A Grammar Library for Information Structure

Sanghoun Song

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Emily M. Bender, Chair

Toshiyuki Ogihara

Fei Xia

Program Authorized to Offer Degree:
UW Linguistics

University of Washington

Abstract

A Grammar Library for Information Structure

Sanghoun Song

Chair of the Supervisory Committee:
Associate Professor Emily M. Bender
Department of Linguistics

This dissertation makes substantial contributions to both the theoretical and computational treatment of information structure, with an eye toward creating natural language processing applications such as multilingual machine translation systems. The aim of the present dissertation is to create a grammar library of information structure for the LinGO Grammar Matrix system (Bender et al. 2010b). Information structure consists of focus, topic, contrast, and background, and refers to how speakers package semantic content they wish to convey to listeners. The information structure of individual sentences is crucial to understanding the cohesiveness of larger segments of text. Despite the crucial role information structure plays in conveying meaning, there is insufficient research on how computational language models might successfully incorporate information structure marking particularly from a multilingual perspective. Part I introduces the current study, and gives some background information. Part II provides cross-linguistic findings about information structure meanings and markings. Part III exploits a naturally occurring text in four languages (e.g. English, Spanish, Russian, and Korean) to formulate a cross-linguistic generalization about distributional properties of information structure. Drawing from these cross-linguistic findings, Part IV shows how information structure can be represented within the HPSG/MRS framework (Pollard and Sag, 1994; Copestake et al., 2005). Part V explores the construction of a grammar library for creating customized grammars incorporating information structure and shows how the information structure-based model improves performance of transfer-based machine translation.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

1/2/3: first/second/third

ABS: absolutive

ACC: accusative

AG: agentive

AUX: auxiliary

CF: contrastive focus

CLF: classifier

CLITIC/CL: clitic

COMP: complementizer

COP: copula

DAT: dative

DECL: declarative

DEF: definite

DET: determiner

DE: *de* in Chinese

DIR: direction

DOBJ: direct object

ERG: ergative

FOC/FC: focus

FUT: future

GEN: genitive

HON: honorific

IMPF/IMP: imperfective

INF: infinite

INT: interrogative

IOBJ: indirect object

LE: *le* in Chinese

LOC: locative

LV: light verb

NEG: negative

NOM: nominative

NONTOP: non-topic

NULL : null marking (zero morpheme)

NUN: (*n*)*un* in Korean

OBJ: object

PART: particle

PAST/PST: past

PERF/PRT: perfective

PL: plural

POLITE: polite

PRES/PRS: present

PROG: progressive

PRON/PRO: pronoun

REFL: reflexive

REL: relative

SG: singular

SHI: *shì* in Chinese

TOP: topic

WA: *wa* in Japanese

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my great adviser, Prof. Emily M. Bender. She introduced me to the study of information structure and provided me a huge help in creating the information structure library. Her advice was really valuable, and her continuous guidance encouraged me to finish my PhD degree.

I am deeply grateful to the other committee members, Prof. Toshiyuki Ogihara and Prof. Fei Xia, for reading my dissertation and providing insightful suggestions. I would also like to thank my GSRs, Prof. Luis Ceze and Prof. Luke Zettlemoyer.

I have received wonderful support from the faculty members at the UW Linguistics department (Prof. Julia Herschensohn, Prof. Sharon Hargus, Prof. Gina-Anne Levow, and many others) both as a student and as a TA/RA. I have also received great assistance from Mike Furr and Joyce Parvi.

I had many helpful discussions with the DELPH-IN developers. I would like to express special thanks to Francis Bond. He was my great supervisor when I served my internship in Japan, and gave me a lot of advices whenever I need. Ann Copestake and Dan Flickinger suggested using ICONS for representing information structure. While discussing with Dan Flickinger, I improved my analysis of how to build up ICONS. Stephan Oepen helped me develop the newer version of ICONS. Antske Fokkens gave me helpful comments on my approach to information structure in the context of grammar modelling. I also had productive discussions with Tim Baldwin, Berthold Crysmann, Lars Hellan, and Petya Osenova.

I have also received important aid from many linguists. Prof. Jong-Bok Kim gave me an opportunity to participate in the KRG project. This led me to the study of HPSG/MRS-based language processing. Prof. Stefan Müller provided me a big help with my corpus study of information structure. Yo Sato let me know his previous study of information structure marking in Japanese and Korean. Bojan Belić helped me understand information structure

properties in Bosnian Croatian Serbian. Prof. Young Chul Jun and Prof. Hae-Kyung Wee read my prospectus and commented on the basic idea of this dissertation. I would like to say a special word of thanks to Prof. Jae-Woong Choe.

I am full of appreciation to fellow students. Michael Wayne Goodman who was my housemate gave me a helping hand in many ways. Joshua Crowgey helped me understand more about the architecture of the LinGO Grammar Matrix system. Woodley Packard made his processor ACE work with ICONS, and I employed it for development and evaluation in this dissertation. Glenn Slayden also made his processor *agree* compatible with ICONS, which I will use in future work. Ned Letcher inspected the initial customization system of information structure and located some problems in the initial library. Varya Gracheva, Marina Oganyan, and Zina Pozen gave me a help with annotating the corpus data in Russian. Lisa Tittle and Maria Burgess annotated the corpus data in English and Spanish. Especially, Lisa Tittle and Varya Gracheva offered me a tremendous help while working on this dissertation. I refined my analysis of information structure sharing ideas with them. Ka Yee Lun helped me a lot regarding Chinese and Cantonese. Naoko Komoto and Sanae Sato provided Japanese judgments in this dissertation. Prescott Klassen and T.J. Trimble gave me a hand for my research, too. Lastly but heartfully, I want to thank Laurie Poulson.

This material is based upon work supported by the National Science Foundation under Grant No. 0644097. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Part I

INTRODUCTION

Chapter 1 presents the motivations and goals of the current work and offers an explanation of the ways in which data and examples are used. Chapter 2 provides an overview of the theoretical underpinnings of the current work: the basic skeleton of within the HPSG (Head-driven Phrase Structure Grammar, Pollard and Sag 1994) and MRS (Minimal Recursion Semantics, Copestake et al. 2005), the fundamentals of grammar engineering, an overview of the LinGO Grammar Matrix customization system (Bender et al. 2010b), and the fundamentals of transfer-based machine translation (Oepen et al. 2007).

Chapter 1

INTRODUCTION

Human languages consist of various structures, among which syntactic structure and semantic structure are particularly well-known. This dissertation is primarily concerned with information structure, and the ways in which it could be leveraged in natural language processing applications.

Information structure is realized by prosodic, lexical, or syntactic cues which constrain interpretation to meet communicative demands within a specific context (Engdahl and Vallduví, 1996). Information structure is comprised of four primary components: focus, topic, contrast, and background. Focus marks what is new and/or important in a sentence, while topic marks what the speaker wants to talk about. Focus and topic are mutually exclusive. Contrast, realized as either contrastive focus or contrastive topic, signals a contextually contrastive set of focus or topic items respectively. That which is not marked as either focus or topic is designated as background information.

Information structure affects the felicity of using a sentence in different discourse contexts, as exemplified in (1).

(1) a. Kim reads the book.

b. It is Kim that reads the book.

c. It is the book that Kim reads.

Though the sentences in (1b-c) are constructed using the same lexical items and describe the same state of affairs as sentence (1a), they differ with respect to the information they focus. 'Kim' is focused in (1b), while 'the book' is focused in (1c). This difference in information structure means that (1b) would be a felicitous answer to *Who is reading the book?* and (1c) would be a felicitous answer to *What is that book?*, but not vice versa.

Furthermore, information structure can be key to finding felicitous translations (Paggio, 1996; Kuhn, 1996). Since languages vary in the ways they mark information structure, a model of information structure meanings and markings is a key component of a well-constructed grammar. For

example, the simple English sentence (2a) can be translated into at least two Japanese allosentences[1] such as (2b) (i.e. (i) with the nominative marker *ga* or (ii) with the so-called topic (and/or contrast) marker *wa*).

(2) a. I am Kim.

b. watashi ga/wa Kim desu.
I NOM/WA Kim COP [jpn]

The choice between the alternatives is largely conditioned by context. Marking on the NP hinges on whether *watashi* 'I' functions as the topic or not. If the sentence is an answer to a question like *Who are you?*, *wa* is preferred. If the sentence is instead a reply to a question like *Who is Kim?*, answering using *wa* sounds unnatural.

This difference in felicity-conditions across languages should be taken into consideration in computational linguistics; especially in machine translation. When machine translation systems cannot accurately model information structure, resulting translations may sound rather awkward or unnatural to native speakers, negatively impacting the performance of the system. It is my firm opinion that successful translation requires reshaping how information is conveyed in accordance with the precepts of the target language and not simply changing words and reordering phrases in accordance with syntactic rules. That level of accuracy in translation requires the incorporation of information structure.

1.1 Motivations

The nature of information structure is less understood than that of syntactic and semantic structures. For many languages the full range of information structure markings remains unknown. Furthermore, the integration of information structure has been rather understudied in computational linguistic investigations to date, despite its potential improving machine translation, Text-to-Speech, automatic text summarization, and many other language processing systems.

There are several opportunities for further exploration of information structure. First, the absence of cross-linguistic findings results in less use of the information in practical systems. In

[1] Allosentences refer to close paraphrases which share truth-conditions (Lambrecht, 1996)

order for language processing systems to provide more fine-grained results and work truly language-independently, lots of knowledge and skills across languages are required (Bender, 2011). Second, distributional findings obtained from language data are still insufficient. Several previous studies exploit language data for the study of information structure, yet the studies use merely monolingual or bilingual texts (Komagata, 1999; Johansson, 2001; Bouma et al., 2010; Hasegawa and Koenig, 2011), so a larger picture of how information structure works across a range of languages is still elusive. Third, existing proposals for representing information structure within a grammatical framework for natural language processing remain somewhat underdeveloped and insufficiently tested. Previous literature, including King (1997), Steedman (2000), and Bildhauer (2007), provides a variety of formalisms to represent information structure, but none of them has been shown to be cross-linguistically valid. Moreover, their formalisms, to my knowledge, have not even been implemented into a practical system in a fully comprehensive way. Lastly, largely for the reasons presented thus far, the potential improvement to machine translation and language processing systems derivable from using information structure has not yet been shown. If the contribution of information structure to improvement of practical applications were to be quantitatively substantiated through experimentation, this motivates further development of information structure-based applications.

1.2 Goals

The central goal of this dissertation is to create a computational model to handle information structure within the HPSG (Head-driven Phrase Structure Grammar, Pollard and Sag 1994) framework, using the MRS (Minimal Recursion Semantics, Copestake et al. 2005) formalism, contributing to the LinGO Grammar Matrix system (Bender et al. 2010b), and from a multilingual perspective. My dissertation is divided into four parts, dedicated to exploring solutions to each of the problems mentioned in the previous subsection individually.

The first part explores various information structure meanings and markings and how they are related to each other within and across different languages. This is done through a review of previous studies on information structure as well as through a survey of various types of languages and their information structure systems. Building on this initial work and additional evidence, a more cross-

linguistically valid explanation of information structure is provided.

The second part utilizes naturally occurring texts in four languages to illustrate their different behaviours with respect to information structure. The languages studied in this section are English, Spanish, Russian, and Korean. As alluded to before, different languages use different phonological, morphological, and syntactic means of expressing information structure. One way to delve into these cross-linguistic differences is to study multilingual texts in which a sentence in one language aligns with a translation in each of the other languages. By investigating multilingual texts, we can determine how information structure markings in different language are related to each other, and find out systematic methods to identify topics and foci in monolingual texts.

The third part presents a formal architecture for representing information structure within the HPSG/MRS-based formalism. The HPSG/MRS-based framework has been used to create extensive computational grammars (e.g. ERG for English (Flickinger 2000), Jacy for Japanese (Siegel and Bender 2002), KRG for Korean (Kim et al. 2011), GG for German (Crysmann 2003, 2005b,a), SRG for Spanish (Marimon 2012), LXGram for Portuguese (Branco and Costa 2010), Norsource for Norwegian (Hellan 2005), and so forth). Moreover, there are several on-going grammar development projects in collaboration with the DELPH-IN (DEep Linguistic Processing with HPSG - INitiative, `http://www.delph-in.net`) consortium. The DELPH-IN group offers an open-source collection of HPSG/MRS-based grammars and software packages, which my dissertation employs.

The fourth part is devoted to the implementation of an information structure-based computational model and evaluates the model. This dissertation is concerned with the LinGO Grammar Matrix system, especially aiming to create a library for information structure and add that library into the customization system (Bender and Flickinger 2005; Drellishak 2009; Bender et al. 2010b). I discuss how the library for information structure is built up and how information structure-based system works for multilingual machine translation.

1.3 Model

The language processing model of the current work is MRS-based, which means the current work aims to incorporate information structure into MRS representation. The main reason for this is to verify that information structure can be used for improving multilingual machine translation system

using the DELPH-IN infrastructure (Oepen et al., 2007). Since the basic model within MRS-based language processing is semantics-based, all ingredients for machine translation, including information structure, should be accessible in semantic representation. The analyses I present here is text-based, but it is also compatible with signal-based NLP tasks such as speech processing. Information structure is often expressed through prosodic marking. Without the help of an automatic acoustic system to resolve prosodic patterns of human sentences, such prosodic behaviours could not be fully represented in text-based processing. To allow for the future incorporation of such phonological information, this model itself should be designed to be extendable to future work in speech processing.[2]

1.4 Data

It is my belief that deep processing of natural languages based on cross-linguistic studies can improve the performance of our language application systems. To this end, this dissertation includes a survey of meanings and markings of information structure across languages which provides a strong background to the theoretical framework, helps focus and constrain search terms when looking for information structure constructions corpus, and gives empirical motivation for extending the customization system to include a model of information structure.

The initial data for these cross-linguistic observations is taken from previous literature on information structure. I augment these findings with two more data sources. The first of them is a repository of student-created grammars developed on the basis of the LinGO Grammar Matrix system, created in fulfillment of a grammar engineering course within the Department of Linguistics at the University of Washington, Linguistics 567 (`http://courses.washington.edu/ling567`, Bender 2007). This collection of grammatical descriptions, entitled Language CoLLAGE (Collection of Language Lore Amassed through Grammar Engineering), are readily available under the MIT license (`http://www.delph-in.net/matrix/language-collage`).[3] This repository of

[2]Although the current work is mainly concerned with text processing, the current model could work (i) through acoustic analysis of speech, (ii) through pre-tagging of information structure, and/or (iii) through mark-up like **boldface** or ALL CAPS.

[3]As of Dec. 14, 2013, Language CoLLAGE consists of 5 languages. These include Malayalam [mal] built in 2007 and Frisian [frr], Lakota [lkt], Miyako [mvi], and Yiddish [ydd] built in 2013. There are many other languages to be curated later.

grammars covers a large variety of language types and a linguistic survey of them could offer valuable insights into the information structure behaviours of less commonly studied languages. Four languages included in the current version of Language CoLLAGE are also used for testing the grammar library for information structure in the present work (§14.5). The second additional source is a multilingual parallel text. Looking into naturally occurring texts allows us to discover mapping types between contextual information and information structure markings. As with other corpus studies, running texts provide counterexamples to theoretic arguments and identify exceptional cases not yet discovered by previous studies. This dissertation exploits a multilingual parallel text in four languages (English, Spanish, Korean, and Russian); *The Adventure of the Speckled Band* out of the *Shamrock Holmes* stories written by Sir Arthur Conan Doyle

1.5 Examples

For ease of exposition, several typeface conventions are employed to represent properties of information structure in examples. First, if a word (or phrase) bears the accent responsible for conveying focus, it is marked in SMALL CAPS. Second, **boldface** denotes an accent conveying topic. Third, $[_f]$ stands for focus projection. For example, in the English Q/A pair in (3), DOG and **Kim** bear the A and B accents (Jackendoff, 1972), respectively, and the focus that DOG (with the A-accent) conveys is projected to the VP *chased the* DOG.

(3) Q: What about Kim? What did Kim do?
A: **Kim** $[_f$ chased the DOG].

Fourth, # means a sentence sounds infelicitous in the given context, though the sentence itself is syntactically legitimate. Finally, ~~strike~~ means either (i) a constituent is semantically and/or informatively empty or (ii) the given utterance cannot be generated from the semantic representation (i.e. MRS).

The examples that previous studies offer, as far as possible, are cited without any change. Thus, glossing conventions may not be consistent across examples. For example, the past morpheme may be glossed as PST in one article or PAST in another. Proper names are not modified at all.[4] Where I have needed to modify an example from the source, the example has been judged by a native

[4]All human languages have well-known personal proper names, which are often used in linguistic literature. For example, *Juan* in Spanish, *Ivan* in Russian, *Zhāngsān* in Chinese, *Taro* in Japanese, *Mia* in Korean, and so on.

speaker of the language. Any sentences provided by native speaker consultants have also been faithfully reproduced. All the examples created for the present study use the gender neutral names *Kim*, *Lee* and *Sandy* for any people, and the name *Fido* for any dog. Every example presented in this dissertation has been (i) taken from literature as is or (ii) checked out by at least one native speaker for each language. In the cases of Korean examples (a language of which I am a native speaker), examples were again, either (i) taken from previous literature or (ii) created by me and judged by another Korean native speaker.

Lexical markers in Korean and Japanese have been dealt with differently in previous literature despite the morphosyntactic similarity between the two languages. As for the lexical markers, I would like to follow the different approach that Jacy (Siegel and Bender 2002) and KRG (Kim et al. 2011) are based on, because this dissertation aims to contribute to DELPH-IN grammars. KRG identifies the lexical markers in Korean (e.g. *i/ka* for nominatives, (*l*)*ul* for accusatives, and (*n*)*un* for topics) as affixes responsible for syntactic (and sometimes semantic) functions of the phrases that they are attached to. In contrast, the lexical markers (e.g. *ga*, *o*, and *wa*) have been treated as adpositions by Jacy, which behave as a syntactic head.[5] Postpositions in Japanese, such as *ga* and *wa*, in different literature are sometimes are attached to NPs with a hyphen (e.g. *inu-ga* 'dog-NOM'), or sometimes are separated by white space (e.g. *inu ga*). In extracted Japanese examples the presence/absence of the hyphen means the presence/absence in the original source. In any Japanese examples created for this project, I make use of white space instead of a hyphen, following the convention of Jacy. Note that, the different glossing format notwithstanding, Japanese lexical markers are all implemented as adpositions (i.e. separate lexical items) in the current work. As for the Korean examples, following the convention in previous literature, hyphens are made use of (e.g. *kay-ka* 'dog-NOM') without any white space before lexical markers. Accordingly, the lexical markers in Korean are dealt with and implemented as affixes, unlike those in Japanese.

After Chapter 9, the information structure of a sentence is mainly represented by means of dependency graphs, indicating the binary relations of each information structure component as a directed, labelled arc between word-forms.

ISO 639-3 codes, such as [spa] for Spanish, [rubs] for Russian, [jpn] for Japanese, [kor] for

[5] On why the Japanese markers are needed to be treated as adpositions rather than affixes, see Siegel (1999) and Yatabe (1999).

Korean, etc., are attached to all examples not in English. Catalogues of languages and examples are provided in Appendix B and C, respectively.

1.6 Terminology

In addition to differences in glossing conventions, there is also some variation in the terminology used by previous research into information structure. First, the distinction between focus *vs.* topic has sometimes been regarded as a relationship between rheme and theme, a distinction originally conceptualized by the Prague School. Within this framework, theme is defined as the element with the weakest communicative dynamism in a sentence, while rheme is defined as the element with the strongest communicative dynamism (Firbas, 1992, p. 72). There have been various studies which incorporate these components of information structure. Halliday (1967) regards theme as one of the major components for articulating information structure. Using slightly different terminologies, Vallduví (1990) considers focus to be the prime factor of information structure. A sentence, in the Vallduví's schema, can be divided into focus and ground, and ground can be divided again into link and tail. Link is roughly equivalent to topic in this dissertation, with tail corresponding to the remaining portion of the sentence. For example, in (4), *the* DOG functions as the focus of the sentence and ***Kim*** *chased* is the ground of the sentence which comprises the link ***Kim*** and the tail *chased*.

(4) Q: What about Kim? What did Kim chase?
A: [[**Kim**]$_{\text{LINK}}$ chased]$_{\text{GROUND}}$ the DOG.

There are also some variation in labels for denoting contrast. Vallduví and Vilkuna (1998) use the term 'kontrast', in order to emphasize that it has a different semantic behaviour from non-contrastive focus. Instead of using such theory-specific terms (e.g. rheme, theme, link, tail, and kontrast), this dissertation, makes use of most widespread and common terms for referring to components of information structure: focus, topic, contrast, and background.

On the other hand, to avoid potential confusion, the present work provides alternate terminology for several morphosyntactic phenomena. First, there are the OSV constructions in English as exemplified in (5b), which are sometimes cited as examples of 'topicalization' in the sense that *Mary* in (5a) is topicalized and preposed.

(5) a. John saw Mary yesterday.

b. Mary, John saw yesterday. (Prince, 1984, p. 213)

Instead, this dissertation calls such a construction 'focus/topic-fronting' taking the stance that the constructions like (5b) are ambiguous. Because a fronted phrase such as *Mary* in (5b) can be associated with either focus or topic, the term 'topicalization' cannot satisfactorily represent the linguistic properties of such a construction. Second, in a similar vein, *wa* in Japanese and (*n*)*un* in Korean have been labelled as 'topic markers' by many previous studies. However, they are not used exclusively to mark topics. They are sometimes employed in establishing contrastive focus. Thus, 'topic-marker' is not an appropriate name.[6] Instead, this dissertation uses just *wa*-marking and (*n*)*un*-marking in order to avoid confusion. In the IGT (Interlinear Glossed Text) format of Japanese and Korean examples, even if the source of the IGT says TOP, they are just annotated as WA and NUN unless there is a particular reason for saying TOP.

