50701016, 50701017, 50701018
AGENT	THEME	RECIPIENT	1	51601012
AGENT	THEME	ATTRIBUTE	2	50901002, 51001005
AGENT	THEME	COAGENT	1	50901009
AGENT	CONTENT	RECIPIENT	2	50601009, 50601020
AGENT	CONTENT	ATTRIBUTE	1	50601014
AGENT	BENEFICIARY	THEME	1	51701012
AGENT	RECIPIENT	LOCATION	3	51301008, 51301017, 51301006

AGENT	RECIPIENT	GOAL	2	51301000, 51301002
AGENT	RECIPIENT	CONTENT	6	50601000, 50601001, 50601006, 50601008, 50601010, 50601018
AGENT	RECIPIENT	RESULT	2	51301018, 51301019
INSTRUMENT	THEME	GOAL	1	50201017
CAUSE	THEME	COTHEME	3	50701001, 50701019, 50701020

Table C.5: Subcategorization Frames for Group nKi verbs

ARG1	ARG2	ARG3	Count	List of SIDs
AGENT	PATIENT	MANNER	5	60101000, 61901000, 60101013, 60701015, 60701017
AGENT	PATIENT	INSTRUMENT	1	60501004
AGENT	PATIENT	LOCATION	1	61901020
AGENT	PATIENT	CONTENT	1	60701010
AGENT	THEME	MANNER	1	62001013
AGENT	THEME	INSTRUMENT	1	62001006
AGENT	THEME	LOCATION	4	61501014, 61501018, 61501019, 61501006
AGENT	THEME	GOAL	3	61501001, 61501003, 61501007
AGENT	THEME	CONTENT	4	61501000, 61501002, 61501011, 61501010
AGENT	THEME	COTHEME	4	61701016, 61701017, 61701018, 61701003
AGENT	THEME	COAGENT	1	60901000
AGENT	THEME	DURATION	1	62001015
AGENT	LOCATION	THEME	1	61301011
INSTRUMENT	THEME	COTHEME	2	61701002, 61701019

Table C.6: Subcategorization Frames for Group nkI verbs

ARG1	ARG2	ARG3	Count	List of SIDs
AGENT	EXPERIENCER	PURPOSE	3	71301010, 71301004, 71301015

AGENT	EXPERIENCER	RESULT	2	71301008, 71301013
AGENT	EXPERIENCER	COAGENT	1	71301018
AGENT	MANNER	THEME	1	71201015
AGENT	PATIENT	MANNER	2	70301004, 70301007
AGENT	PATIENT	INSTRUMENT	4	71801007, 70401002, 71801009, 70401007
AGENT	PATIENT	PURPOSE	5	70401016, 70401018, 70401020, 70401022, 70401023
AGENT	PATIENT	EXTENT	4	71801013, 71801011, 71901013, 71801005
AGENT	PATIENT	DURATION	1	71901003
AGENT	INSTRUMENT	GOAL	1	70501023
AGENT	THEME	MANNER	3	70701003, 70701005, 70701015
AGENT	THEME	INSTRUMENT	10	70501029, 71001000, 72001004, 71701005, 71701008, 71001011, 70501012, 71001016, 71001020, 70501021
AGENT	THEME	CAUSE	2	71201008, 71201020
AGENT	THEME	LOCATION	4	70601000, 72001002, 71001014, 71001015
AGENT	THEME	SOURCE	5	70701002, 71001004, 72001006, 71001007, 70701017
AGENT	THEME	PURPOSE	1	71001013
AGENT	THEME	CONTENT	4	72001000, 72001003, 72001013, 71501007
AGENT	THEME	RESULT	5	71001009, 72001011, 72001012, 72001016, 72001020
AGENT	THEME	ATTRIBUTE	1	71201002
AGENT	RESULT	EXPERIENCER	1	71301019

Table C.7: Subcategorization Frames for Group Nki verbs