1.7 Outline

This dissertation is structured as follows. Chapter 2 provides the essential background including the basic skeleton of HPSG and MRS, the fundamentals of grammar engineering, and an overview of the LinGO Grammar Matrix customization system. Part II provides a cross-linguistic survey of information structure: Chapter 3 lays out the meanings each component of information structure conveys, and Chapter 4 looks into three forms of expressing information structure: namely prosody, lexical markings, and sentence position. Chapter 5 discusses the discrepancies in meaning-form mapping of information structure. Part III employs a multilingual parallel text: Chapter 6 gives an explanation of the process of annotation and evaluation. Chapter 7 presents several cross-linguistic findings with respect to strategies for marking information structure. Part IV investigates how information structure can be represented in HPSG/MRS: First, several previous studies on information structure are surveyed in Chapter 8. After that, I propose the definition of a new constraint type and feature hierarchy for modelling information structure in HPSG/MRS. ICONS (mnemonic for Individual CONStraints) are presented as an extension to MRS in Chapter 9 and 10. Chapter 9 presents the fundamentals of representing information structure via ICONS, and Chapter 10 goes into the

[6] Chapter 5 looks into the discrepancies between forms and meanings in information structure in detail.

particulars of how ICONS works with some sample derivations. Next, Chapter 11 shows how information structure in multiclausal utterances can be represented via ICONS, and Chapter 12 delves into several means of expressing information structure with reference to ICONS. Finally, Chapter 13 explores how focus projection can be supported by underspecification. Part V addresses the computational implementation of the grammar library for information structure: Chapter 14 builds up a grammar library for information structure. Chapter 15 addresses how machine translation can be improved using ICONS. In Chapter 16, I conclude this dissertation and present the outlook for future research promised by the results of this dissertation.

Chapter 2

BACKGROUND

This chapter provides necessary background for contextualizing the current research. HPSG (Head-driven Phrase Structure Grammar, Pollard and Sag 1994) framework is summarized in §2.1. The MRS (Minimal Recursion Semantics, Copestake et al. 2005) formalism for meaning representation is surveyed in §2.2. §2.3 gives a bird's-eye view of grammar engineering with special reference to the DELPH-IN (Deep Linguistic Processing with HPSG - Initiative, `http://www.delph-in.net`) consortium. The reference formalism that the DELPH-IN grammars and processors are using is described in Type Definition Language (hereafter, TDL). TDL is briefly explained in §2.4. The current work especially aims to make a contribution to the LinGO Grammar Matrix system (Bender et al. 2010b). §2.5 introduces how the LinGO Grammar Matrix system works. §2.6 addresses how transfer-based machine translation operates.

2.1 Head-driven Phrase Structure Grammar

In a nutshell, HPSG has the following characteristics.

(1) a. sign-based

b. surface-oriented and monostratal

c. unification-based

d. lexicalized (head-driven)

e. using typed feature structures

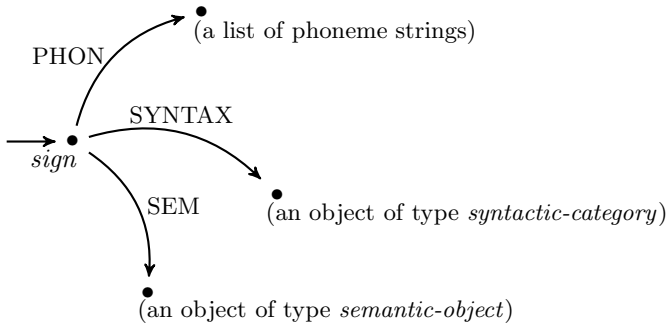
f. avoiding redundancies

First of all, HPSG is essentially conceptualized as *sign*[1] as sketched out in (2) (Sag et al., 2003, p. 16).[2]

[1]HPSG, of course, includes much richer information into the formalism than the traditional 'sign' (De Saussure, 1931).

[2]The schema presented in (2) is the former version presented in Pollard and Sag (1987). In the later version, SYNTAX

(2)



Since *sign* is the most general element representing linguistic structure in HPSG, all linguistic units (e.g. words, phrases, and sentences) within the HPSG framework are regarded as instances of *sign*.

Second, one of the most important motivations in HPSG is surface-orientedness, which means that the grammar looks at the surface form as it appears, not using additional empty elements. HPSG also does not assume movement (i.e. monostratality), which makes the difference from transformation-based grammars such as the GB (Government and Binding) theory (Chomsky, 1993) and the Minimalist Program (Chomsky, 1995). Different linguistic layers (e.g. phonology, syntax, semantics, etc.) are synthetically represented in a single structure (i.e. *sign*).

Third, as with LFG (Lexical Functional Grammar, Bresnan 2001), HPSG employs unification as the combinatory mechanism. Linguistic information is gathered incrementally and unified in such a way that coreference to the same value is modelled as token-identity of that value.

Fourth, HPSG is lexicon-oriented like LFG, in that its framework moves much more information away from the grammatical rules into the lexicon. HPSG deals with syntactic or semantic categories by using lexical information. The lexical information plays a crucial role. Well-formed phrase structure rules should obey two principles; (i) the Head-Feature Principle and (ii) the Subcategorization Principle. The Valence Principle in the later version of HPSG is an extension of the Subcategorization Principle. These principles are defined as follows.

(3) a. Head Feature Principle: The HEAD value of any headed phrase is structure-shared with the HEAD value of the head daughter. (Pollard and Sag, 1994, p. 34)

b. Subcategorization Principle: In a headed phrase (i.e. a phrasal sign whose DTRS value is of sort *head-struc*), the SUBCAT value of the head daughter is the concatenation of the phrase's SUB-

---

and SEM are merged into SYNSEM (Pollard and Sag, 1994). Thus, in the recent HPSG formalism, *syntactic-category* is handled under SYNSEM|CAT(egory), while *semantic-object* is treated under SYNSEM|CONT(ent).

CAT list with the list (in order of increasing obliqueness) of SYNSEM values of the complement daughters. (Ibid. p. 34)

c. The Valence Principle: Unless the rule says otherwise, the mother's values for the VAL feature (SPR and COMPS) are identical to those of the head daughter. (Sag et al., 2003, p. 106)

The HPSG-based formalism also uses typed feature structures (henceforth, TFS), which provide a method for expressing various types of grammatical constraints in an abstract way in order to capture linguistic generalizations. The current work largely depends on the DELPH-IN version of TFS in the context of HPSG-based grammar engineering (Copestake, 2002). The main properties of the DELPH-IN TFS are summarized in (4).

(4) a. Connectedness and Unique Root: A typed feature structure has a unique root node; apart from the root, all nodes have one or more parent nodes.

b. Unique Features: Any node may have zero or more arcs leading out of it, but the label on each (that is, the feature) must be unique.

c. No Cycles: No node may have an arc that points back to the root node or to a node that intervenes between it and the root node.

d. Types: Each node must have a single type, which must be present in a type hierarchy.

e. Finiteness: A typed feature structure must have a finite number of nodes. (Copestake, 2002, p. 45–46)

TFSs essentially consist of two components. One is the supertypes of the current type, and the other is a set of attribute-value pairs. The key concept in TFS is inheritance. Each constraint is hierarchically built up and a type completely inherits all constraints specified in its supertypes. One node in the hierarchy can inherit information from multiple upper nodes (i.e. multiple inheritance). Properties of type hierarchies are as follows.

(5) a. Unique Top: There is a single hierarchy containing all the types with a unique top type.

b. No Cycles: There are no cycles in the hierarchy.

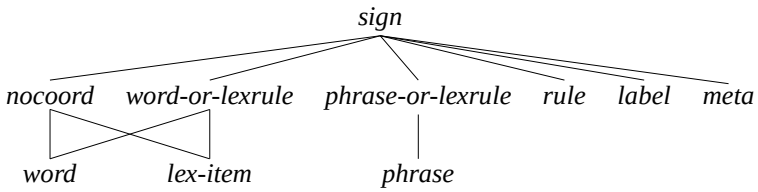
c. Unique Greatest Lower Bounds: Any two types in the hierarchy must either be incompatible, in which case they will not share any descendants, or they are compatible, in which case they must have a unique highest common descendant. (Copestake, 2002, p. 39)

A feature structure is dealt with in HPSG as a set of attribute-value pairs. The pairs between them are traditionally represented in the form of an Attribute-Value Matrix (henceforth, AVM). For example, the minimal structure of sign and its hierarchy in `matrix.tdl` of the LinGO Grammar Matrix system (explained in §2.5) are defined as follows:

(6) a.

sign	
STEM	*list*
KEY-ARG	*bool*
SYNSEM	*synsem*
ARGS	*list*
INFLECTED	*inflected*

b.



An AVM representation of sign (6a) makes it explicit that every *sign* has a SYNSEM structure. *Sign* is divided into several subtypes in a hierarchy (6b). *Word-or-lexrule* has two subtypes *word* and *lex-item*, and they also inherit form *nocoord*. The other major subtype of *sign* is *phrase-or-lexrule*, which is the supertype of *phrase*.

One of the main motivations within the HPSG framework is reducing redundancy in description. In particular, redundancies are seriously rejected in the formal analysis of the HPSG-based studies. The creation of hierarchical lexical categories in HPSG cuts down on redundancies (Flickinger, 1987): (a) the type-hierarchical system of lexical categories, and (b) hierarchically organizing lexical rules for inflectional and derivational morphology.

2.2 Minimal Recursion Semantics

MRS (Minimal Recursion Semantics (Copestake et al. 2005), or sometimes called "Meaning Representation System") is a framework for computational modeling of semantic representation. This dissertation represents information structure in MRS via ICONS (Individual CONStraints). That is, representation of information structure is incorporated into MRS (Meaning Representation System, in this context). This is an important departure from previous work in which MRS was conceived as a (possibly underspecified) representation of a truth-condition associated with a sentence.

There are two distinct characteristics of MRS representations: First, MRS introduces a flat representation expressing meanings by feature structures. Second, MRS takes advantage of underspecification (for handling quantifier scopes and others), which allows flexibility in representation. In MRS description, it is important to represent the meanings of a sentence in an efficient manner for a practical purpose. The main criteria MRS is grounded upon are as follows.

(7) a. Expressive Adequacy: The framework must allow linguistic meanings to be expressed correctly.

b. Grammatical Compatibility: Semantic representations must be linked cleanly to other kinds of grammatical information (most notably syntax).

c. Computational Tractability: It must be possible to process meanings and to check semantic equivalence efficiently and to express relationships between semantic representations straightforwardly.

d. Underspecifiability: Semantic representations should allow underspecification (leaving semantic distinctions unresolved), in such a way as to allow flexible, monotonic resolution of such partial semantic representations. (Copestake et al., 2005, p. 281–282)

The minimal components of MRS include HOOK, RELS, and HCONS as shown in (8).[3]

(8) a.
$$\begin{bmatrix} \textit{mrs} & \\ \text{HOOK} & \textit{hook} \\ \text{RELS} & \textit{diff-list} \\ \text{HCONS} & \textit{diff-list} \end{bmatrix}$$

b.
$$\begin{bmatrix} \textit{hook} & \\ \text{LTOP} & \textit{handle} \\ \text{INDEX} & \textit{individual} \\ \text{XARG} & \textit{individual} \end{bmatrix}$$

c.
$$\begin{bmatrix} \text{RELS} \left\langle ! \ldots, \begin{bmatrix} \textit{relation} & \\ \text{LBL} & \textit{handle} \\ \text{PRED} & \textit{string} \\ \text{ARG0} & \textit{individual} \end{bmatrix}, \ldots ! \right\rangle \end{bmatrix}$$

d.
$$\begin{bmatrix} \text{HCONS} \left\langle ! \ldots, \begin{bmatrix} \textit{qeq} & \\ \text{HARG} & \textit{handle} \\ \text{LARG} & \textit{handle} \end{bmatrix}, \ldots ! \right\rangle \end{bmatrix}$$

First of all, HOOK keeps track of the attributes that need to be externally visible upon semantic composition, whose minimal components are included in (8b). The value of LOP (Local TOP) is the handle of the relation or relations with the widest fixed scope within the constituent. The value of INDEX is the index that a word or phrase combining with this constituent might need access to. The value of ARG (external ARGument) is identified with the index of a semantic argument which

[3] AVMs in (8), in which a difference list (i.e. *diff-list*) is used as the value of RELS and HCONS, are the grammar-internal representations of MRS as feature structures. When MRSs are used as an interface representation, they use *list* rather than *diff-list*, and do not involve feature structures.

serves as the subject in raising and control constructions. Second, REL is a bag of Reps (Elementary Predicates), whose type is a *relation*. Each *relation* has at least three attributes: LB (Label), PR ED (Predicate), and ARG0 (ARGument #0). The value of LBL is a *handle*, which represents the current EP. The value of PRED is normally a string, such as '_dog_n_1_rel', '_bark_v_rel', etc.[4] The value of ARG0 is either *ref-ind* for EPs introduced by nominals or *event-ind* for EPs introduced by verbals, adjectives, adverbs, and adpositions. Depending on the semantic argument structure of an EP, more ARGs can be introduced. For example, intransitive verbs (e.g. *bark*) additionally have ARG1, transitive verbs (e.g. *chase*) have ARG1 and ARG2, and ditransitive verbs (e.g. *give*) have ARG1, ARG2, and ARG3. Finally, HCONS represents partial information about scope. The value of HCONS is a bag of *qeq* (equality modulo quantifier) constraints.

More recently, alternative representations of MRS have been suggested for ease of utilizing the MRS formalism for a variety of language applications, which include RMRS (Robust MRS, Copestake 2007), and DMRS (Dependency MRS, Copestake 2009). Compared to MRS, RMRS involves the functionality of underspecification of relational information, which facilities shallow techniques in language processing (e.g. NP chunking). DMRS, which makes use of a dependency style representation, is another variant of MRS. DMRS, designed to facilitate machine learning algorithms, mainly aims to remove redundancies that (R)MRS may have (Eveleigh, 2010). This dissertation makes use of the conventional version of MRS, but the dependency style representation DMRS deploys is introduced for ease of explication.

2.3 Grammar Engineering

This subsection briefly looks over what grammar engineering is and how it has been studied.

In a nutshell, grammar engineering is the process of creating machine-readable implementations of formal grammars. In order to look at the whole picture of grammar engineering, it is necessary to understand what language is. Since the early days of generative study of linguistics language has been defined as (i) an infinite set of (ii) strings accepted as (iii) grammatical by (iv) native speakers, and grammar engineering has embraced this definition. (i) Given that the number of sentences in human language is assumed to be nonfinite, grammar engineering takes the generative capacity

[4]The PRED value can be a type, particularly for incorporating lexical semantics (i.e. wordnet) into the meaning representation (Francis Bond, p.c.). Besides, even though the PRED value is treated as a string, it is structured.

of grammar into account in sentence-generation as well as sentence-parsing. (ii) Most work in grammar engineering is based on text-processing in that language is regarded as a set of strings. (iii) Since formulating grammatical well-formedness in a language is crucial, grammar engineering is fundamentally concerned with constructing a linguistically-precise and broad-coverage grammar. (iv) Finally, grammaticality has to be judged by native speakers. The judgment can be made either by linguistic intuition or sometimes with reference to language data such as corpora. Methodology based on intuition and methodology based on data complement each other in grammar engineering.[5]

The main goal of grammar engineering is to build up reusable computational grammar resources. Ideally, the empirical description of the data is linguistically motivated. A grammar is to be described in a linguistically well-elaborated way, and on a large scale to cover the linguistic phenomena in a human language. Second, the described grammar should be able to run on a computer in order to prove its mathematical tractability as well as its potential for utilization. Third, the constructed grammar has to be reusable for other studies with varied research goals.[6]

Grammar engineering utilizes various types of linguistic data such as (machine-readable) dictionaries, corpora (Nichols et al., 2010; Song et al., 2010), test-suites (Oepen, 2001), treebanks (Oepen et al., 2004; Bond et al., 2006), and wordnets (Bond et al., 2009; Pozen, 2013). Grammars can also be constructed grounded on linguistic literature as reference grammars; existing grammars (a) for other languages on one platform or (b) for one language on other platforms. Grammar engineering includes parsing and generation, which can be used for several practical applications such as machine translation, grammar checking, information extraction, question-answering, etc.

Within the field of grammar engineering, there are several competing theories of grammar, including HPSG, LFG, CCG (Combinatory Categorial Grammar, Steedman 2001), and TAG (Tree-Adjoining Grammar, Joshi and Schabes 1997). HPSG, which employs typed feature structures

---

[5]Baldwin et al. (2005), in the context of grammar engineering, discuss this spirit in an overall sense and conduct an evaluation using the ERG (English Resource Grammar, Flickinger 2000) and the BNC (British National Corpus, Burnard 2000). They substantiate the interaction of two sources of linguistic findings: namely grammaticality judgments and corpus data. Their conclusion is that combining the two types of evidence facilitates exploring their interaction in more detail.

[6]Bender (2008) offers an explanation about how grammar engineering can be used for linguistic hypothesis testing: "[L]anguages are made up of many subsystems with complex interactions. Linguists generally focus on just one subsystem at a time, yet the predictions of any particular analysis cannot be calculated independently of the interacting subsystems. With implemented grammars, the computer can track the effects of all aspects of the implementation while the linguist focuses on developing just one." (Ibid. p. 16)

as a mathematical foundation, has been used for creation of reusable computational grammars in many languages. Those who study grammar engineering within the HPSG framework have cooperated with each other, by forming a consortium called DELPH-IN.[7] DELPH-IN, in the spirit of open-source NLP (Pedersen, 2008), provides research and development outcomes in a readily available way. These are largely gathered in the LOGON repository (`http://moin.delph-in.net/LogonTop`).[8] LOGON includes a collection of DELPH-IN grammars (e.g. ERG (Flickinger 2000), Jacy (Siegel and Bender 2002), KRG (Kim et al. 2011), etc.), processors (e.g. LKB (Copestake 2002), PET (Callmeier 2000), etc.), and other related software (e.g. [incr tsdb()] (Oepen 2001)).

The current work is largely dependent upon the collaborated results of the DELPH-IN consortium. First, I make use of the DELPH-IN formalism to construct the HPSG/MRS-based information structure library from a multilingual perspective on grammar engineering. Second, the comprehensive DELPH-IN grammars (i.e. resource grammars, such as the ERG and the Jacy) are often referred to during the construction. Finally, in this dissertation, I utilize the DELPH-IN tools to check the feasibility of what I propose and conduct several types of evaluations.

2.4 Type Description Language

The grammatical fragments the current work creates are described in Type Description Language (TDL). For facilitate of the syntax, this subsection provides a summary of TDL.

TDL aims to describe feature structures within constraint-based grammars (`http://moin.delph-in.net/DelphinTutorial/Formalisms`). TDL has been partially simplified and partially extended in the reference formalism of DELPH-IN. Thus, all processors in the DELPH-IN collection: LKB (Copestake 2002), PET (Callmeier 2000), ACE (`http://sweaglesw.org/linguistics/ace`), and *agree* (Slayden 2012), are fully compatible with TDL. The syntax of TDL in the DELPH-IN formalism has three components: (i) multiple type inheritance, (ii) attribute-value

[7] There are other initiatives based on HPSG as well as other frameworks, such as CoreGram for HPSG-based implementations (`http://hpsg.fu-berlin.de/Projects/CoreGram.html`) using the TRALE system (`http://www.sfs.uni-tuebingen.de/hpsg/archive/projects/trale`), and ParGram in LFG-based formalism (`http://pargram.b.uib.no`). There are also other HPSG-based grammars such as Enju for English (`http://www.nactem.ac.uk/enju`, Miyao and Tsujii 2008), and a Chinese grammar constructed in a similar way to Enju (Yu et al. 2010).

[8] Note that not all DELPH-IN resources are in the LOGON repository. For example, the collection of *Language CoLLAGE* is not in the repository, but is readily available (`http://www.delph-in.net/matrix/language-collage`).

constraints, and (iii) coreference. For example, (9) indicates that the current type inherits from two supertypes and that the value of an attribute SYNSEM|HEAD should be consistent with the value of the head daughter's SYNSEM|HEAD.

```
(9) type-name := supertype-name-1 & supertype-name-2 &
      [ SYNSEM.CAT.HEAD #head,
        HEAD-DTR.SYNSEM.CAT.HEAD #head ].
```

One of the frequently used data structures in TDL is list. For instance, a list <a,b,c> can be represented as follows.

```
(10) [ FIRST a,
       REST [ FIRST b,
              REST [ FIRST c,
                     REST e-list ] ] ]
```

Lists sometimes need to work more flexibly to allow concatenation, append, removal, etc. For these operations, the DELPH-IN formalism utilizes difference lists (*diff-list*). This structure maintains a pointer to the last element of the list. Analogously to (10), a difference list <!a,b,c!> can be represented as in (11).

```
(11) [ LIST [ FIRST a,
              REST [ FIRST b,
                     REST [ FIRST c,
                            REST #last ] ] ],
       LAST #last ]
```

2.5 The LinGO Grammar Matrix System

The LinGO Grammar Matrix (Bender et al. 2010b) is an open source starter kit for the rapid development of HPSG/MRS-based grammars. The grammars created by the LinGO Grammar Matrix system are to be (i) rule-based, (ii) scalable to broad-coverage, and (iii) cross-linguistically comparable. The main idea behind the system is that the common architecture simplifies exchange of analyses among groups of developers, and a common semantic representation speeds up implementation of multilingual processing systems such as machine translation.

The core components of the LinGO Grammar Matrix are implemented as a customization system, a web-based application for creating HPSG fragments on the basis of user input (Bender and Flickinger, 2005; Drellishak, 2009; Bender et al., 2010b). The customized grammars are described in TDL and are executable on DELPH-IN software such as [incr tsdb()]. The system is made up of (i) a core grammar in `matrix.tdl` containing types and constraints that are useful for modelling

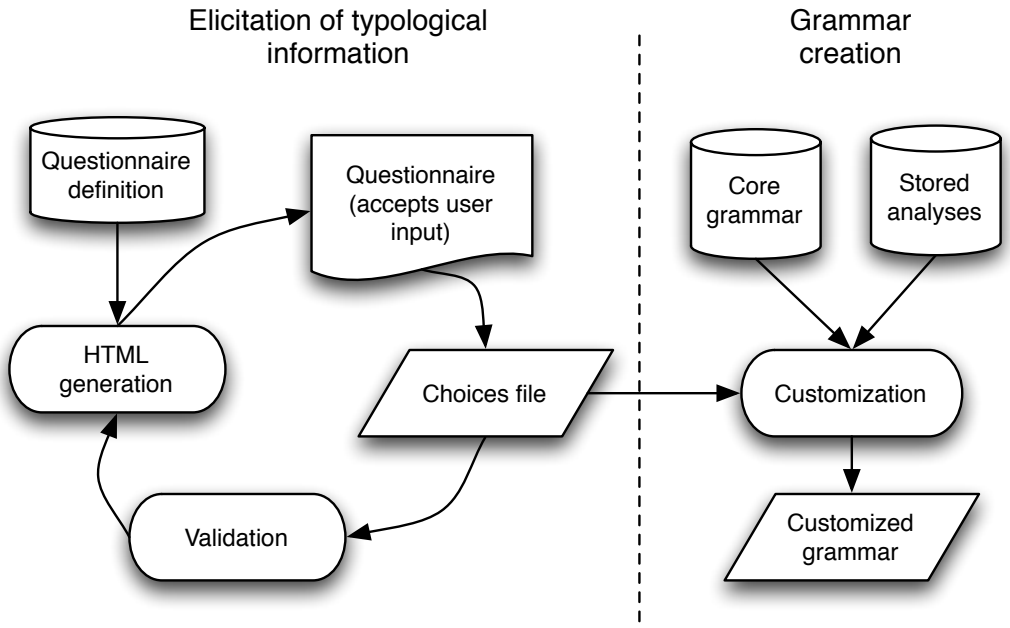


Figure 2.1: The LinGO Grammar Matrix customization system

phenomena in all human languages, and (ii) linguistic libraries for widespread, but non-universal language phenomena (Bender and Flickinger, 2005; Drellishak, 2009).

Figure 2.1, reproduced from Bender et al. (2010b), shows how the LinGO Grammar Matrix customization system works on the basis of user input. The questionnaire that the users give answers to is converted into a file named `choices`, which is validated to check out consistency at every single step. The choices file serves as the specification that the customization system refers to for creating an HPSG/MRS-based grammar. Using the choices made by the users, the process of customizing grammars is accomplished with two types of grammatical components: (i) a core grammar, and (ii) stored analyses (i.e. phenomena-specific libraries).

The core grammar in the LinGO Grammar Matrix is based on (12), to which this dissertation adds several more attributes.

(12)
$$\begin{bmatrix} \textit{sign} & & & & & \\ \text{STEM} & \textit{list} & & & & \\ \text{PHON} & \textit{phon} & & & & \\ \text{SYNSEM} & \begin{bmatrix} \text{LOCAL} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \textit{cat} & & \\ \text{HEAD} & \textit{head} & \\ \text{VAL} & \begin{bmatrix} \text{SUBJ} & \textit{list} \\ \text{COMPS} & \textit{list} \\ \text{SPR} & \textit{list} \\ \text{SPEC} & \textit{list} \end{bmatrix} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \textit{mrs} & & \\ \text{HOOK} & \begin{bmatrix} \text{GTOP} & \textit{handle} \\ \text{LTOP} & \textit{handle} \\ \text{INDEX} & \textit{individual} \\ \text{XARG} & \textit{individual} \end{bmatrix} \\ \text{RELS} & \textit{diff-list} \\ \text{HCONS} & \textit{diff-list} \end{bmatrix} \end{bmatrix} \\ \text{NON-LOCAL} & \begin{bmatrix} \text{SLASH} & \textit{0-1-dlist} \\ \text{QUE} & \textit{0-1-dlist} \\ \text{REL} & \textit{0-1-dlist} \end{bmatrix} \end{bmatrix} \\ \text{ARGS} & \textit{list} \\ \text{INFLECTED} & \textit{inflected} \end{bmatrix}$$

In addition, the stored analyses treat linguistic phenomena common to many languages. In other words, there is a small set of analyses that handle most languages, including word order (Fokkens, 2010), yes/no questions (Bender and Flickinger, 2005), and sentential negation (Crowgey, 2012). Many languages have a linguistic system for representing tense and aspect (Poulson, 2011) as well as marking person, number, and gender (Drellishak, 2009). Quite a few languages employ a morphological paradigm with some linguistic constraints (Goodman, 2013). Other phenomena, which are widely used in natural languages, have also been integrated into the customization system; for example, coordination (Drellishak and Bender, 2005), cognitive status (Bender and Goss-Grubbs, 2008), and argument optionality (Saleem, 2010; Saleem and Bender, 2010).

The starter grammar that the LinGO Grammar Matrix customization system creates on the basis of user input makes two contributions to grammar engineering. First, the starter-grammar is useful to those who have an interest in testing linguistic hypotheses within the context of a small implemented grammar (Bender et al., 2011). Second, starter grammars serve as a departure point to those who want to construct broad-coverage implemented grammars, and sometimes presents directions

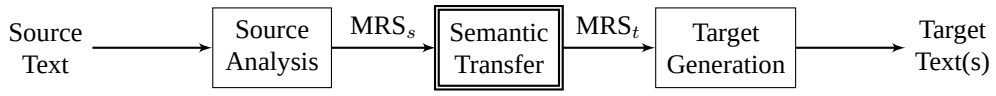


Figure 2.2: HPSG/MRS-based MT architecture

for improvement to an existing grammar. Thus far, the LinGO Grammar Matrix has been used to construct new HPSG/MRS-based grammars (e.g. BURGER (BUlgarian Resource Grammar – Efficient and Realistic, Osenova 2011) for Bulgarian), and to improve existing grammars (e.g. KRG2 for Korean, Song et al. 2010).

2.6 Transfer-based Machine Translation

The basic method I employ for testing machine translation herein is built in the symbolic approach to machine translation, which normally consists of three stages: namely (i) parsing, (ii) transfer, and (iii) generation. Since MRS is not an interlingua (i.e. a meaning representation language in which the representations are identical for all languages), using MRS for machine translation requires an independent stage to convert one MRS into another MRS. This stage is called transfer, and is carried out between parsing and generation.

Figure 2.2, adapted from Oepen et al. (2007) and Song et al. (2010), is illustrative of the MRS-based architecture of machine translation. The first step (i.e. parsing) analyses a sentence with a computational grammar for the source language, whose output is a form of semantic representation such as a (near) logical form. The output of the first step serves as the source of the next step (i.e. transfer), which is called MRS$_s$ (i.e. an input MRS). The transfer module converts the source representation obtained from the parsing process into another type of representation compatible with the target language, which is called MRS$_t$ (i.e. an output MRS). MRS$_t$ is used as the source for the final step (i.e. generation), which constructs one or more surface form(s) conveying the same meaning as the semantic representation. As a consequence, the two surface forms in the source language and the target language are compatible with a common meaning representation.

2.7 Summary

This chapter provided background knowledge for contextualizing this dissertation. The fundamental grammar theory used is HPSG. HPSG is unification-based, lexicon-oriented, and concerned with surface forms. HPSG uses typed feature structure and avoids redundancies in description. The meaning representation system employed for the purpose of computational modelling is MRS. MRS introduces a flat representation expressing meanings by way of feature structures and makes the best use of underspecification in order to facilitate flexibility in representation. The present work follows the notion of grammar engineering for creating and using the information structure-based computational model. Critically, all proposals and implementations in this dissertation are founded on the infrastructure the DELPH-IN consortium provides. The present work models information structure in Type Description Language and thereby creates a cross-linguistically valid library of information structure for the LinGO Grammar Matrix system. This model can be used to improve performance of transfer-based machine translation.

Part II

MEANINGS AND MARKINGS OF INFORMATION STRUCTURE

This part scrutinizes information structure meanings and markings from a typological perspective. Chapter 3 investigates components of information structure, such as focus, topic, contrast, and background, paying attention to definition, subtypes, linguistic properties, and diagnostic tests. Chapter 4 addresses forms of expressing information structure: prosody, lexical markers, syntactic positioning. Chapter 5 presents some cases in which meanings and markings of information structure are discrepant.

Chapter 3

MEANINGS OF INFORMATION STRUCTURE

The main parameters of information structure are focus, topic, contrast, and background (Lambrecht, 1996; Gundel, 1999; Féry and Krifka, 2008). (i) Focus means what is new and/or important in the sentence. (ii) Topic refers to what the sentence is about. (iii) Contrast applies to a set of alternatives, which can be realized as either focus or topic. (iv) Background is neither focus nor topic.

The main criterion for classifying components of information structure in this dissertation is the linguistic forms. If there is a way to express an information structure meaning linguistically, that meaning can be regarded as a cross-cutting category that participates in information structure. This criterion is also applied to the taxonomy of information structure in each language. If a language has a linguistic means of expressing a type of information structure meaning, the corresponding component is assumed to function as an information structure value in the language.

The current analysis of information structure meanings builds on the following findings: (i) Every sentence has at least one focus, because new and/or important information plays an essential part in information processing in that all sentences are presumably communicative acts (Engdahl and Vallduví, 1996; Gundel, 1999). (ii) Sentences do not always have a topic (Büring, 1999), which means there are topicless sentences in human language. (iii) Contrast, contra Lambrecht (1996),[1] is treated as a component of information structure given that it can be linguistically expressed. (iv) Sometimes, there is a linguistic item to which neither focus nor topic are assigned (Büring, 1999), which is called background (a.k.a. 'tail' in the schema of Vallduví and Vilkuna (1998)) hereafter.

Building upon the taxonomy presented above, this dissertation makes three fundamental assumptions. Two of them largely come from Engdahl and Vallduví (1996), and the last one, to my knowledge, is original to this dissertation. First, focus and topic cannot overlap with each other in a

[1] Lambrecht (1996, p. 290–291) says "Given the problems involved in the definition of the notion of contrastive, I prefer not to think of this notion as a category of grammar. To conclude, contrastiveness, unlike focus, is not a category of grammar but the result of the general cognitive processes referred to as conversational implicatures."

single clause.[2] That means there is no constituent that plays both roles at the same time in relation to a specific predicate. The information structure meaning of a constituent within a clause should be either, or neither of them (i.e. background). Second, as constituents that can receive prosodic accents presumably belong to informatively meaningful categories (Lambrecht, 1996), contentful words and phrases either bear their own information structure meanings or assign an information structure meaning to other constituents. Finally, just as informatively meaningful categories exist, there are also lexical items to be evaluated as informatively meaningless. The informatively void items themselves cannot be associated with any component of information structure, though they can function in forming information structure.

Each following section is structured as follows. First, the fundamental notion and the definition of each component are provided. Second, each component is classified into several subtypes, with special reference to the ways it can be marked in linguistic forms. Next, linguistic properties relating to each component are discussed. Finally, linguistic tools to identify the meaning and marking of each component are surveyed.

3.1 Information Status

Before discussing information structure meanings in turn, it is necessary to go over information status such as givenness (i.e. new *vs*. old). It is my position that information structure interacts with but is distinct from information status.

Information status has been widely studied in tandem with information structure (Gundel, 2003). For instance, Halliday (1967) claims that focus is not recoverable from the preceding discourse because what is focused is new. To take another instance, Cinque (1977) argues the leftmost NPs and PPs in dislocation constructions have a restriction on information status in some languages. According to Cinque, in Italian, a required condition for placing a constituent to the left peripheral position of a sentence is that the constituent should deliver old information. Thus, NPs and PPs conveying new information cannot be detached from the rest of a sentence in Italian. In this context, it has been assumed that information status has to do with information structure meanings in that new information bears focus, and topic is something given in the context. However, there are more

[2]Chapter 11 looks into two or more different information structure values that a constituent can have with respect to different clauses (i.e. multiclausal constructions).

than a few counterexamples to this generalization (Erteschik-Shir, 2007): New information can occasionally convey topic meaning, and likewise focus does not always carry new information.

Definiteness, upon which the choice of determiners is dependent, has also been assumed to have an effect on articulation of information structure: Definite NPs carry old information, and indefinite NPs carry new information. Thus, it has been thought that indefinite NPs cannot be the topic of a sentence, unless used for referencing generics (Lambrecht, 1996). In particular, topic-comment structures have a tendency to assign a topic relation to only definite NPs. Kuroda (1972), for instance, claims *wa*-marked NPs in Japanese, widely assumed to deliver topic meaning, can be translated into only either definite or indefinite non-specific NPs in English, while *ga*-marked NPs (i.e. ordinary nominatives) do not have such a correspondence restriction in translation. A similar phenomenon can be found in Chinese. Chinese employs three types of word orders such as SVO (unmarked), SOV, and OSV, but the ordering choice is influenced by the definiteness of the object: The preverbal object in SOV and OSV constructions seldom allows an indefinite non-specific expression.[3]

(1) a. wo zai zhao yi-ben xiaoshuo.
I at seek one-CL novel
'I am looking for a novel.' (SVO) [cmn]

b. *wo yi-ben xiaoshuo zai zhao.
I one-CL novel at seek (SOV) [cmn]

c. *yi-ben xiaoshuo, wo zai zhao.
one-CL novel I at seek (OSV) [cmn] (Huang et al., 2009, p. 200)

However, there are quite a few counterarguments to the generalization that topic is always associated with definiteness. Erteschik-Shir (2007) argues that the correspondence between marking definiteness and topichood is merely a tendency. This argument can be supported in several languages. First, Yoo et al. (2007), exploiting a large English-Korean bilingual corpus, verify there is no clear-cut corresponding pattern between (in)definite NPs in English and the NP-marking system (e.g. *i/ka* for nominatives *vs*. (*n*)*un* for topics or something else) in Korean. Thus, we cannot say that the correlation between expressing definiteness and topichood is cross-linguistically true. Second,

[3]In (1), tones are not displayed, because that is how they were in the source text.

since some languages (e.g. Russian) seldom use definite markers, we cannot equate definiteness with topichood at the surface level. It is clear that definiteness is presumably language universal. Every language has (in)definite phrases in the interpretation, even though they are not necessarily overtly expressed in some languages. What I take notice of is overt marking systems of definiteness in some languages. For instance, distinctions between different types of determiners (e.g. *the*/*a*(*n*) in English) do not have a one-to-one correspondence with information structure components. That is, at least in English, not all NPs specified with *the* deliver a topic meaning, and likewise NPs with *a*(*n*) sometimes have topic meaning in certain circumstances.

I conclude that information status is neither a necessary nor a sufficient condition for identifying information structure; the relationship between them is just a tendency in some languages and quite language-specific. For this reason, in this dissertation, I downplay the discussion of information status, and instead pay more attention to information structure.

3.2 Focus

3.2.1 Definition

Focus, from a pragmatic point of view, refers to what the speaker wants to draw the hearer's attention to (Erteschik-Shir, 2007; Féry and Krifka, 2008). Lambrecht (1996) regards the basic notion of focus as (2).

(2) a. Pragmatic Presupposition: the set of presuppositions lexicogrammatically evoked in an utterance which the speaker assumes the hearer already knows or believes or is ready to take for granted at the time of speech. (Ibid. p. 52)

b. Pragmatic Assertion: the presupposition expressed by a sentence which the hearer is expected to know or believe or take for granted as a result of hearing the sentence uttered. (Ibid. p. 52)

c. Focus: the semantic component of a pragmatically structured proposition whereby the assertion differs from the presupposition. (Ibid. p. 213)

In a nutshell, focus means what speakers want to say importantly and/or newly, which is influenced by both semantics and pragmatics. Building upon (2), the current work represents information structure within the MRS (Minimal Recursion Semantics, Copestake et al. 2005) formalism. In the following subsection, different approaches to the taxonomy of focus are provided on different levels

of classification (syntactic, semantic, and pragmatic). Among them, I mainly adapt Gundel (1999)'s classification, because it is based on linguistic markings and properties of focus. In other words, semantic focus (a.k.a. non-contrastive focus) and contrastive focus are distinguishably marked in quite a few languages, and they exhibit different linguistic behaviours from each other across languages.

3.2.2 Subtypes of Focus

Lambrecht (1996)

Lambrecht classifies focus into three subtypes depending on how focus meaning spreads into larger phrases; (a-i) argument focus, (a-ii) predicate focus, and (a-iii) sentential focus. The main classification criterion Lambrecht proposes is sentential forms, which suggests that how a sentence is informatively articulated largely depends on the scope that the focus has in a sentence. For argument focus, the domain is a single constituent such as a subject, an object, or sometimes an oblique argument. Predicate focus has often been recognized as the second component of 'topic-comment' constructions. That is, when a phrase excluding the fronted and topicalized constituent is in the focus domain, the rest of the sentence is an instance of predicate focus. Sentential focus's domain is the entire sentence, in which topic disappears.

This notion has been developed in quite a few studies. For instance, Paggio (2009) offers a type hierarchy for sentential forms, looking at how components of information structure are articulated and ordered in a sentence.[4] In the taxonomy of Paggio, there are two main branches: namely focality and topicality. As subtypes of focality, Paggio presents narrow focus and wide focus. Note that the argument focus is not the same as the narrow focus. The former means that an argument (i.e. NP) of the predicate is marked as the focus of the clause, while the latter means that a single word is marked as the focus of the clause. Thus, non-nominal categories such as verbs, adjectives, and even adverbs can be narrowly focused. The same goes for the distinction between predicate focus and wide focus. The predicate focus literally means the predicate plays the core role of focus, and the focus is spread onto the larger VPs. What wide focus has in common with predicate focus is that they always involve focus projection, but the core of wide focus can be various types of categories

[4]The type hierarchy that Paggio proposes is presented in Chapter 8 with discussion about which implication it has on the current work.

including nominal ones (e.g. common nouns, proper names, pronouns, etc.). In other words, (i) argument focus is a subset of narrow focus; predicate focus is a subset of wide focus, and (ii) a narrow focus is not necessarily an argument focus; a wide focus does not necessarily involves a predicate focus.

É. Kiss (1998)

É. Kiss, in line with alternative semantics (Rooth 1992), suggests a distinction between (b-i) identificational focus and (b-ii) informational focus.

(3) An identificational focus represents a subset of the set of contextually or situationally given elements for which the predicate phrase can potentially hold. (É. Kiss, 1998, p. 245)

(3) implies identificational focus has a relation to a powerset of the set consisting of all the elements in the given context. Thus, the elements in the alternative set of identificational foci are already introduced in the context, while those of informational foci are not provided in the prior context. The difference between them can be detected in the following sentences in Hungarian; (4a-b) exemplify identificational focus and informational focus, respectively.

(4) a. Mari EGY KALAPOT nézett ki magának.
Mary a hat.ACC picked out herself.ACC
'It was A HAT that Mary picked for herself.' [hun]

b. Mari ki nézett magának EGY KALAPOT.
Mary out picked herself.ACC a hat.ACC
'Mary picked for herself A HAT.' [hun] (Ibid. p. 249)

According to É. Kiss, (4a) sounds felicitous in a situation in which Mary was trying to pick up something at a clothing store, which implies that she chose only one hat among the clothes in the store, and nothing else. (4b), by contrast, does not presuppose such a collection of clothes, and provides just new information that she chose a hat. In other words, there exists an alternative set given within the context in (4a), which establishes the difference between identificational focus and informational focus.

Gundel (1999)

Gundel, mainly from a semantic standpoint, divides focus into (c-i) psychological focus, (c-ii) semantic focus (a.k.a. non-contrastive focus), and (c-iii) contrastive focus. Psychological focus, according to Gundel's explanation, refers to the current center of attention, and has to do with unstressed pronouns, zero anaphora, and weakly stressed constituents. Among the three subtypes that Gundel presents, my dissertation takes only the last two as the subtypes of focus, because psychological focus seems to be a component of information status, rather than information structure.

Gundel offers some differences between semantic focus and contrastive focus. First, semantic focus is the most prosodically and/or syntactically prominent.[5] This is in line with Givón (1991)'s claim that the most important element in a cognitive process naturally has a strong tendency to be realized in the most marked way. This property of focus is also argued by Büring (2010) as presented in (5).

(5) Focus Prominence: Focus needs to be maximally prominent. (Büring, 2010, p. 277)

Second, semantic focus does not necessarily bring an entity into psychological focus, whereas contrastive focus always does. Finally and most importantly, contrastive focus has no influence on the truth-conditions, while the semantic focus is truth-conditionally sensitive.

Gussenhoven (2007)

Gussenhoven classifies focus in English into seven subtypes in terms of its functional usage within the context. These include (d-i) presentational focus, (d-ii) corrective focus, (d-iii) counterpresupposition focus, (d-iv) definitional focus, (d-v) contingency focus, (d-vi) reactivating focus, and (d-vii) identificational focus. (d-i) Presentational focus is a focused item corresponding to *wh*-words in questions. (d-ii) Corrective focus and (d-iii) counterpresupposition focus appear when the speaker wants to correct an item of information that the hearer incorrectly assumes. In this dissertation, these subtypes are regarded as contrastive focus such that the correction test can be used as a tool to vet contrastive focus (Gryllia, 2009). (d-iv) Definitional focus and (d-v) contingency focus, which

[5]Of course, contrastive focus is also prosodically and/or morphosyntactically marked. What is to be noted is that semantic focus is assigned to the most prominent constituent in a sentence.

usually occur with an individual-level predicate, aim to inform the hearer of the attendant circumstances: For example, *Your* EYES *are* BLUE. states that the eye-color of the hearer is generically blue. (d-vi) Reactivating focus, unlike other subtypes of focus, is assigned on given information and is realized by the syntactic device called focus/topic fronting in this dissertation. Finally, (d-vii) identificational focus (É. Kiss, 1998) is realized within clefts (e.g., *It is* JOHN *who she dislikes.*). The taxonomy provided by Gussenhoven has its own significance in that it shows the various functions that focus performs, but this dissertation does not directly use it. Gussenhoven's subtypes seem to be about the way in which focus is used to different communicative ends (i.e. pragmatics), which is not synonymous with focus as defined in the current work. Additionally, these distinctions are not directly and linguistically marked in some languages. Recall that I restrict the subtypes of information structure components to those which are signaled by linguistic marking in at least one language.

Summary

Regarding the subtypes of focus, this dissertation draws primarily from Gundel (1999), except for psychological focus which has a more relevance to information status. Focus is classified into semantic focus (a.k.a. non-contrastive focus) and contrastive focus. There are two reasons. First, in quite a few languages, they are distinctively expressed via different lexical markers or different positions in a clause. Second, they show clearly different behaviours. In particular, semantic focus is relevant to truth-conditions, while contrastive focus is not. Lambrecht (1996) provides a classification in terms of how a sentence is configured. The classification has to do with focus projection and the ways in which a core of focus spread onto a larger phrase. The classification that É. Kiss (1998) proposes is not also applied to the basic taxonomy of focus in this dissertation, but the distinction (i.e. identificational *vs*. informational) is reviewed in the analysis of cleft constructions. Gussenhoven (2007)'s subtypes show various properties of focused elements, but they are not also straightforwardly incorporated into the components of focus herein. This is mainly because they are seldom linguistically distinguishable.

3.2.3 Linguistic Properties of Focus

There are four major properties of focus realization; (i) inomissibility, (ii) felicity/truth-conditions, and (iii) focus sensitivity.

Inomissibility

Information structure is a matter of how information that a speaker bears in mind is articulated in a sentence. Thus, formation of information structure has to do with selecting the most efficient way to convey what a speaker wants to say. Focus is defined as what is new and/or important in an utterance and necessarily refers to the most marked element in an utterance (Gundel, 1999; Büring, 2010). If the maximally prominent information is missing, a conversation becomes a void and infelicitous. As a result, focus can never be dropped from the utterance with reference to context.[6] For this reason, inomissibility has been commonly regarded as the universal factor of focus realization in previous literature (Lambrecht, 1996; Rebuschi and Tuller, 1999): only non-focused constituents can be elided. Lambrecht, for instance, suggests an ellipsis test. In (6), *John* and *he* convey topic meaning, and *he* can be elided as shown in (6A2). In contrast, if the subjects are focused, elision is disallowed as shown in (7A2).

(6) Q: What ever happened to John?
A1: John married ROSA, but he didn't really LOVE her.
A2: John married ROSA, but didn't really LOVE her.

(7) Q: Who married Rosa?
A1: JOHN married her, but he didn't really LOVE her.
A2: *?JOHN married her, but didn't really LOVE her. (Lambrecht, 1996, p. 136)

I argue that (8) is the most important property of focus.

(8) Focus is an information structure component associated with an inomissible constituent in an utterance.

This property can also be straightforwardly applied to contrastive focus. Constituents associated

[6]Note that this distinguishes focus as a component of information structure from the information status *in focus*, since referents that are *in focus* can often be referred to with zero anaphora (Gundel, 2003).

with contrastive focus cannot be elided, either.[7] This is the main distinction between contrastive focus and contrastive topic. As mentioned before, contrast is realized as either contrastive focus or contrastive topic. In other words, a constituent conveying contrastiveness should be either of these. In some cases, because many languages use the same marking system to express both contrastive focus and contrastive topic and they share a large number of properties, it would be hard to discriminate them using existing tests. However, when we test whether a constituent is omissible or not, they are distinguishable. A constituent with contrastive focus cannot be dropped, whereas one with contrastive topic can. This difference between them is exemplified in Chapter 5 (p. 90) with reference to discrepancies between meanings and markings.

On the other hand, the fact that focus can be assigned to only contextually inomissible constituents logically entails another theorum that dropped elements in subject/topic-drop languages can never be evaluated as conveying focus meaning. It is well-known that subjects in subject-drop languages such as Italian and Spanish can be dropped. What should be noted is that there is a constraint on dropping subjects that hinges on information structure. Cinque (1977, p. 406) argues that subject pronouns in Italian are omissible everywhere unless the subjects give new information (i.e. focus from the perspective of this dissertation).

I argue that *pro*-drop is relevant to expressing information structure, mainly focusing on argument-optionality (Saleem, 2010; Saleem and Bender, 2010): Some languages often and optionally drop an NP with non-focus meaning. That is, dropped arguments in *pro*-drop languages must be non-focus. *Pro*-drop can be divided into two subtypes: subject-drop and topic-drop.[8] Topic-drop has been explained in relation to subject-drop. Typical examples are shown in (9) (a set of multilingual translations, taken from *The Little Prince*). (9) are answers in English, Spanish, Korean, and (Mandarin) Chinese to a *wh*-question like *What are you doing there?*.

(9) a. I am drinking.

---

[7] In terms of the HPSG formalism, because contrastive focus is also a specific type of focus, linguistic features plain focus involves are directly inherited into contrastive focus.

[8] It is clear that we cannot equate topics with a single grammatical category like subjects at least in English, Spanish, Korean, and Chinese. Linguistic studies, nonetheless, have provided ample evidence that topics and subjects have a close correlation with each other across languages: Subjects normally are the most unmarked topics in most languages (Lambrecht, 1996; Erteschik-Shir, 2007). Therefore, in more than a few cases, it is not easy to make a clear-cut distinction between subject-drop and topic-drop, which stems from the fact that subjects display a tendency to be interpreted as topics.

b. Ø Bebo.
(I) drink.1SG [spa]

c. Ø swul masi-n-at.
(I) alcohol drink-YPRES-DECL [kor]

d. (wǒ) hē jǐu.
I drink alcohol [cm]

The subjects in (9a-d) are all first person, and also function as the topic of the sentences. The different languages have several different characteristics. (i) The use of 'I' in (9a) is obligatory in English. (ii) The subject in Spanish, a morphologically rich language, can be freely dropped as shown in (9b). (iii) The subject in Korean is also highly omissible as shown in (9c), though Korean does not employ any agreement in the morphological paradigm. (iv) Chinese, like English, is morphologically impoverished, and also is like Korean in that it does not have inflection on the verb according to the subject. The subject in Chinese (e.g. *wǒ* 'I' in (9d)) can be dropped as well. The subjectless sentences exemplified in (9c-d) in Korean and Chinese have been regarded as instances of topic-drop in quite a few previous studies in tandem with the subjectless sentences in subject-drop languages (e.g. Spanish) (Li and Thompson, 1976; Huang, 1984; Yang, 2002; Alonso-Ovalle et al., 2002).

Felicity-conditions and Truth-conditions

Felicity is conditioned by how a speaker organizes an utterance with respect to a particular context. For this reason, information structure generally affects felicity conditions. That is, information structure should be interpreted with respect to the contexts in which an utterance of a particular form can be successfully and cooperatively used.

Information structure has often been studied in terms of allosentences. These are close paraphrases which share truth-conditions (Lambrecht, 1996).[9] Engdahl and Vallduví (1996) begin their analysis with a set of allosentences though they do not use that terminology. Allosentences (10a-b) differ in the way their content is packaged: (10a) in which the object is focused is an appropriate answer to a question like '*What does he hate?*,' while (10b) in which the verb is focused is

[9]Truth-condition is sensitive to what the speaker is saying about the world.

not. Propositions in (10a-b) have in common what they assert about the world (i.e. the same truth-condition), but differ in the way the given information is structured.

(10) a. He hates CHOCOLATE.

b. He HATES chocolate.

c. Chocolate he LOVES. (Engdahl and Vallduví, 1996, p. 2)

In a nutshell, allosentences are sentences which differ only in felicity-conditions. Although a set of allosentences is comprised of exactly the same propositional content, the sentences convey different meanings from each other, and the differences are caused by how focus is differently expressed.

Information structure can also impact truth conditions (Partee, 1991; Gundel, 1999). Partee (1991) provides the following example, originally taken from Halliday (1970). (11a-b)[10] convey different meanings depending on which item the A-accent (H* in the ToBI format, (Bolinger, 1961; Jackendoff, 1972)) falls on. In other words, the focused items in (11a-b) differ, and they causes differences in truth-conditions.

(11) a. Dogs must be CARRIED.

b. DOGS must be carried. (Partee, 1991, p. 169)

(11a) means everyone who brings a dog should carry the dog in his or her arms. (11b) implies it is not allowed to go around without a dog. For instance, if a pedestrian who has no dog sees a sign of *Dogs must be carried!* in front of a building, and interprets it as (11b), it is likely that (s)he will hesitate to enter the space.

Focus Sensitivity

Beaver and Clark (2008) claim that focus sensitive items deliver complex and non-trivial meanings which differ from language to language, and their contribution to meaning is rather difficult to elicit. What is notable with respect to focus sensitive items is that if there is an item whose contribution to the truth-conditional semantics (e.g., where it attaches in the semantic structure) is focus-sensitive,

[10]In the notion of formal semantics, they are analyzed as (a) MUST(dog(x) & here(x), x is carried) and (b) MUST(here(e), a dog or dogs is/are carried at e) or MUST(you carry x here, you carry a dog here), respectively (Partee, 1991, p. 169).

then changes in information structure over what is otherwise the same sentence containing that item should correlate with changes in truth-conditions.

Focus sensitive items related to truth-conditions include modal verbs (e.g. *must* in (11)), frequency adverbs (e.g. *always*), counterfactuals, focus particles (e.g. *only*, *also*, and *even*), and superlatives (e.g. *first*, *most*, etc.) (Partee, 1991). To take an example, (12b-c), in which two focus-sensitive operators *most* and *first* are used, do not share the same truth-conditions.

(12) a. The most students got BETWEEN 80 AND 90 on the first quiz.

b. The most students got between 80 and 90 on the FIRST quiz. (Partee, 1991, p. 172)

Table 3.1 helps understand the difference between (12a-b) in truth-conditions.

Table 3.1: Quiz scores

	1st	**2nd**
Kim	85	85
Lee	87	88
Sandy	95	87

If three students Kim, Lee, and Sandy received the scores presented in Table 3.1 in their first and second quizzes, (12b) is false. Sandy got 95 in the first quiz, while all three students got between 80 and 90 in the second quiz. Since the second one is the situation in which the scores are more clustered between 80 and 90, (12b) is not true.

3.2.4 Tests for Focus

As exemplified several times so far, *wh*-questions have been commonly used to probe meaning and marking of focus (Lambrecht, 1996; Gundel, 1999). The phrase answering the *wh*-word of the question is focused in most cases; the focused part in the reply may be either a word (i.e. narrow focus, or argument focus), a phrase consisting of multiple words (i.e. wide focus, or predicate focus), or a sentence including the focused item (i.e. all focus, or sentence focus). For instance, if a *wh*-question is given like *What barks?*, the corresponding answer to the *wh*-word bears the A-accent, such as *The* DOG *barks.*

It seems clear that using *wh*-questions is a very reliable test for identifying focus in the sense that we can determine which linguistic means for marking are used in a language. Yet, there are also instances in which *wh*-questions cannot be used. In particular, it is sometimes problematic to use *wh*-questions to locate focused constituents in running texts, which do not necessarily consist of Q/A pairs. Moreover, even when we know which marking is responsible for focus in a language, it is sometimes troublesome to pinpoint which element contributes focus meaning unless the marking system is overtly expressed, such as in clefting constructions. For instance, since the primary way to express information structure meanings in English is prosody, *wh*-questions are unlikely to be determinate when analyzing written texts in English. This problem is also raised by Gracheva (2013), who utilizes the Russian National Corpus for a study of contrastive structures in Russian. She points out that it is troublesome to apply existing tests of information structure, such as *wh*-questions, to naturally occurring speech. This is because, when working with running text, it is actually impossible to separate a single sentence from the context and test it independently.

In order to make up for the shortcomings of the *wh*-test, I make use of one more tool: the deletion test. As illustrated in the previous subsection (§3.2.3), one of the important linguistic properties of focus is inomissibility. The deletion test is particularly useful for corpus studies that annotate information structure values in naturally occurring texts (Part III).[11]

3.2.5 Summary

Focus signals what is new and/or important in a sentence, and constituents associated with focus cannot be elided in the given context (i.e. are inomissible). Previous literature proposes several taxonomies of focus subtypes: Lambrecht (1996) suggests (a-i) argument focus (a subset of narrow focus), (a-ii) predicate focus (a subset of wide focus), and (a-iii) sentential focus. É. Kiss (1998) classifies focus into (b-i) identificational focus which appears in cleft constructions in English and other languages and (b-ii) informational focus which occurs in ordinary sentences with a specific marking (e.g. the A-accent). Gundel (1999) proposes (c-i) psychological focus, (c-ii) semantic focus (a.k.a. contrastive focus), and (c-iii) contrastive focus. Because psychological focus is likely to be

[11] Another test for focus is identifying the strongest stress (Rebuschi and Tuller, 1999), but Casielles-Suárez (2004) provides a counterexample to this test. Casielles-Suárez reveals that primary stress does not always guarantee the focus even in English. In particular, finding the more remarkable stress is not available for this dissertation that basically aims at text processing. Thus, only two tests, namely the *wh*-test and the deletion test, are employed in this dissertation.

more related to information status, this dissertation is exclusively concerned with the last two subtypes. Semantic focus is sensitive to truth-conditions, while contrastive focus is not. Gussenhoven (2007) divides focus in English into seven subtypes: (d-i) presentational focus, (d-ii) corrective focus, (d-iii) counterpresupposition focus, (d-iv) definitional focus, (d-v) contingency focus, (d-vi) reactivating focus, and (d-vii) identificational focus. Focus has an effect on felicity-conditions and sometimes on truth-conditions when focus sensitive items appear in a sentence. The most well-known test to identify focus is *wh*-questions, but it is not straightforwardly applicable to naturally occurring texts. For this reason, I employ the deletion test, leveraging the fact that focused items are inomissible.

3.3 Topic

3.3.1 Definition

Topic refers to what a sentence is about (Strawson, 1964; Lambrecht, 1996; Choi, 1999), which can be defined as (13).

(13) An entity, E, is the topic of a sentence, S, iff in using S the speaker intends to increase the addressee's knowledge about, request information about, or otherwise get the addressee to act with respect to E. (Gundel, 1988, p. 210)

There is an opposing point of view to this generalization. Vermeulen (2009) argues that what the sentence is about is not necessarily the topic of the sentence. Vermeulen does not analyzes the subject *he* in (14A) as the topic of in the sentence, though the sentence itself is about the subject. According to the definition presented by Vermeulen, *he* is just an anaphoric item that merely refers back to the so-called discourse topic *Max* in (14Q).

(14) Q: Who did Max see yesterday?
A: He saw Rosa yesterday.

However, this narrow definition is not adopted by this dissertation for two reasons: First, Vermeulen's argument runs counter to the basic assumption presented by Lambrecht (1996), who asserts that a topic has to designate a discourse referent internal to the given context. Second, if the answer (14A), given to a question like *Who did Max see yesterday?*, is translated into Korean in

which the (*n*)*un* marker is used in complementary distribution with the nominative marker *i*/*ka*, *ku* 'he' can be combined with only the (*n*)*un* marker as shown in (14′).

(14′) Q: Maysu-nun/ka ecey mwues-ul po-ass-ni.
Max-NUN/NOM yesterday what-ACC see-PST-INT
'What did Max see yesterday.' [kor]
A: ku-nun/#ka ecey losa-lul po-ass-e.
he-NUN/NOM yesterday Rosa-ACC see-PST-DECL
'He saw Rosa yesterday.' [kor]

That is to say, though *he* in (14A) is an anaphoric element connecting to the discourse topic *Max*, it can function as the topic in at least one language with relatively clear marking of topic.

There is another question partially related to (13): Are there topicless sentences? There are two different viewpoints on this. For instance, Erteschik-Shir (2007) argues that every sentence has a topic, though the topic does not overtly appear, and the topic that covertly exists is the so-called stage topic. According to Erteschik-Shir's claim, topic is always given in sentences in human language, because topic is relevant to knowledge in possession of the hearer(s). In contrast, Büring (1999) argues that topic may be non-existent, in terms of sentential forms. Büring assumes that sentences, in terms of information structure, are composed of focus, topic, and background, and a sentence may be either an all-focus construction, a bipartite construction (i.e. lacking topic), or a tripartite construction consisting of all three components, including background. In fact, their arguments may not be different from each other, because Erteschik-Shir lays more focus on psychological status, and Büring focuses on form. Since not all entities that cognitively exist are overtly expressed in human language, there may be a topic which does not take place in the surface form. Furthermore, Erteschik-Shir also seems not to contest the argument that there are two different types of structures, such as bipartite and tripartite. This dissertation follows Büring's argument, because I am interested in mapping linguistic forms to information structure meaning.

3.3.2 Subtypes of Topic

Topics are divided into two subtypes: non-contrastive topic, and contrastive topic, given that contrast is one of the cross-cutting categories in information structure. Non-contrastive topic is renamed *aboutness-topic* in the current work, in line with Choi's claim that aboutness is the core concept of

regular topics. Contrastive topic has been relatively understudied in comparison with other components of information structure. Contrastive topic has been largely addressed in Japanese and Korean with special reference to *wa* and (*n*)*un* (a.k.a. topic markers) (Kuno, 1973; Choi, 1999). Additionally, Arregi (2003) argues that clitic left dislocation in Spanish and other languages is a syntactic operation to articulate contrastive topic.

The present work regards these two components as the major subtypes, but Féry and Krifka (2008) present another subtype: frame-setting topics. This terminology is not directly included into the taxonomy of information structure meanings (i.e. the type hierarchy of information structure) in this dissertation, but its linguistic constraints are incorporated into the information structure library. This is mainly because the linguistic behaviours of frame-setting topic are mostly syntactic, rather than semantic.

Frame-setting topic is language-universally associated with sentence-initial adjuncts (Lambrecht, 1996, p. 118) though not all sentence-initial adjuncts are necessarily frame-setting topics (i.e., the relation is not bidirectional). In other words, frame-setting topics have one constraint on sentence positioning; they should be sentence-initial. Chafe (1976) defines frame-setting topic as an element which sets "a spatial, temporal or individual framework within which the main predication holds". Féry and Krifka (2008) give an example of frame setting as shown in (15), in which the sentence talks about the subject *John*, but is only concerned with his health. Thus, the aboutness topic is assigned to *John*, but frame setting narrows down the aspect of description.

(15) Q: How is John?
A: {Healthwise}, he is FINE (Féry and Krifka, 2008, p. 128).

As mentioned above, the present work does not regard frame-setting topics as one of the cross-cutting components that semantically contribute to information structure. Since a frame-setting topic is close to a syntactic operation expressing information structure, hereafter it is called just a frame-setter. Various types of categories can be used as a frame-setter. First, the 'as for ...' construction in English serves as a frame-setter given that it has the same property as ordinary frame-setters as exemplified in (15) and (16).

(16) Q: How is John?
A: {As for his health}, he is FINE (Féry and Krifka, 2008, p. 128).

Second, adverbial categories (e.g. *Healthwise* in (15A)) can sometimes serve as frame-setters. For instance, regarding (i), in (17) in German, where *gestern abend* 'yesterday evening' and *körperlich* 'physically' are fronted, the frame-setting topic is assigned to the adverbials. This property is conjectured to be applicable to all other languages, according to the notion presented in Chafe (1976) and Lambrecht (1996).

(17) a. Gestern abend haben wir Skat gespielt.
yesterday evening have we Skat played
'Yesterday evening, we played Skat.'

b. Körperlich geht es Peter sehr gut.
Physically goes it Peter very well
'Physically, Peter is doing very well.' [ger] (Dipper et al., 2007, p. 169)

An adjunct NP can sometimes be a frame-setter, if and only if it appears in the sentence-initial position. In a Japanese sentence (18), the genuine subject of the sentence is *supiido suketaa* 'speed skater', while *Amerika* 'America' restricts the domain of what the speaker is talking about. Note that since frame-setters are realized as a topic, they are normally realized by the *wa*-marking.

(18) Amerika wa supiido suketaa ga hayai
America WA speed skater NOM fast
'As for America, the speed skaters are fast.' [jpn]

Therefore, the first NP combined with *wa* is interpreted as topic, whereas the second one with the nominative marker *ga*, which functions as the subject, does not convey topic meaning. Likewise, adjunct clauses which set the frame to restrict the temporal or spatial situation of the current discourse also have a topic relation with the main clause. Haiman (1978) and Ramsay (1987) argue that sentence-initial conditional clauses function as the topic of the whole utterance. The same goes for sentence-initial temporal clauses. For instance, in (19), taken from *The Little Prince* written by *Antoine de Saint-Exupéry*, the entire temporal clause *when he arrived on the planet* is dealt with as the frame-setter of the sentence.

(19) When he arrived on the planet, he respectfully saluted the lamplighter.

In sum, frame-setters, presumably across languages (Chafe, 1976; Lambrecht, 1996), must show up before anything else. The role of frame-setters is assigned to sentence-initial constituents, which

narrows down the domain of what the speaker is talking about. It can be assigned to various types of phrases including adverbs and even adjunct clauses. The syntactic restrictions on frame-setters are not reflected in the classification of topics, because the present work intends to provide a semantics-based classification of information structure components. The only semantic distinctions between frame-setter and other topics are orthogonal to information structure. That is, frame-setters are adjuncts, while other topics are mostly arguments. Both have the same type of role in the information structure.

3.3.3 Linguistic Properties of Topic

This subsection discusses several linguistic properties which should be taken into consideration in the creation of a computational model for realization of topics; (i) scopal interpretation relying on the topic relation, (ii) clausal constraints, (iii) multiple topics, and (iv) verbal topics.

Scopal Interpretation

Many previous studies argue that topics take wide scope. For instance, according to Büring (1997), if a rise-fall accent contour in German co-occurs with negation, the prosodic marking disambiguates a scopal interpretation. For example, (20a) would have two scopal readings if it were not for prosodic marking, but in (20b), in which '/' and '\' stand for rise and fall respectively, there is only a single available meaning.

(20) a. Alle Politiker sind nicht korrupt.
all politicians are not corrupt
(a) $\surd\forall>\neg$ 'For all politicians, it is not the case that they are corrupt.'
(b) $\surd\neg>\forall$ 'It is not the case that all politicians are corrupt.'

b. / ALLE Politiker sind NICHT \ korrupt.
all politicians are not corrupt
*$\forall>\neg$, $\surd\neg>\forall$ (Büring, 1997, p. 175)

Clausal Constraints

Topics can appear in non-matrix clauses, but there are some clausal constraints. Lambrecht (1996, p. 126) offers an observation that some languages mark the difference in topicality between matrix

and non-matrix clauses by morphosyntactic means. To take a well-known example, Kuno (1973) argues that the topic marker *wa* in Japanese tends not to be attached to NPs in embedded clauses, and Korean seems to have the same tendency in my intuition. Yet, a tendency is just a tendency. Some subordinate clauses are evaluated as containing topic. (21) presents two counterexamples from Korean.

(21) a. hyangki-nun coh-un kkoch-i phi-n-ta.
scent-NUN good-REL flower-NOM bloom-PRES-DECL
'A flower with a good scent blooms.' [kor]

b. Chelswuka-ka insayng-un yuhanha-tako malha-yss-ta.
Cheolsoo-NOM life-NUN limited-COMP say-PST-DECL
'Cheolsoo said that life is limited.' [kor] Lim (2012, p. 229)

First, Lim argues that (*n*)*un* can be used in a relative clause as given in (21a) when the (*n*)*un*-marked NP conveys a contrastive meaning (i.e. a contrastive focus in this case). The relative clause in (21a), which modifies the following NP *kkoch* 'flower', conveys a meaning like *The flower smells good, but contrastively it does not look so good.*, and (*n*)*un* attached to *hyangki* 'scent' is responsible for the contrastive reading. If *hyangki* is combined with a nominative marker *ka*, instead of *nun*, the sentence still sounds good as presented in (21′a), but the contrastive meaning becomes very weak or just disappear.

(21′) a. hyangki-ka coh-un kkoch-i phi-n-ta.
scent-NOM good-REL flower-NOM bloom-PRES-DECL
'A flower with a good scent blooms.' [kor]

Second, if the main predicate is concerned with speech acts (e.g. *malha* 'say') as provided in (21b), non-matrix clauses can have topicalized constituents.

The relationship between topic and clausal types has been discussed in previous literature with special attention to the so-called root phenomena (Haegeman, 2004; Heycock, 2007; Bianchi and Frascarelli, 2010; Roberts, 2011). Roberts, for instance, provides several English examples in which the topic shows up in the non-matrix clauses.

(22) a. Bill warned us that **flights to Chicago** we should try to avoid. (Emonds, 2004, p. 77)

b. It appears that **this book** he read thoroughly. (Hooper and Thompson, 1973, p. 478)

c. I am glad that **this unrewarding job**, she has finally decided to give up. (Bianchi and Frascarelli, 2010, p. 69)

These examples imply that left-dislocated constituents can appear in embedded clauses even in English, if the main predicate denotes speech acts (e.g. *warn* in (22a)), quasi-evidentials (e.g. *it appears* in (22b)), or (semi-)factives (e.g. *be glad* in (22c)).

This property that topics have in non-matrix clauses needs to be considered when I build up a model of information structure for multiclausal utterances. The relevant constraints are reexamined in Chapter 11 in detail.

Multiple Topics

Bianchi and Frascarelli (2010) argue that aboutness topics (A-Topics in their terminology) can appear only once, while other types of topics, such as contrastive topics (C-Topics), can turn up multiple times. This dissertation looks at the difference in terms of discrepancies between marking and meaning of information structure. As discussed previously at the beginning of this chapter, topic-marked elements may or may not occur in a single clause. Notably, they can appear multiple times as exemplified in (23).

(23) Kim-un chayk-un ilk-ess-ta.
Kim-NUN book-NUN read-PST-DECL
'Kim read the book.' [kor]

However, the (*n*)*un*-marked NPs in (23) do not carry the same status with respect to information structure. The (*n*)*un*-marked subject *Kim-un in situ* in (23) can be either an aboutness topic or a contrastive topic, because Korean is a topic-first language (Sohn, 2001). In contrast, the (*n*)*un*-marked object *chayk-un* has a contrastive meaning, because (*n*)*un*-marked non-subjects *in situ* is normally associated with contrastive focus (Choi, 1999; Song and Bender, 2011). In short, (*n*)*un*-marked (a.k.a. topic-marked) constituents can occur multiple times, but not all of them are necessarily associated with topic. Thus, the so-called topic marker is not an appropriate name at least for (*n*)*un* in Korean (and *wa* in Japanese). §5.1 provides more discussion on meanings that (*n*)*un* and *wa*-marked items in Korean and Japanese carry.

Verbal Topics

Topic-marking on verbal items is rare, but a cross-linguistic survey on information structure markings provides one exceptional case: Paumarí does employ a verbal topic marker, which cannot co-occur with a nominal topic marker in the language (Chapman, 1981). Therefore, this dissertation assumes that topic can be assigned to verbal items, and the possibility is one of the language-specific parameters. This property needs to be considered when I describe and implement the web-based questionnaire (§14.1). In the questionnaire, I let users choose a categorical constraint on focus and topic. Although topics are normally assigned to NPs, users are able to choose verbal markings of topic.

3.3.4 Tests for Topic

Given that the argument that aboutness is the semantic core of topics is supported by a lot of distributional findings as well as theoretic argument, Reinhart (1981) and Choi (1999) suggest a diagnostic to identify topic: namely the *tell-me-about* test. For instance, a reply to *Tell me about the dog.* will contain a word with the B-accent (L+H*) in English, such as *The **dog*** BARKS. This test can be validly used across languages. For example, the word corresponding to *tell-me-about* in Korean should not be with the case markers that have the non-topic relation.

(24) Q: ku kay-ey tayhayse malha-y cwu-e.
the dog-DAT about talk-COMP give-IMP
'Tell me about the dog.' [kor]

A: ku kay-#ka/nun cacwu cic-e
the dog-NOM/NUN often bark-DECL
'The **dog** often barks.' [kor]

Nonetheless, there are a few opposing claims (Vermeulen, 2009), and several tests have been devised that take notice of the relationship between topichood and aboutness. Roberts (2011), in line with Reinhart (1981) and Gundel (1985), provides four paraphrasing tests for topic in English as follows, which differ subtly in their felicity-conditions. If a left-dislocated NP, such as *Coppola* in *Coppola, he ...*, conveys a meaning of topic, the constituent can be paraphrased as at least one of the constructions presented below.

(25) a. **About** Coppola, he said that he found him to be ...

b. **What about** Coppola? He found him to be ...

c. **As for** Coppola, he found him to be ...

d. **Speaking of** Coppola, he found him to be ... (Roberts, 2011, p. 1916)

As Roberts explains, those tests may not be straightforwardly applicable to other languages because translations can vary in accordance with fairly delicate differences in felicity. Oshima (2009) suggests using the *as-for* test to test for the topic in Japanese, which can be translated into *ni-tsuite-wa*. The test can be defined as (26) with examples in Japanese given in (27). For example, *Ken-wa* in (27a) and *Iriasu-wa* (27c) are evaluated as containing topic meaning because they pass the *as-for* test.

(26) The *as for* test: If an utterance of the form: [$_{S_1}$ ... X ...] can be felicitously paraphrased as [*As for* X, S_2] where S_2 is identical to S_1 except that X is replaced by a pronominal or empty form anaphoric to X, X in S_1 is a topic. (Oshima, 2009, p. 410)

(27) a. Ken-wa Iriasu-o yomi-mashi-ta.
Ken-WA Iliad-ACC read-POLITE-PAST
'Ken read Iliad.' [jpn]

b. Ken-ni-tsuite-wa, Iriasu-o yomi-mashi-ta.
Ken-ni-tsuite-WA Iliad-ACC read-POLITE-PAST
'As for Ken, he read Iliad.' [jpn]

c. Iriasu-wa Ken-ga yomi-mashi-ta.
Iliad-WA Ken-NOM read-POLITE-PAST
'As for Iliad, Ken read it.' [jpn]

d. Iriasu-ni-tsuite-wa, Ken-ga yomi-mashi-ta.
Iliad-ni-tsuite-WA Ken-NOM read-POLITE-PAST
'As for Iliad, Ken read it.' [jpn] (Ibid. p. 410)

Given that aboutness is the semantic core of topic from a cross-linguistic view in order to to scrutinize aboutness, employs the diagnostics presented by Roberts (2011) as exemplified in (25). In this dissertation, topic is assumed to be assigned to an entity that can pass one of the paraphrasing tests in Roberts.

3.3.5 Summary

Topic refers to what the speaker is talking about. There are basically two subclasses of topics: aboutness topic and contrastive topic. Aboutness topic is non-contrastive and cannot appear more than once. Contrastive topic is assigned to constituents that contribute to an entailment of alternative sets and can be omitted. In addition, there is a syntactic operator that has to do with realization of topic: frame-setter. It appears only sentence-initially, and restricts the domain of what the speaker says. The tests for topics are the *tell-me-about* test for identifying topic-marking and the paraphrasing test to vet meaning of topic in a sentence.

3.4 Contrast

3.4.1 Definition

Contrast is treated as a cross-cutting information structure component in this dissertation, which contributes the entailment of an alternative set (Molnár, 2002; Krifka, 2008). Contrast can never show up out of the blue (Erteschik-Shir, 2007, p. 9). This is because the existence of an alternative set within the discourse is essential for contrast.

Lambrecht (1996) regards 'contrastiveness' as a merely cognitive concept, yet there are quite a few counterexamples to the claim from a cross-linguistic perspective which show contrast is a linguistic category. Some languages employ specific markers or syntactic means of expressing contrastiveness.

As with non-contrastive topic and focus, contrast may be marked by virtually any linguistic means (prosodic, lexical, and/or syntactic) across languages, and the same device may mark both non-contrastive and contrastive constituents. For example, Gundel (1999, p. 296) argues that placing a constituent in a specific sentence position (e.g. the sentence-initial position in English) can be used to mark either non-contrastive focus or purely contrastive focus. Topic can also have a contrastive meaning, and sometimes non-contrastive topic and contrastive topic share the same linguistic means as both non-contrastive focus and contrastive focus are prominently marked in the prosodic pattern in English. For example, Korean which employs (*n*)*un* can express contrastive topic as shown in (28); the answer conveys an interpretation like *I surely know Kim read a book, but I think Lee, contrastively, might not have.*

(28) Q: Kim-kwa Lee-nun mwuess-ul ilk-ess-ni?
Kim-and Lee-NUN what-ACC read-PST-INT
'What did Kim and Lee read?' [kor]

A: Kim-un chayk-ul ilk-ess-e.
Kim-NUN book-ACC read-PST-DECL
'Kim read a book.' [kor]

Some languages, unlike English and Korean, have a linguistic means marking contrast in a distinctive way from non-contrastive topic and focus. For instance, Vietnamese uses a contrastive-topic marker *thì* (Nguyen, 2006), exemplified in (29). The contrast function is shown by the alternative set evoked in (29), while the distinctiveness from focus is shown by the fact that *thì*-marked NPs cannot be used to answer *wh*-questions (Ibid.).

(29) Nam thì đi Hà Nội
Nam THI go Ha Noi
'Nam goes to Hanoi(, but nobody else).' [vie] (Nguyen, 2006, p. 1)

We can also find syntactic marking of contrast in several languages. In Standard Arabic, for instance, contrastively focused items are normally preposed to the sentence-initial position, while non-contrastively focused items which convey new information (i.e. semantic focus in Gundel (1999)'s terminology) are *in situ* with a specific pitch accent, as exemplified in (30a-b) respectively (Ouhalla, 1999).

(30) a. RIWAAYAT-AN ʔallat-at Zaynab-u
novel-ACC wrote-she Zaynab-NOM
'It was a NOVEL that Zaynab wrote.'

b. ʔallat-at Zaynab-u RIWAAYAT-an
wrote-she Zaynab-NOM novel-ACC
'Zaynab wrote a NOVEL.' [arb] (Ibid. p. 337)

Similarly, in Portuguese, contrastive focus precedes the verb, while non-contrastive focus follows the verb (Ambar, 1999). If *a tarte* 'a pie' conveys a contrastive meaning, it cannot be preceded by the verb *comeu* 'ate' as exemplified in (31A).

(31) Q: Que comeu a Maria?
what ate the Mary
'What did Mary eat?'

A: #A Maria comeu a tarte.
the Mary ate the pie.
'Mary ate the pie (What else she ate, I don't know.)' [por] (Ibid. p. 28–29)

In Russian, contrastive focus is preposed, while non-contrastive focus shows up clause-finally (Neeleman and Titov, 2009). For example, in (32), JAZZ-PIANISTA 'jazz-pianist' in initial position shows a contrast with *jazz-gitarista* 'jazz-guitarist'.

(32) JAZZ-PIANISTA mal'čiki slyšali vystuplenie (a ne jazz-gitarista).
jazz-pianist.GEN boys listened performance.ACC (and not jazz-guitarist.GEN)
'The boys listened to the performance of the jazz pianist.' [rus] (Neeleman and Titov, 2009, p. 519)

3.4.2 Subtypes of Contrast

Contrast can be used with either focus or topic, resulting in two subtypes: contrastive focus and contrastive topic. These may co-occur in a single clause. For example, in (33) taken from van Valin (2005), MARY and SALLY are contrastively focused, whereas **book** and **magazine** are contrastively topicalized.

(33) Q: Who did Bill give the book to and who did he give the magazine to?

A: He gave the **book** to MARY and the **magazine** to SALLY. (van Valin, 2005, p. 72)

3.4.3 Linguistic Properties of Contrast

In addition to the distributional facts presented in §3.4.1, which substantiate the existence of contrast as a component of information structure, there is also an argument that contrast behaves differently from non-contrastive focus (or topic) in the semantics.

Regarding the difference between contrastive focus and non-contrastive focus, Gundel (1999) provides several differences, as already presented in §3.2.2. The different behaviour between them is also exemplified in (34) taken from Partee (1991). (34a) can be ambiguously interpreted depending on where the accent is assigned. (34d), in which $_{CF}$ stands for a specific accent responsible for

contrastive focus, has the same truth-conditions as (34c), not (34b), because the city with the largest demonstrations in November 1989 was Budapest.

(34) a. The largest demonstrations took place in Prague in November (in) 1989.

b. The largest demonstrations took place in PRAGUE in November (in) 1989.

c. The largest demonstrations took place in Prague in NOVEMBER (in) 1989.

d. The largest demonstrations took place in $\text{PRAGUE}_{\text{CF}}$ in NOVEMBER (in) 1989. (Gundel, 1999, p. 301–302)

Nakanishi (2007), based on the taxonomy of Kuno (1973), compares contrastive topic with thematic topic (i.e. in Nakanishi's terminology, a.k.a. non-contrastive topic or aboutness topic in this dissertation) in Japanese from several angles, in that *wa* can be used for either the theme or the contrastive element of the sentence. From a distributional viewpoint, the non-contrastively *wa*-marked constituents can be either anaphoric (i.e. previously mentioned) or generic, whereas the contrastive element with *wa* can be generic, anaphoric, or neither. (35) is an example in which a contrastively *wa*-marked NP conveys neither an anaphoric interpretation nor a generic one.

(35) oozei-no hito-wa paatii-ni kimasi-ta
many-GEN people-WA party-to come-PST
ga omosiroi hito-wa hitori mo imas-en-desita.
but interesting people-WA one-person even be-NEG-PST
'Many people came to the party indeed but there was none who was interesting.'
[jpn] (Kuno, 1973, p. 270)

From a phonological stance, if *wa* is used for a thematic interpretation, the highest value of F0 contour after *wa* is as high as or even higher than the highest value before *wa*. In contrast, if it denotes a contrastive meaning, the former is much lower than the latter, which co-occurs with a dramatic downslope of F0 contour. From a semantic perspective, it turns out that the two version of the marker have different scopal interpretations when they co-occur with negation. The scopal interpretation driven by the relationship with focus and negation was originally captured by Büring (1997) as exemplified earlier in (20). Nakanishi, in line with the claim of Büring, compares two types of *wa*-marked topics in Japanese as shown in (36): Thematic *wa* in (36a) and contrastive *wa* in (36b) have the opposite scopal reading to each other.

(36) a. Minna-wa ne-nakat-ta.

everyone-WA sleep-NEG-PAST
'Everyone didn't sleep.'

(thematic *wa*) $\surd\forall>\neg$, $*\neg>\forall$

b. [Minna-wa]$_{T}$ ne-[nakat]$_{F}$-ta.

everyone-WA sleep-NEG-PAST
'Everyone didn't sleep.'

(contrastive *wa*) $*\forall>\neg$, $\surd\neg>\forall$ (Nakanishi, 2007, p. 187–188)

Compared to non-contrastive topics, contrastive topics tend to have relatively weak constraints on positioning and the selection of NP. Choi (1999) provides an analysis of scrambling (i.e. OSV) in Korean, which reveals that contrastive focus can freely scramble, while completive focus (a.k.a. non-contrastive focus or semantic focus) cannot scramble. Erteschik-Shir (2007) on the other hand argues that in Danish contrastive topic can be associated with non-specific indefinites, whereas non-contrastive topic cannot, as shown in (37). *En pige* 'a girl' in (37a) cannot play the non-contrastive topic role, because its interpretation is non-specifically indefinite. In contrast, *et museum* 'a museum' in (37b) can be the topic of the sentence, because it has an alternative *en kirke* 'a church'.

(37) a. #En pige mødte jeg i går.

a girl met I yesterday

'I met a girl yesterday.'

b. Et museum besøgte jeg allerede i går, en kirke ser jeg først i morgen.

a museum visited I already yesterday a church see I only tomorrow
'I visited a museum already yesterday, I will see a church only tomorrow.' [dan] (Erteschik-Shir, 2007, p. 8–9)

3.4.4 Tests for Contrast

Gryllia (2009) provides six tests to vet the meaning and marking of contrastive topic and contrastive focus as follows (p. 42–43).[12]

(38) a. *Wh*-questions: A contrastive answer is not compatible with a common *wh*-question.

b. Correction test: A contrastive focus can be used to answer a yes-no question correcting part of the predicate information of the question.

[12]For more elaborated explanation and examples for each of them, see Chapter 3 in Gryllia (2009). This subsection, for brevity, provides only the definition and representative examples focusing on correction test.

c. Choice test: When answering an alternative question, one alternate is contrasted to the other.

d. Accommodation focus test: When the discourse is accommodated in such a way that the initial *wh*-question can be interpreted as containing a positive and a negative question (e.g. "*who came?*" "*who did not come?*"), then the focus in the answer is contrastive.

e. Substitution test: If two terms are interpreted with a 'List Interpretation', then they can be substituted with 'the former' and 'the latter'.

f. Right dislocation: Contrast is incompatible with right dislocation.

g. Implicit subquestion test: (i) When a *wh*-question can be split into subquestions and the answer is organized per subquestion, then, there is a contrastive topic in the answer. (ii) When a question can be interpreted as containing more than one implicit subquestion, and the answer addresses only one of these subquestions, rather than the general question, then, this answer contains a contrastive topic.

Some of the diagnostics above, however, are not cross-linguistically valid. In non-Indo-European languages, such as Korean, only some work. For example, the *wh*-question test does not work for English and Korean in the same manner, as exemplified below.

(39) Q: Who came?

A: Well KIM came, I know that much, but I can't tell you about anyone else.

(40) Q: nwuka o-ass-ni?
who come-PST-INT
'Who came?' [kor]

A: Kim-i/un o-ass-e.
Kim-NOM/NUN come-PST-DECL
'Kim came.' [kor]

Kim with (*n*)*un* in (40A) can be an appropriate answer to the question, and it involves a contrastive interpretation (i.e. conveying a meaning like *I know that at least Kim came, but I'm not sure whether or not others came.*). In this case, the replier alters the information structure articulated by the questioner arbitrarily in order to offer a more informative answer to the solicited question. Note that contrastiveness is basically speaker-oriented (Chang, 2002). In other words, contrast is primarily

motivated by the speaker's necessity to attract the hearer's special attention at a particular point in the discourse. Thus, speakers may change the stream of information structure as they want.[13]

The right dislocation test, on the other hand, seems valid in Korean as well. The (*n*)*un*-marked NP can be used in the right dislocation constructions in Korean as given in (41Q1). Yet, if an alternative set is entailed as shown in (41Q2), right dislocation sounds absurd.

(41)Q1: cangmi-nun ettay?
rose-NUN about
'How about the rose?' [kor]

A: coh-a, cangmi-nun.
good-DECL rose-NUN
'It's good, the rose.' [kor]

Q2: kkoch-un ettay?
flower-NUN about
'How about the flowers?' [kor]

A1: #coh-a, cangmi-nun.
good-DECL rose-NUN [kor]

A2: cangmi-nun coh-a.
rose-NUN good-DECL [kor]

The most convincing and cross-linguistically applicable tests among the tests in (38) is the correction test, as exemplified in (42) in Italian (43) in Greek, and (44) in Korean.

(42) Q: L' ha rotto Giorgio, il vaso?
it has broken Giorgio the vase
'Has Giorgio broken the vase?' [ita]

A: $[\text{Maria}]_{\text{C-Foc}}$ ha rotto il vaso.
Maria has broken the vase
'It is Maria who has broken the vase.' [ita] (Gryllia, 2009, p. 32)

[13] Young Chul Jun, p.c.

(43) Q: Thelis tsai?
want.2SG tea.ACC
'Would you like tea?' [ell]

A1: Ohi, thelo [kafe]$_{\text{C-Foc}}$.
no want.1SG coffee.ACC
'No, I would like coffee.' [ell]

A2: Ohi, [kafe]$_{\text{C-Foc}}$ thelo
no coffee.ACC want.1SG [ell] (Ibid. p. 44)

(44) Q: chayk ilk-ess-ni?
book read-PST-INT
'Did you read a book?' [kor]

A: ani, capci-lul/nun ilk-ess-e.
no magazine-ACC/NUN read-PST-DECL
'No, (but) I read a magazine.' [kor]

Gussenhoven (2007), in a similar vein, suggests corrective focus as a subtype of focus in English, which appears when a constituent bears focus meaning with an alternative (i.e. contrast) as presented below.

(45) A: What's the capital of Finland?

B: The CAPital of FINland is [HELsinki]$_{\text{FOC}}$

A′: The capital of Finland is OSlo.

B′: (NO.) The capital of Finland is [HELsinki]$_{\text{CORRECTIVE}}$ (Gussenhoven, 2007, p. 91)

Gussenhoven also provides a similar example in Navajo. Navajo has two negative modifiers; one is neutral, *doo ... da* in (46a), and the other expresses corrective focus, *hanii* in (46b). That is, *hanii* serves to mark a contrastive focus in Navajo.

(46) a. Jáan doo chidí yiyíiłchø'-da.
John NEG car 3RD-PAST-wreck-NEG
'John didn't wreck the car.' [nav]

b. Jáan hanii chidí yiyíiłchø'.
John NEG car 3RD-PAST-wreck
'JOHN didn't wreck the car (someone else did).' [nav] (Ibid. p. 91)

Wee (2001) proposes the conditional test, which means contrastive topic can be paraphrased into a conditional clause as exemplified in (47B″).[14] That is, (*n*)*un* which can convey contrast meaning in Korean can be altered into a conditional marker *lamyen*, which also has an alternative set drawn by *nobody* in (47A) and functions to make a correction to the presupposition given in (47A).

(47) A: Nobody can solve the problem.

B: Peter would solve the problem.

B′: Peter-nun ku muncey-lul phwul-keya.
Peter-NUN the problem-ACC solve-would
'Peter would solve the problem.' [kor]

B″: Peter-lamyen, ku muncey-lul phwul-keya.
Peter-if the problem-ACC solve-would
'If Peter were here, he would solve the problem.' [kor]

von Fintel (2004) and Kim (2012a) suggest a test for contrast called "Hey, wait a minute!", which serves to cancel or negate presupposed content in the previous discourse. In other words, the contrastive marking acts as the key for correcting the inaccurate part in a presupposition. Likewise, Skopeteas and Fanselow (2010), exploring focus positions in Georgian, define contrastive focus as a "corrective answer to truth value question". This definition is also in line with my argument that the correction test can be reliably used to vet contrastive focus.

This dissertation makes use of the correction test to scrutinize contrast. However, that means that recognizing corrections is the only use of contrastive focus. Note that use for corrections is a sufficient condition for expressing contrastive focus, but not a necessary condition.

Lastly, because foci are inomissible while topics are not, if a constituent that passes the correction test cannot be elided, it is evaluated as conveying contrastive focus. If a constituent passes the correction test but can be dropped, it is regarded as contrastive topic.

---

[14] Hae-Kyung Wee, p.c.

3.4.5 Summary

I regard contrast as one of the cross-cutting components that participate in information structure, because some languages employ a specific means of expressing contrast. Contrast contributes to the entailment of an alternative set in a given context. For this reason, contrast cannot show up out of the blue. There are two subtypes: namely contrastive focus and contrastive topic. While contrastive focus cannot be eliminated from a sentence, contrastive topic can be dropped. The cross-linguistically valid test for identifying contrast is the correction test.

3.5 Background

We can say a constituent is in the background when it conveys a meaning of neither focus nor topic. In terms of linguistic forms, it typically does not involve additional marking but may be forced into particular positions in a sentence. Background is in complementary distribution to both topic and focus; thus, it adds no information structure meaning to the discourse. Focus, topic, and background are mutually exclusive, and thereby they cannot be overlap with each other.

Background can often be found in cleft sentences. Clefts refer to (copula) constructions consisting of a main clause and a dependent clause (e.g. a relative clause), in which a constituent in the main clause is narrow-focused.[15] The narrow foci in cleft constructions can be easily identified by means of deletion test. As noted before, focus means a constituent that can never be elided, which is one of the main distinguishable behaviours that focus has from topic and background. Thus, any other constituent in (48-49), except for the narrowly focused ones *Kim* and *from her*, can be freely eliminated.

(48) Q: Who reads the book?
A1: It is Kim that reads the book.
A2: It is Kim.
A3: Kim.

[15]The focused item in clefts does not need to be an argument focus, because non-nominal categories such as adverbs and PPs can sometimes take place in the main clause of clefts as given in (49).

(49) Q: Where did you have my address?

A1: It was from her that I had your address.

A2: It is from her.

A3: From her.

Clefts typically put the part of the sentence after the focused item into background. Since the remaining part of sentence (a.k.a. cleft clause) such as *that reads the book* and *that I had your address* in each cleft sentences can be freely dropped, they can be regarded as either topic or background. Moreover, the constituents in cleft clauses are rarely (*n*)*un*-marked in Korean, as shown in (50).

(50) a. ku chayk-ul/*un ilk-nun salam-i/un Kim-i-ta.
the book-ACC/NUN read-REL person-NOM/NUN Kim-COP-DECL
'It is Kim that reads the book.' [kor]

b. Kim-i/*un ilk-nun kes-i/un ku chayk-i-ta.
Kim-NOM/NUN read-REL thing-NOM/NUN the book-COP-DECL
'It is the book that Kim reads.' [kor]

As discussed thus far, (*n*)*un* in Korean assigns either topic or contrast, or both (i.e. contrastive topic) to the word it is attached to. Yet, (*n*)*un* cannot be used within the cleft clauses as shown in (50). Thus, NPs in the cleft clauses are usually identified as background (i.e. non-focus and non-topic, simultaneously), at least in Korean. However, cleft clauses can contain a focused constituent in some languages as exemplified in (51). Thus, we cannot say cleft clauses are always in background. More discussion about cleft clauses is given in Chapter 12 (§12.4.3).

(51) Q: Does Helen know JOHN?

A: It is John/JOHN she DISLIKES.

Q: I wonder who she dislikes.

A: It is JOHN she dislikes. (Gussenhoven, 2007, p. 96)

3.6 Summary

This chapter reviewed the primary components of information structure (focus, topic, contrast and background), including definitions of the concepts and explorations of sub-classifications, associated linguistic phenomena and potential tests for focus, topic and contrast. First, information status

is not reliable means of identifying information structure. The relationship between the two is a tendency. Second, I define focus as what is new and/or important in a sentence, and specify that a constituent associated with focus cannot be eliminated from the sentence. There are two subtypes of focus; one is semantic focus, lacking a contrastive meaning, and the other is contrastive focus. Tests to vet focus marking and meaning include *wh*-questions and the deletion test. Third, topic is defined as what a speaker is talking about. While every sentence presumably has at least one focus, topic may or may not appear in the surface form. There are two subtypes, which include aboutness topic (a.k.a. thematic topic or non-contrastive topic) and contrastive topic. Frame-setters, serving to restrict the domain of what is spoken (temporal, spatial, conditional, manner, etc.), are always external (not an argument of the predicate) and sentence-initial. Given that the semantic core of topic is aboutness, the tools for identifying topics are *tell-me-about* test and several paraphrasing tests such as *as for ...*, *speaking of ...*, and *(what) about ...*. Next, contrast always entails an alternative set, which can be realized as either contrastive focus or contrastive topic. The most reliable and cross-linguistically valid diagnosis for contrast is the correction test, because correction necessarily requires an alternative. Finally, background is neither focus nor topic, and any constituent associated with it can be freely elided without loss of information delivery. These cross-linguistic findings help create an annotation guideline in the corpus study (Part III), provide linguistic generalizations to be used in creating HPSG/MRS-based constraints on information structure. Moreover, they are also used to design the library of information structure for the LinGO Grammar Matrix system.

Chapter 4

MARKINGS OF INFORMATION STRUCTURE

The main goal of this chapter is to find the range of possible expressions with respect to information structure. Different languages employ different marking systems for expressing information structure. The linguistic means of conveying information structure meanings includes: (i) prosody, (ii) lexical markings, (iii) syntactic positioning, and (iv) combinations of these (Gundel, 1999). This chapter explores how these meanings are specifically realized in various languages, and finds a cross-linguistic generalization about the marking system in a systemic manner. This finding not merely contributes to typological studies of human languages, but also carries weight with implementing a grammar library for information structure. Because the users' input on the LinGO Grammar Matrix customization system refers to the linguistic forms in their languages (Bender and Flickinger, 2005; Drellishak, 2009; Bender et al., 2010b), it is important for creation of a library to systematize linguistic realization in a fine-grained way.

This chapter is structured as follows. §4.1 begins with an explanation of my methodology for gathering data related to information structure markings and systematizing them. §4.2 addresses prosodic marking of expressing information structure. The present work does not directly implement constraints on prosodic patterns into the system, but this dissertation presents flexible representation for them to set the ground work for further developed system. §4.3 looks into lexical markers responsible for focus and topic from a cross-linguistic viewpoint. These are classified into three subclasses: affixes, adpositions, and modifiers. §4.4 surveys positioning constraints on information structure components in human language.

4.1 Methodology

My method consists of three steps: (a) referencing to three types of sources, (b) listing means of expressing information structure in each language, and (c) drawing a taxonomy of the marking system in a bottom-up way on the basic of the collected data.

The initial step was data collection. (a) My sources included (a-i) previous literature on information structure, (a-ii) reference grammar books, and (a-iii) human consultants. (a-i) There is a wealth of previous work analyzing realization of focus and topic, which provides the basic data for my generalization. (a-ii) Since 2010, grammar developers of *Language CoLLAGE* (`http://www.delph-in.net/matrix/language-collage`) have examined how information structure is articulated in their languages. Their languages are highly significant for understanding the systemized nature of human languages (Bender, 2007). In order to include insights from these grammars into my model, I first each developer's survey of information structure. I then went through any literature they referenced (a-iii) When two or more claims about the linguistic realizations of focus and topic in a language were in discord with each other, I consulted linguists who spoke the language in question as a mother tongue.

(b) After collecting data, I created tables indicating which forms are used in which languages to express information structure. Each table includes (b-i) name of the language, (b-ii) ISO 639-3 code, (b-iii) the source(s) I referred to, (b-iv) prosodic marking, (b-v) lexical marking, and (b-vi) syntactic marking. In each table, there are twelve cells to be filled out for (b-iv) to (b-vi): four meanings (non-contrastive focus, contrastive focus, non-contrastive topic, and contrastive topic) and three possible means (prosody, lexicon, and syntax). I found no language in which all the cells were filled out. There are two main reasons for empty cells, information that is unknown and and categories that are not applicable for a given language. For many languages, a specific way of marking focus, topic, and contrast remains unknown. If I did not find any information about a potential slot in the table, I left the cell empty, meaning 'unknown'. On the other hand, 'not applicable' means the marking system is not used in the language. For example, prosody is not responsible for expressing focus in some languages.

(c) Finally, I constructed a cross-linguistic taxonomy of information-structure markings using the collected tables. According to Gundel (1999) and Féry and Krifka (2008), there are three ways to mark information structure in human language: namely (c-i) prosody, (c-ii) lexical markers, and (c-iii) syntactic positioning. These three are added into the taxonomy as the main branches. The taxonomy was crosschecked by Büring (2010)'s typological view. Büring classifies languages into six subgroups in terms of focus realization: (i) boundary languages (e.g. Chicheŵa, Bengali, Japanese, and English), (ii) (relaxed) edge languages (e.g. Spanish, Italian, and Hungarian), (iii)

strict position languages (e.g. Armenian, Basque, Georgian, etc.), (iv) mixed languages (e.g. most of Slavic languages, etc.), (v) particle languages (e.g. Chickasaw), and (vi) non-marking languages (e.g. Hausa).

Ultimately, all the data from the original tables was included, and then I established a hierarchical relationship among them. The languages I have surveyed hitherto include 46 languages in a variety of language families plus English. They are presented in Appendix B and Appendix C.

4.2 Prosody

In much of the previous work on this topic, prosody has been presumed to be the universal means of marking information structure (Gundel, 1999; Büring, 2010, among many others). Many previous papers have studied information structure with special reference to how prosody marks information structure. Bolinger (1958) argues that there are two types of pitch accents in English; the A and B-accents (i.e. H* and L+H* in the ToBI format respectively). Jackendoff (1972) creates a generalization about the correlation between pitch accents and information structure components: The A and B-accents in English are responsible for marking constituents as focus and topic respectively.

(1) If a phrase P is chosen as the focus of a sentence S, the highest stress in S will be on the syllable of P that is assigned highest stress by the regular stress rules. (Jackendoff, 1972, p. 247)

The way in which A and B accents structure information is exemplified in (2), in which SMALL CAPS stands for the A-accent **boldface** does for the B-accent, respectively. The constituent semantically associated with aboutness bears the B-accent in English. Note that aboutness is regarded as the locus of realization of topic in this dissertation. On the other hand, the constituent corresponding to the *wh*-word in the question bears the A-accent, which gives a focus meaning.

(2) Q: What about Kim? What did Kim read?
A: **Kim** read the BOOK.

In the following subsections, from three points of view, I enter into the details about incorporating prosodic information into grammatical structures. This is done with an emphasis on application in the creation of an information structure library as a tool for grammar engineering.

4.2.1 Prosody as a Widespread Means of Marking

Since Jackendoff (1972), quite a few studies have tried to capture a connection between prosodic patterns and information structure in many languages, including English (Steedman, 2000), German (Büring, 2003), Portuguese (Frota, 2000), Japanese and Korean (Ueyama and Jun, 1998), and so on. Then, do we have to assume that every language employs prosody for marking information structure? Is it a language-universal way to express focus and/or topic? The answer is no, and in fact there are some counterexamples.

My cross-linguistic survey of the literature revealed that some languages have no means of expressing information structure through prosody. For instance, it is reported that Yucatec Maya employs no prosodic marking for expressing information structure. Instead, syntactic functions indicate these relations without an interaction with prosody (Kügler et al., 2007). In Akan, prosoodic patterns also have little to do with expressing focus, and instead a focused item must occupy the clause-initial position with one of several morphological markers (Drubig, 2003). Likewise, Catalan, in which syntactic operation is responsible for marking information structure, has a rather weak (or even null) correlation between prosody and information structure meanings (Engdahl and Vallduví, 1996). Hence, the assumption that prosody is a language-universal means of marking information structure is not valid. That is to say, using prosody for expressing information structure is clearly widespread, but not universal (Drellishak, 2009).

4.2.2 Conditions between Prosody and Information Structure

There seems to be no clear consensus with respect to mappings between prosody and information structure even in English. Contra to Jackendoff's claim, (i) Kadmon (2001), Büring (2003), and Oshima (2008) argue that the B-accent is specifically responsible for contrastive topic, rather than topic in a broad sense. (ii) Steedman (2000), on the other hand, argues that the B-accent marks theme. Steedman additionally associates information structure meanings with boundary tones. (iii) Hedberg (2006) regards the B-accent as a contrastive marker for both focus and topic (i.e. either contrastive focus or contrastive topic). (iv) More recently, Constant (2012) explores how semantic and pragmatic behaviour is influenced by a specific prosodic pattern 'rise-fall-rise' in English. The intonation contour can be transcribed in the ToBI format as [L*+H L- H%], as illustrated in (3).

That is, there are three components: The first 'rise' corresponds to [L*+H], 'fall' to [L-], and the second 'rise' to [H%].[1]

(3) A: Why isn't the coffee here?
B: I don't know. I was*expecting* there to be coffee ...
L*+H L- H% (Constant, 2012, p. 409)

Constant investigates the correlations between 'rise-fall-rise' intonation and contrastive topic intonation that Büring, Büring, and Oshima, proposes. Constant denies the previous attempt to unify the two which had claimed the former was a subclass of the latter.

Amongst the varied claims, I follow Hedberg (2006)'s argument for restricting two types of pitch accents in English. This is because Hedberg's classification is firmly based on an acoustic analysis of naturally occurring spoken data (Hedberg and Sosa, 2007), and empirical evidence is important for the creation of accurate models.

The debate presented above is largely concerned with which prosodic pattern has which effect on information structure. That is, they delve into is the nature of mapping between prosody and information structure. However, there exist some circumstances in which prosody is not involved in articulation of information structure (even in English).

Féry and Krifka (2008) argue prosodic patterns are not obligatorily related to information structure even in English. For example, the association between prosody and focus can be canceled in certain contexts, such as with Second Occurrence Focus. A second occurrence focus is an expression that falls within the scope of a focus sensitive operator (e.g. *only* in English), but is a repeat of an earlier focused occurrence (Partee, 1999; Beaver et al., 2007; Féry and Ishihara, 2009). The repeatedly focused item prosodically differs from the previously focused one (i.e. ordinarily focused); it is normally devoid of a specific pitch accent responsible for marking focus. Given that *vegetables* in (4b) is combined with a focus sensitive item *only*, it would be interpreted as containing focus meaning, but that meaning is already given in (4a). (4) is a clear counterexample to Halliday (1967)'s claim that what is focused should carry new information in that 'vegetables' in (4b) is already mentioned. In addition, while the *vegetables* in (4a) bears an A-accent, the repeated one in (4b) lacks

[1]The main argument Constant (2012) provides is that the 'rise-fall-rise' intonation involves a regular conventional implicature, acting as a focus sensitive quantifier over assertable alternative propositions. Conventional implicature is not treated in this dissertation. For more information about it, see Lee (2003b).

the pitch accent. According to Féry and Krifka, "there are only weak correlates of accent, and no pitch excursions in the postnuclear position" (Ibid. p. 132). This means that the focus meaning in this case is not directly invoked by the A-accent.

(4) a. Everyone already knew that Mary only eats [vegetables]$_{F}$.

b. If even [Paul]$_{F}$ knew that Mary only eats [vegetables]$_{SOF}$,
then he should have suggested a different restaurant. (Partee, 1999, p. 215–216)

This indicates that prosodic patterns are not reliable enough to establish a rigorous rule to restrict information structure. In other words, prosodic prominence is merely a tendency; it is neither a sufficient nor a necessary condition for conveying information structure meanings even in languages whose markings are largely dependent on prosody (e.g. English) (Rochemont, 1986; Drubig, 2003).

4.2.3 Flexible Representation

Although the correlates between prosody and information structure are controversial, it is true that prosody is the most widespread way of expressing information structure in many languages. It is necessary, then, to think of how to represent prosodic information into the formalism for creating computational model of information structure. Given that our processing system is usually text-based, currently it is almost impossible for us to resolve the phonological patterns of sentences, including intonation contour and pitch accents. Kuhn (1996) in the same context suggests an under-specified representation for information structure, noting that even prosodic marking of information structure often yields ambiguous meanings, which cannot in general be resolved in sentence-based processing.

4.2.4 Summary

My position to prosodic marking in the current work is as follows: It is clear that prosody makes a contribution to information structure in many languages, even if the relationship between prosodic marking and information structure is complicated. However, in some contexts, especially processing of texts that were originally written (rather than transcribed speech), we do not have access to prosodic information anyway. The best way to handle prosodic marking is to allow for underspeci-

fication in such a way that prosodic information can be later added into the formalism. In principle, this would allow for refining the representation monotonically.

4.3 Lexical Markers

According to my cross-linguistic survey, there are three subtypes of lexical markers that assign information structure roles; (i) affixes, (ii) adpositions, and (iii) modifiers.

Quite a few languages have specific affixes to signal focus, topic, and contrast, as exemplified in the following Rendile[2] examples, in which two affixes are used to express an argument focus (i.e. *é* by an enclisis process) and a predicate focus (i.e. *á* by a proclisis process) respectively (Lecarme, 1999).

(5) a. ínam-é yimi
boy-FOC came
'THE BOY came.' [rel]

b. ínam á-yimi
boy FOC-came
'The boy CAME.' [rel] (Lecarme, 1999, p. 277)

Some languages use affixes responsible for topic meanings; for instance, (*n*)*un* in Korean is used to signal information structure meanings (*contrast-or-topic* in this dissertation), and is in complementary distribution with ordinary case morphemes (e.g. *i*/*ka* for nominatives, (*l*)*ul* for accusatives).[3]

(6) ku kay-nun cic-e
DET dog-NUN bark-DECL
'The **dog** barks.' [kor]

Unlike the focus affixes used in (5) (i.e. *é* and *á*) which directly signal the information structure roles of the constituent, (*n*)*un* in Korean is not deterministic. The word which (*n*)*un* is attached to can be ambiguously interpreted. This is addressed in §5.1 in detail.

---

[2]Cushitic, Afro-Asiatic, spoken in northern Kenya

[3]In spite of morphosyntactic similarity, the lexical markers in Korean, such as (*n*)*un*, *i*/*ka*, and (*l*)*ul*, are treated as suffixes in KRG (Korean Resource Grammar, Kim et al. 2011), while those in Japanese, such as *wa*, *ga*, and *o*, are dealt with as separate words in Jacy (Siegel and Bender 2002).

Clitics are also often employed to express information structure, too. A clitic, which is somewhere between morpheme and word, is a lexical item that is syntactically independent, but phonologically dependent. Clitics used for information structure markings can be subclassed into two types; one is an adposition, and the other is a modifier. Adpositions, for example, are responsible for information structure markings in Japanese (i.e. the case markers, such as *ga* for nominatives and *o* for accusatives, and *wa* for contrast or topic).

(7) inu wa hoeru.
dog WA bark
'The **dog** barks.' [jpn]

On the other hand, clitics that have nothing to do with case marking can also be used as lexical markers for information structure. They are regarded as modifiers in this dissertation. For instance, Man (2007) presents two types of Cantonese lexical particles that modify an NP for marking information structure roles: *aa4* and *ne1* as the topic marker and *aa3*, *laa1*, and *gaa3* as the focus marker, respectively.

(8) a. nei1 bun2 syu1 **aa4** ngo5 tai2gwo3 hou2do1 ci3
DEF CLF book PART 1.SG read.EXP many times
'As for this book, I have read it for many times.' [yue]

b. keoi5 **aa3** bun2 syu1 ngo5 bei2zo2
3.SG PART CLF book 1.SG give.PERF
'It is him/her who I have given the book to.' [yue] (Man, 2007, p. 16)

There are many examples in which clitics are made use of to designate the topic and/or the focus in other languages, too. Cherokee,[4] for example, employs a clitic *=tvv* as the focus marker, which immediately follows the first word of the sentence as shown below (Montgomery-Anderson, 2008).

(9) a. ayv=tvv yi-tee-ji-hnooki
1PRO=FC IRR-DST-1A-sing:IMM
'I amd going to sing it.' [chr]

b. noókwu=tvv ji-tee-a-asuúla-a
now=FC REL-DST-3A-wash.hands:IMM-IMM
'He just washed his hands.' [chr] (Montgomery-Anderson, 2008, p. 152)

[4]a native American language (Iroquoian), still spoken in Oklahoma and North Carolina

From the cross-linguistic facts presented so far, this dissertation suggests lexical markers expressing information structure consist of three subitems; (i) affixes, (ii) adpositions, and (iii) modifiers.[5]

The differences among them are as follows: First, (i) affixal markers such as (*n*)*un* in Korean always behave dependently within the morphological system as shown in (6). In contrast, adpositions (e.g. lexical markers in Japanese) and modifiers (e.g. particles in Cantonese (8) and Cherokee (9)) are dealt with as separate words in the language. Second, if a language employs a non-affixal marker to express information structure, there are two options: (ii) If a non-suffixal marker is used to express information structure and the language employs adpositions, the marker is regarded as an adposition, too. In other words, when the language makes use of case-marking adpositions, and the adpositions are in complementary distribution with a lexical marker responsible for information structure markings in the language (as in Japanese), the marker is subtyped as an adposition. (iii) Otherwise, the lexical marker is regarded as a modifier.

According to my survey, there are four constraints on lexical markers for information structure. They are presented in the following subsections.

4.3.1 Multiple Markers

Human languages can have two or even more lexical markers for expressing either focus or topic, and the markers have different syntax from each other. Turning back to the Rendile example (5), *é* is used for nominals, while *á* is a verbal focus marker. There are similar cases in other languages, too: For example, Akan employs two focus markers; one is *na* that appears only in sentential replies, and the other is *a* that shows up only with a short answer. (Drubig, 2003, p. 4).

(10) Q: Hena na Ama rehwehwɛ?
who FOC Ama is-looking-for
'Who is it that Ama is looking for?' [aka]

A1: Kofi na *(Ama rehwehwɛ)
Kofi FOC Ama is-looking-for [aka]

[5]Somebody may claim that what I regard as an adposition in a given language is actually a modifier or vice versa. However, I am concerned with finding the full range of potential ways to mark information structure. This enables the users of the LinGO Grammar Matrix system to have flexibility in describing what they see in their language following Poulson (2011)'s meta-modeling idea.

A2: Kofi a (*Ama rehwehwɛ)
Kofi FOC
'(It is) KOFI (that Ama is looking for)' [aka] (Ibid. p. 5)

Sometimes, the lexical markers can be multiply used at the same time with some constraints: Schneider (2009) argues that Abma has four markers of expressing information structure, which include *ba* as a comment marker, and *tei* as a focus marker. Both *ba* and *tei* can appear before the predicate to designate comment plus focus (i.e. predicate focus), but the latter should be immediately preceded by the former as presented in (11) below.

(11) ... ba tei te ba=i=te Liwusvet=nga.
COMM FOC 3SG.PFV NEG.1=be=PART Liwusvet=NEG.2
'... but it wasn't Liwusvet.' [app] (Ibid. p. 5)

4.3.2 Positioning Constraint

The second one is related to the position of lexical markers. They can occur either following or preceding (or both) a phrase that is assigned an information structure role by the markers. For instance, in Rendile, *é* in (5a) is a suffix, and *á* in (5b) is a prefix. (12) is an example in Buli, in which the focus marker *kà* precedes the the focused constituent. In contrast, the focus marker *nyā* in Ditammari is preceded by the focused constituent, as shown in (13). Both languages belong to the language family of Niger-Congo/Gur. Those examples show us that the position of markers expressing information structure can vary from language to language.

(12) Q: What did the woman eat?

A: ò ŋɔ̀b kà túé.
3.SG eat FM beans
'She ate BEANS.' [bwu] (Féry and Krifka, 2008, p. 133)

(13) Q: What did the woman eat?

A: ò dī yātũrà nyā.
3.SG eat beans FM
'She ate BEANS.' [tbz] (Ibid. p. 133)

4.3.3 Categorical Restriction

There is a categorical restriction on the phrases with which lexical markers can be combined. Phrases can be nominal, verbal, and even adverbial; for instance, adverbial categories in Korean and Japanese can be *wa* and (*n*)*un*-marked. However, not all content words are necessarily able to be combined with the lexical markers in the lexical marking types of languages, which means whether or not a category can be lexically marked with information structure values has to be handled as a language-specific variable. Choice of lexical markers can also be dependent on category; in Rendile as shown in (5), an affix *é* is attached to only nouns such as *ínam* 'boy', while a prefix *á* is exclusively used with verbs such as *yimi* 'came'. That means, each lexical marker has a constraint on which category it can be used for, which also needs to be represented as lexical information.

4.3.4 Interaction with Syntax

In some languages that employ lexical markers for expressing information structure, the lexical markers have an interaction with syntactic operation. One well known case of this interplay between lexical markers and syntactic positioning is scrambling constructions in Korean and Japanese (Choi, 1999; Ishihara, 2001). In Akan, focused items obligatorily (i) occupy sentence-initial position and (ii) immediately precede focus markers such as *na* and *a* as already illustrated in (10) (Drubig, 2003, p. 4). A similar phenomenon can be found in the Buli example (12), too. According to Féry and Krifka (2008), if a focused constituent is sentence-initial, the focus marker *kà* is optionally used. Cherokee, as mentioned before with example (9), employs a clitic *tvv* to signal focus, and the focused constituent with *tvv* should be followed by any other constituents in the sentence (i.e. it is clause-initial).

4.4 Syntactic Positioning

Information structure roles are often associated with specific positions in a clause. As is well-documented, the realization of information structure has much to do with word order (van Valin, 2005; Mereu, 2009, among many others). The relationship between information structure and variations in word order can be cross-linguistically captured. For example, although word order in Spanish is relatively free in comparison with English, there are still ordering constraints in Spanish

that hinge on information structure (Zubizarreta, 1998; Zagona, 2002). Moreover, every language presumably has one or more syntactic device(s) for expressing information structure (Li and Thompson, 1976).

Before discussing specific positions that each component occupies, it is necessary to look into how information is structured in the basic word order in a language. Languages have different unmarked focus positions, depending largely, but not entirely, on their neutral word order. For example, in English, narrow focus on the object is a case of unmarked narrow focus, while narrow focus on the subject is a case of marked narrow focus. An ordinary example of a narrow focus can be found in Q/A pairs in which the object plays the role of focus as provided in (14).

(14) Q: What did Kim read?
A: Kim read the BOOK.

van Valin (2005) captures a generalization about the relationship between word order type and the most unmarked position of narrow focus: In SVO languages, it is the last position in the core clause (e.g. English) or the immediate postverbal position (e.g. Chicheŵa). In verb-final languages, the unmarked focus position is the immediate preverbal position (e.g. Korean and Japanese). In VOS languages, it is the immediate postverbal position (e.g. Toba Batak).

In this dissertation, I do not place an information structure constraint on sentences in the unmarked word order for two reasons.

First, the clause-initial items in subject-first or V2 languages are ambiguous when it comes to focus/topic fronting. For instance, note (15) in Yiddish. Given that declarative clauses in Yiddish are both SVO and V2 (Jacobs, 2005), the constituent that occurs in the sentence-initial position is the subject in the default word order. What is to be considered at the same time is that focus/topic fronting is very productively used in Yiddish as exemplified below (Jacobs, 2005).

(15) a. Der lerər šrajbt di zacn mit krajd afn tovl.
The teacher writes the sentences with chalk on the blackboard (neutral) [ydd]

b. Di zacn šrajbt der lerər mit krajd afn tovl.
the sentences writes the teacher with chalk on the blackboard
'It's the sentence (not mathematical equations) that the teacher is writing with chalk on the blackboard.' [ydd]

c. mit krajd šrajbt der lerər di zacn afn tovl.

with chalk writes the teacher the sentences on the blackboard
'It's with chalk (not with a crayon) that that the teacher is writing the sentence on the blackboard.'
[ydd]

d. afn tovl šrajbt der lerər di zacn mit krajd.

on the blackboard writes the teacher the sentences with chalk
'It's on the blackboard (not the notepad) that that the teacher is writing the sentence with chalk.'
[ydd] (Jacobs, 2005, p. 224)

Thus, without reference to the context, we cannot clearly say which information structure meaning the subject carries when the sentence is in V2 order. That is, the subject *Der lerər* in (15a) may or may not be associated with focus. Another example can be found in Breton (a V2 language). In the Q/A pair, what is focused in (16A) is the fronted item *Marí* (the rheme and the new information in Press (1986)'s terminology). In this case, the word order of the sentence is SVO.

(16) Q: Pív a wel Yanníg?
who sees Yannig
'Who sees Yannig?' [bre]

A: Marí a wel Yanníg
Marie sees Yannig
'Maries sees Yannig.' [bre] (Press, 1986, p. 194)

However, the sentence *Marí a wel Yanníg* itself, if it were not for the contextual information, sounds ambiguous. Press argues that in the sentence *Yanníg* could well be the subject of the sentence (i.e. in an OVS order). If *Yanníg* is the subject, focus is assigned to the fronted object *Marí*. In other words, a Breton sentence *Marí a wel Yanníg* conveys two potential meanings like either *It is Marie who sees Yannig.* (when the sentence is SVO) or *It is Marie who Yannig sees.* (when the sentence is OVS). Note that (16A) in which the focus is associated with the subject can be ambiguous given that Breton is a V2 language; that is, the subject, in itself, can be either interpreted as focus or just unknown. In my proposal presented later, the information structure value of the constituents *in situ* (e.g. the subjects in (15) and (16A)) has to remain underspecified.

Second, the unmarked focus positions in different languages also deeply interact with phonological variation. Ishihara (2001) argues that two types of stresses have an effect on the unmarked position; one is N-stress (Nuclear stress), and the other is A-stress (Additional stress). According

to Ishihara, the A-stress is not required, while every sentence presumably bears the N-stress, and the position of the N-stress is rather fixed in a language.[6] Thus, the N-stress is realized in the same position almost invariably even if some constituents move forward or backward (e.g. inversion, scrambling, etc.). For example, the following sentences in Japanese (17) and Ondarroa Basque (18), in which ´ and ˆ stand for the N-stress in each language, show the position of N-stress (i.e. preverbal in both languages) does not shift to reflect the change in word order.

(17) a. Taro-ga kyoo hón-o katta
Taro-NOM today book-ACC bought
'Taro bought a book today.' [jpn]

b. Taro-ga hon-o kyóo katta
Taro-NOM book-ACC today bought [jpn] (Ishihara, 2001, p. 145)

(18) a. Jonek Mîren ikusí ban.
John.ERG Miren see.TU AUX.PST
'Jon saw MIREN. [eus]

b. Miren Jônek ikusí ban.
Miren.ERG John see.TU AUX.PST
'JON saw Miren.' [eus] (Ibid., originally taken from Arregi (2000, p. 22))

N-stress has a tendency to fall on the preverbal position in OV languages as shown in *hón-o* and *kyóo* (17) and *Mîren* and *Jônek* in (18), while it tends to fall on the postverbal position in VO languages (e.g. English). By contrast, since A-stress lays an additional emphasis on a specific word, its position can vary depending on what the speaker wants to emphasize (i.e. focus). With respect to presence of the A-stress, Ishihara proposed a rule: Any material that follows an A-stress must be deaccented.

Combining the three factors presented thus far together, (i) basic word order, (ii) N and A-stresses, and (iii) the unmarked position for narrow focus, we can come up with the reason why an object normally bears the focus of a sentence in an unmarked way at least in the languages presented so far. The A-stress, as mentioned, does not show up unless it is necessary for the speaker

[6] In fact, Ishihara (2001) offers this argument based on a lot of previous phonological studies, but not seeing a large number of languages (e.g. Japanese, Korean, Basque, etc.). Thus, we may not say that these rules are meant to be universals. Nonetheless, Ishihara's argument still has a significance in that it is well discussed how different types of sentential stresses impact forming information structure of sentences in a default word order.

to emphasize something. In the absence of the A-stress, the word with N-stress is the most stressed constituent in the sentence. The N-stress has a strong tendency to fall on the object in both OV and VO languages. In addition, subjects have a strong tendency to be topics. Most languages have a part of the syntactic structure which is the unmarked position for topics, and subjects tend to fall in that part of the syntactic structure (Lambrecht, 1996). Hence, the unmarked marking of focus tends to fall on objects.

This dissertation does not deal with the unmarked position of topic and focus. We cannot identify them in an overt manner without a deterministic clue that reveals the information structure meanings. The different positions of focus in the next section are not in the most neutral word order in each language.

4.4.1 Focus Position

Some languages assign a specific position to signal the meaning of focus. It is noticeable that the position in this case is exclusively motivated by the necessity to mark narrow focus on a single constituent in the non-neutral word order. For example, if a language employs SVO by default, and the canonical focus position of the language is clause-final, then the object in SVO is not considered as necessarily containing focus. This is because sentences in the default word order allow for all possibilities in information structure.

According to my survey (and also Féry and Krifka (2008)), there are four positions that human languages employ to designate narrow focus; (i) clause-initial, (ii) clause-final, (iii) preverbal, and (iv) postverbal. In the following subsections, each position is exemplified and the languages that use the strategy are enumerated.

Clause-initial Position

First, narrow focus can be assigned to the clause-initial position in some languages, including English (e.g. focus/topic fronting constructions (Prince, 1984)), Ingush (Nichols, 2011), Akan (Drubig, 2003), Breton (Press, 1986), Yiddish (Jacobs, 2005), and Hausa (Hartmann and Zimmermann, 2007; Büring, 2010).

The representative example can be found in Ingush.[7] Ingush is a head-final language except for predominantly V2 order in main clauses (Nichols, 2011). In (19), the first element in each sentence is associated with focus.

(19) a. Cuo diicar suona jerazh.
3s.ERG D.tell.WP 1s.DAT these
'*She* told me them (=stories) to me.' (focus on *she*) [inh]

b. Suona diicar cuo yzh.
1s.DAT D.tell.WP 3s.ERG 3p
'She told *me* them (=stories) to me.' (focus on *me*) [inh] (Nichols, 2011, p. 687)

Hausa is also known as using the initial position for marking focus (Büring, 2010). As exemplified in a Q/A pair presented in (20Q-A1), the constituent replying to the *wh*-question appears first in Hausa. Focus in Hausa can also be realized *in situ* as shown in (20A2). That is to say, there are two types of foci in Hausa: namely *ex situ* focus (20A1) and *in situ* focus (20A2) (Hartmann and Zimmermann, 2007).

(20) Q: Mèe sukà kaamàa?
what 3PL.REL.PERF catch
'What did they catch?' [hau]

A1: **Kiifii** (nèe) sukà kaamàa.
fish PRT 3PL.REL.PERF catch
'They caught FISH.' [hau]

A2: Sun kaamàa **kiifii**.
3PL.ABS.PERF catch fish
'They caught FISH.' [hau] (Hartmann and Zimmermann, 2007, p. 242–243)

There are then two types of languages with respect to focus position. One obligatorily places focused elements in a specific position, and the other optionally does. Hausa is the latter type, while Ingush seems to fall into the former type.

English belongs to the second type of languages, in which positioning focus in an overt way is not mandatory. Even if a language does not always assign focus to the clause-initial position, the

[7] a Northeast Caucasian language, spoken in Ingushetia and Chechnya

language can sometimes make use of clause-initial focus, which is called focus/topic-fronting in this dissertation.[8] In human languages, old information is sometimes focus-marked. In (21), the replier wants to say that *she* does not merely know *John*, but dislikes him.[9]

(21) Q: Does she know JOHN?
A: JOHN she DISLIKES. (Gussenhoven, 2007, p. 96)

Hence, an English sentence in which the object is not *in situ* (e.g., *John she dislikes.*), if we do not consider the accents, can be read ambiguously (e.g., either *It is John who she dislikes.* or *As for John, she dislikes him.*). These matters are revisited in the next chapter in terms of discrepancies between meaning and marking of information structure. For the moment, suffice it to say that the clause-initial position can be employed to narrowly mark the focus of the sentence in many languages including English.

Clause-final Position

Second, narrow focus can be licensed in clause-final position in some languages. These include Russian (Neeleman and Titov, 2009),[10] Bosnian Croatian Serbian (Bojan Belić, p.c.), American Sign Language (Petronio, 1993; Churng, 2007), and some Chadic languages such as Tangale and Ngizim (Drubig, 2003). For example, in Russian, if (i) a constituent corresponds to the *wh*-word in a given question, and thereby is narrowly focused and (ii) the accent does not designate the focus, it can occupy the clause-final position as presented below.[11]

(22) Q: Kto dal Kate knigu?
who gave Kate.DAT book.ACC
'Who gave a book to Kate?'

A: Kate knigu dala ANJA.
Kate.DAT book.ACC gave Anna
'ANNA gave a book to Kate.' (focus on the subject)

---

[8] As mentioned several times, this kind of syntactic operation is often called topicalization (Prince, 1984; Man, 2007).

[9] Another example is already given in (4), which is called Second Occurrence Focus (§4.2.2).

[10] In Russian, non-contrastive focus (i.e. semantic-focus in the taxonomy of this dissertation) shows up sentence-finally, whereas contrastive focus is fronted (Neeleman and Titov, 2009).

[11] The second answer in (22) is in the most unmarked word order in Russian.

Q: Čto Anja dala Kate?
what.ACC Anna gave Kate.DAT
'What did Anna give to Kate?'

A: Anja dala Kate KNIGU.
Anna gave Kate.DAT book.ACC
'Anna gave a BOOK to Kate.' (focus on the direct object)

Q: Komu Anja dala knigu?
who.DAT Anna gave book.ACC
'Who did Anna give a book to?'

A: Anja dala knigu KATE.
Anna gave book.ACC Kate.DAT
'Anna gave a book to KATE.' (focus on the indirect object) [rus] (Neeleman and Titov, 2009, p. 515)

Russian, in which the most unmarked word order is SVO, is known for its free word order of constituents. However, Rodionova (2001), exploring variability of word order in Russian declarative sentences, concludes that the word order in Russian is influenced by different types of focus: namely narrow, predicate, and sentential focus.

The same phenomenon holds in Bosnian Croatian Serbian as exemplified in (23); compared to (23a) in an unmarked word order in the language (SVO), the subject in (23b) *Slavko* is postposed to mark focus meaning overtly through syntax.[12]

(23) a. Slavk-o vid-i Olg-u
Slavko.M-SG.NOM see-3.SG Olg-3.F.SG.ACC
'Slavko sees OLGA' (the unmarked word order) [hbs]

b. Olg-u vid-i Slavk-o
Olga.F-SG.ACC see-3.SG Slavko.M-SG.NOM
'SLAVKO sees Olga.' (focus on the subject) [hbs]

Preverbal Position

Third, the (immediately) preverbal position is one of the sites that signal focus. Languages that assign narrow focus to the preverbal position include Basque (Ortiz de Urbina, 1999), Hungarian

[12](23) was examined by Bojan Belić (p.c.).

(É. Kiss, 1998; Szendrői, 2001), Turkish (İşsever, 2003), and Armenian (Comrie, 1984; Tamrazian, 1991, 1994; Tragut, 2009; Megerdoomian, 2011). Basque, for instance, is a language in which focus marking heavily depends on sentence positioning. This is similar to Catalan (Vallduví, 1992; Engdahl and Vallduví, 1996) and Yucatec Maya (Kügler et al., 2007). The syntactic device for marking narrow focus in Basque is to assign focus immediately to the left of the verb as exemplified in (24). While (24a) conveys neutral information structure (i.e., all constituents are underspecified from the view of this dissertation.), in (24b-c), the subject *Jonek* 'Jon', being adjacent to the verb *irakurri* 'read', should be read as conveying focus meaning.

(24) a. Jonek eskutitza irakurri du
Jon letter read has
'Jon has read the letter.' (SOV) [eus]

b. Jonek irakurri du eskutitza
Jon read has letter
'JON has read the letter.' (SVO) [eus]

c. Eskutitza, Jonek irakurri du
letter Jon read has
'JON has read the letter.' (OSV) [eus] (Ortiz de Urbina, 1999, p. 312)

Crowgey and Bender (2011) also employs the *wh*-test for identifying focus in Basque: Both (25b-c) are grammatical sentences in Basque, but (25c) cannot be used as an answer to (25a). This distinction in felicity-conditions shows that focused constituents should appear in the immediately preverbal position.

(25) a. Liburu bat nork irakurri du?
book one.ABS.SG who.ERG.SG.FOC read.PERF 3SGO.PRES.3SGA
'Who has read one book?' [eus]

b. Liburu bat Mirenek irakurri du.
book one.ABS.SG Mary.ERG.SG.FOC read.PERF 3SGO.PRES.3SGA
'Mary has read one book.' [eus]

c. Mirenek liburu bat irakurri du.
Mary.ERG.SG.FOC book one.ABS.SG read.PERF 3SGO.PRES.3SGA
'Mary has read one book.' [eus] (Crowgey and Bender, 2011, p. 48–49)

Hungarian is one of the most well-known languages for the fixed focus position. The constituent order in Hungarian can be schematized as '(Topic*) Focus V S O' (Büring, 2010), as exemplified in (26).

(26) a. Mari fel hívta Pétert.
Mary-NOM VM rang Peter-ACC [hun]

b. MARI$_F$ hívta fel Pétert.
Mary-NOM rang VM Peter-ACC [hun]

c. *MARI fel hívta Pétert.
Mary-NOM VM rang Peter-ACC
'Mary rang up Peter' [hun] (Szendrői, 1999, p. 549)

(26a) is encoded as the basic word order, in which a marker *fel* occurs between the subject *Mari* 'Mary' and the main verb *hívta* 'rang'. If *Mari* is focused, the verb *hívta* should immediately follow the focused item as given in (26b), and if not, it sounds bad as given in (26c). É. Kiss (1998) states that focus in Hungarian can appear either *in situ* or immediately preverbally.[13] Szendrői (2001) argues that the focus in Hungarian tends not to be *in situ*, and that preverbal positioning has to be phonologically licensed as marked with small caps above.

According to Tamrazian (1991), Armenian also places focused constituents in the immediately preverbal position: Both sentences (27a-b) sound natural in Armenian, but the first one is in the basic word order without a focused element. In contrast, the preverbal item SURKIN in (27b) is focused, which is signaled by the adjacent auxiliary *e*. The auxiliary *e* should immediately follow the focused item. For instance, (27c) in which an accent falls on SURKIN but *e* appears after the main verb *sirum* 'like' is ill-formed.

(27) a. siranə surikin sirum e
Siran(NOM) Surik(ACC) like is
'Siran likes Surik' [hye]

b. siranə SURIKIN e sirum
Siran(NOM) Surik(ACC) is like
'Siran likes SURIK' [hye]

[13]That indicates informational focus and identificational focus, respectively. According to É. Kiss (1998), the preverbal focus in Hungarian (i.e. identificational focus) is almost the same as cleft constructions in English.

c. *siranə SURIKIN sirum e
Siran(NOM) Surik(ACC) like is [hey] (Tamrazian, 1991, p. 103)

Postverbal Position

Finally, the (immediate) postverbal position is responsible for marking narrow focus in several languages. These include Portuguese (Ambar, 1999), Toba Batak, and Chicheŵa (van Valin, 2005). In one example of this, Ambar claims that the non-contrastive focus is preceded by the verb in Portuguese. An example is presented below, in which the focused item *a Joana* (functioning as the subject) follows the verb *comeu* 'ate'. If the subject with focus meaning precedes the verb, the sentence sounds infelicitous in the context, as shown in (28A3-A4).

(28) Q: Quem comeu a tarte?
who ate the pie
'Who ate the pie?' [por]

A1: Comeu a Joana.
Ate the Joana

A2: A tarte comeu a Joana.

A3: #A Joana comeu.

A4: #A Joana comeu a tarte. [por] (Ambar, 1999, p. 27)

4.4.2 Topic Position

Topic is also associated with a specific position in some languages. For example, according to Ambar (1999), topics in Portuguese cannot follow the verb as shown in (29).

(29) Q: Que comeu a Maria?
what ate the Mary
'What did Mary eat?' [por]

A1: Comeu a tarte.

A2: A Maria Comeu a tarte.

A3: #A tarte Comeu a Maria.

A4: #Comeu a Maria a tarte. [por] (Ambar, 1999, p. 28)

In (29), *Maria* 'Mary' plays the topic role in the answers. The word should either disappear (as shown in (29A1)) or precede the verb *comeu* 'ate' (as presented in (29A2)). The sentences in which the topic preceded by the verb sound infelicitous (as provided in (29A3-A4)).

Topic-first Restriction

The canonical position of topic has been assumed to be sentence-initial in some previous studies. In fact, quite a few languages have been reported to have a strong tendency to use topic-fronting. Nagaya (2007) claims that topics in Tagalog canonically appear sentence-initially, Chapman (1981) says topics in Paumarí appear sentence-initially, and Casielles-Suárez (2003) states that topics should be followed by the focus (i.e. *topic-focus*) in the canonical word order in Spanish. In Bosnian Croatian Serbian, if a constituent such as *Olg-u* is given in the previous sentence as the focus as shown in (30a), it appears sentence-initially in the following sentence such as (30b) when functioning as the topic. Since focused constituents in that language appear in the clause-final position (as mentioned earlier (§4.4.1)), *mi* 'we' in (30b) is associated with focus as marked in small caps in the translation.[14] That is, in Bosnian Croatian Serbian, topics appear first, and foci occur finally.

(30) a. Slavk-o vid-i Olg-u
Slavko.M-SG.NOM see-3.SG Olg-3.F.SG.ACC
'Slavko sees OLGA' [hbs]

b. Olg-u vid-imo i mi
Olg.F-3 SG.ACC as well 1.PL.NOM
'WE see Olga, too' [hbs]

Other studies, however, indicate that topics are not necessarily sentence-initial (Erteschik-Shir, 2007; Féry and Krifka, 2008). According to Erteschik-Shir's analysis, topic-fronting is optional in Danish, and topics can be marked either in an overt way (i.e. topicalization in (31a)) or *in situ* as shown in (31b).

(31) a. Hun hilstepå Ole. **Ham** havde hun ikke mødt før...
She greeted Ole. **Him** had she not met before

[14](30) was also checked out by Bojan Belić (p.c.). He said *i* 'as well' in (30b) enforces the focus effect on *mi* 'we' in the final position. That means *i* in the sentence behaves as a focus particle, similarly to 'also' in English.

b. Hun hilstepå Ole. Hun havde ikke mødt **ham** før...
She greeted Ole. She had not met **him** before (Erteschik-Shir, 2007, p. 7)

Erteschik-Shir asserts that so-called topicalization in Danish, which dislocates the constituent playing the topic role to the left periphery, is used only for expressing the topic in an overt way. In other words, topics in Danish are not necessarily sentence-initial.

However, in some languages including Japanese and Korean, it is the case that (non-contrastive) topics are required to be sentence-initial (Maki et al., 1999; Vermeulen, 2009). Maki et al. argue that a *wa*-marked phrase can be interpreted as a topic if and only if it turns up in initial position. Otherwise, the *wa*-marked phrase in a clause-internal position should be evaluated as conveying a contrastive meaning.

(32) a. John-wa kono hon-o yonda.
John-WA this book-ACC read
'As for John, he read this book.' [jpn]

b. Kono hon-wa John-ga yonda.
this book-WA John-NOM read
'As for this book, John read it.' [jpn]

c. John-ga kono hon-wa yonda.
John-NOM this book-WA read
'John read this book, as opposed to some other book.'
'*As for this book, he read this it.' [jpn] (Maki et al., 1999, p. 7–8)

The same goes for Korean in my intuition. Féry and Krifka (2008) provide a *prima facie* counterexample to this claim as shown in (33), in which *disethu* 'dessert' is combined with (*n*)*un*.

(33) nwukwuna-ka disethu-nun aiswu khwulim-ul mek-ess-ta.
everyone-NOM dessert-NUN ice.cream-ACC eat-PST-DECL
'As for dessert, everyone ate ice cream.' [kor] (Féry and Krifka, 2008, p. 130)

However, (*n*)*un* is not always compatible with the information structure meaning of topic; that is, there is a mismatch between form and meaning. The (*n*)*un*-marked *disethu* in (33), in my intuition, surely fills the role of contrastive topic, rather than aboutness topic. Contrastive topics cross-linguistically have no constraint on position in word order (Erteschik-Shir, 2007; Roberts, 2011). In conclusion, aboutness topics in Korean and Japanese should be sentence-initial.

Building on the analyses presented so far, this dissertation argues that the canonical position of aboutness topics is language-specific: In some languages such as Japanese and Korean aboutness topics must appear in the initial position, while in other languages such as Danish they do not.

Right Dislocation

It is necessary to take one more non-canonical topic position into account. Topics can also appear sentence-finally. This phenomenon is called right dislocation (Cecchetto, 1999; Law, 2003), sentence-final topic (Féry and Krifka, 2008), anti-topic (Chafe, 1976; Lambrecht, 1996), or post-posing (Kim, 2011b).

(34) a. Left dislocation: This book, it has the recipe in it.

b. Right dislocation: You should go to see it, that movie. (Heycock, 2007, p. 185–186)

Gundel (1988) regards this construction as a peculiar construction within the comment-topic structure as opposed to the ordinary topic-comment structure. There must be an intonational break (i.e. a prosodic phrase marked as $_{p}$) which separates the topic from the prior parts of the given sentence. Such constructions exist cross-linguistically as exemplified in (35[15]-36[16]).

(35) a. kumyen, kuke-n com saki-nte.
if.so that-NUN a.little fraud-be.SEM
'If so, that is a kind of fraud, I think.' [kor]

b. kumyen, com saki-nte, kuke-n
if.so a.little fraud-be.SEM that-NUN
'If so, that is a kind of fraud, I think.' [kor] (Kim, 2011b, p. 223–224)

(36) a. ((Go loupo)$_{P}$ (nei gin-gwo gaa)$_{P}$, ([ni go namjan ge]$_{T}$)$_{P}$)$_{I}$.
CLF wife 2.SG see-EXP DSP this CLF man DSP
'The wife you have seen, of this man.' [yue]

[15]The topic marker *n* in (35) is an allomorph of (*n*)*un*, which mostly shows up in spoken data.

[16]Féry and Krifka (2008) state a boundary tone that created by the lexical markers responsible for information structure meanings (e.g. *ge* in Cantonese as given in (36a)) allows the topic to be added into the final position.

b. ((Pierre l' a mangée)$_{P}$, ([la pomme]$_{T}$)$_{P}$)$_{I}$.
Peter it-ACC has eaten, the apple
'Peter has eaten the apple.' [fra] (Féry and Krifka, 2008, p. 130)

Despite the difference in positioning, right dislocation has much in common with left dislocation. At first appearance, right dislocation looks like a mirror image of left-dislocation, in that the topic is apparently separate from the main clause and it is not likely that there is a missing function in the preceding sentence. In fact, Cecchetto (1999) proposes the so-called mirror hypothesis, which implies right dislocation is actually tantamount to a mirror image of left-dislocation.

This dissertation hence regards right dislocation as a non-canonical variant of left dislocation. Lambrecht (1996) provides a counterargument to this hypothesis, but the difference between left/right dislocations in Lambrecht's analysis looks contextual, rather than the result of a morphosyntactic operation. As this dissertation is not directly concerned with pragmatic constraints, the mirror hypothesis is still applicable to the current work. The difference between them seems to be trivially influenced by the degree of speaker's attention to the conversation: Left-dislocation would be used for the purpose of restricting the frame of what the speaker wants to talk about in advance, whereas right dislocation is just an afterthought performing almost the same function. A piece of evidence that supports this argument is provided by a corpus study which exploits a monolingual but fully naturally occurring text. Kim (2011b) scrutinizes several spoken corpora in Korean, and concludes that right dislocation (postposing, in his terminology) such as (35b) is largely conditioned by how accessible and/or urgent the information is: If the information is not uttered within several neighboring preceding sentences and is thereby less accessible in the speaker's consciousness, it tends to be easily postposed. That implies the choice between left and right dislocation is affected by only contextual conditions.

4.4.3 Contrast Position

Contrastive topics have a weaker constraint on order than non-contrastive topics (i.e. aboutness topics) (Erteschik-Shir, 2007; Bianchi and Frascarelli, 2010). Contrastive topics have a tendency to precede aboutness topics in some languages (Bianchi and Frascarelli, 2010), but this generalization has not been verified in all languages. Second, regarding sentence positioning of contrastive focus, there are two types of languages. The first, in which contrastive focus share the same po-

sition with non-contrastive focus, is more common. The typical language for this type is English, in which contrastive focus is not distinguishable from non-contrastive focus in terms of sentence position. The second type of language has two distinctive positions among the ordinary focus positions given earlier; (i) clause-initial, (ii) clause-final, (iii) preverbal, and (iv) postverbal. The languages that belong to this type, as presented before, include Georgian (preverbal *vs*. postverbal, (Skopeteas and Fanselow, 2010)), Portuguese (preverbal *vs*. postverbal, (Ambar, 1999)), Russian (clause-initial *vs*. clause-final, (Neeleman and Titov, 2009)), Ingush (immediately preverbal *vs*. clause-initial, (Nichols, 2011)), and so on. For example, (37) shows preverbal focus and postverbal focus in Georgian.

(37) a. kal-i kotan-s u-q'ur-eb-s.
woman-NOM pot-DAT (IO.3)OV-look.at-THM-PRS.S.3.SG [kat]

b. kal-i u-q'ur-eb-s kotan-s.
woman-NOM (IO.3)OV-look.at-THM-PRS.S.3.SG pot-DAT
'The woman looks at the pot.' [kat] (Skopeteas and Fanselow, 2010, p. 1371)

According to Skopeteas and Fanselow, both sentences in (37) are legitimate in Georgian. The difference between them is where the narrowly focused item appears in a sentence; either in the immediately preverbal position or in a postverbal position. That is, *kotan-s* in (37a) is a preverbal focus, while the subject *kal-i* and the object *kotan-s* in (37b) can be interpreted as focus (i.e. preverbal focus and postverbal focus, respectively). Skopeteas and Fanselow draw a conclusion that focus in the preverbal position normally bears contrastiveness (i.e. contrastive focus). Thus, the positions that non-contrastive focus and contrastive focus canonically occupy are different in Georgian.

This distinction between two types of foci requires the grammar library for information structure to provide an option: (a) whether the language uses the same position for both, and (b) if not, which component occupies which position.

On the other hand, one language can have two (or more) types of forms of expressing contrastive meaning, and this has to be considered in developing a grammar library for information structure. For example, Ingush marks contrastive focus by two means; one uses a clitic *=m*, and the other is expressed via word order, as exemplified in (38a-b) respectively.

(38) a. Suona=m xoza di xet, hwuona myshta dy xaac (suona).
1s.DAT=FOC nice day think, 2s.DAT how D.be.PRS know.PRS (1s.DAT)
'I don't know what you think, but *I* think it's a nice day.' [inh] (Nichols, 2011, p. 721)

b. Pacchahw **uqazahw** hwa-voagha
king here DX.V.come.PRS
'The king is coming *here* (he was expected to go somewhere else).' [inh] (Nichols, 2011, p. 690)

According to Nichols, using a clitic as given in (38a) is motivated by the necessity to express contrastive meaning in a more marked way. The ordinary contrastive focus, as shown in (38b) where focus is in boldface, occupies the immediate preverbal position in Ingush, and this position is different from the non-contrastive focus position, which occurs clause-initially.

In sum, the canonical position for contrastive focus is language specific; contrastive focus can either share the same position with non-contrastive focus (e.g. English, Greek (Gryllia, 2009), etc.) or show up in another position (e.g. Portuguese (Ambar, 1999), Russian (Neeleman and Titov, 2009), Georgian (Skopeteas and Fanselow, 2010), Ingush (Nichols, 2011), etc.). Contrastive topics, which can turn up anywhere (Erteschik-Shir, 2007; Bianchi and Frascarelli, 2010), do not have such a rigid restriction on position.

4.5 Summary

There are three linguistic forms of expressing information structure: (i) prosody, (ii) lexical markers, and (iii) sentence positioning. Prosody is a widespread way, but some languages do not employ prosody for marking focus and topic. The best way to handle prosodic marking in the current work is to allow for underspecification in such a way that prosodic information can be added into formalism. Lexical markers can be affixes, adpositions, and modifiers. The last two can be commonly categorized as clitics, but they are differentially dealt with in the current work. Information-structure marking adpositions are in complementary distribution with ordinary case-marking adpositions in a language. As for sentence positioning, my basic argument is that information structure of sentences in the basic word order has to be underspecified. When a constituent is *ex situ* and narrowly focused, four positions can be used: clause-initial, clause-final, preverbal, and postverbal. Topics canonically appear sentence-initially in some languages, but the topic-first restriction is not necessarily applied to all languages (i.e. language-specific). Contrastive focus may or may not share the same posi-

tion as non-contrastive focus (i.e. semantic focus). Lastly, contrastive topic does not enforce strong constraints on position across languages.

Chapter 5

DISCREPANCIES BETWEEN MEANING AND MARKING

Bolinger (1977) claims that one meaning per one form and vice versa (i.e. an isomorphism between formal and interpretive domains) is the most natural status of human languages. Natural human languages, however, provide many counterexamples to this notion. To take a typical example, homonymy and polysemy are examples of a single forms ability to convey two or more meanings. Moreover, mismatches between meaning and form can sometimes be caused by grammatical elements. For example, English shows discrepancies between form and meaning in counterfactuals, which refer to constructions in which the speaker does not believe the given proposition expressed in the antecedent is true. The most well-known factor which deeply contributes to the counterfactual meaning in many languages is the past tense morpheme, such as '-ed' in English (Iatridou, 2000). The past tense morpheme in counterfactuals (a.k.a. fake past tense) does not denote an event that actually happened in the past as exemplified in (1). Thus, the mapping relationship between morphological forms and their meaning in counterfactual sentences is not the same as that in non-counterfactual sentences.

(1) a. If he were smart, he would be rich.
(conveying "He isn't smart." and "He isn't rich.")

b. I wish I had a car.
(conveying "I don't have a car now." (Iatridou, 2000, p. 231–232))

As with other grammatical phenomena, information structure also exhibits some discrepancies in form-meaning mapping. This chapter presents several types of mismatches between the forms that express information structure and the information structure meanings conveyed by those forms.

5.1 Ambivalent Lexical Markers

In some languages, one lexical marker can correspond to meanings of several components of information structure (i.e. no one-to-one correspondence between form and meaning). A mismatch

caused by lexical markers shows up in Japanese and Korean. As is well-known, *wa* in Japanese and (*n*)*un* in Korean are regarded as lexical markers to express topic, but they can also sometimes be used for conveying contrastive focus.

(2) Q: Kim-i onul o-ass-ni?
Kim-NOM today come-PST-INT
'Did Kim come today?' [kor]

A: ani. (Kim-un) ecey-nun o-ass-e.
No. Kim-NUN yesterday-NUN come-PST-DECL
'No. Kim came yesterday.' [kor]

The lexical marker (*n*)*un* in Korean appears twice in (2A); one occurance is with the subject *Kim*, and the other is combined with an adverb *ecey* 'yesterday'. Although the same lexical marker is used, they do not share the same properties of information structure. It is clear that topic is assigned to *Kim-un* in that the word is already given in the question and it is optional as indicated by the parentheses. By contrast, the (*n*)*un*-marked *ecey* is newly and importantly mentioned by the replier, and thereby it should be evaluated as containing a meaning of focus rather than topic. Moreover, if *ecey-nun* disappears, the answer sounds infelicitous within the context, which clearly implies it is focused. Recall that I define focus as an information structure component associated with an inomissible constituent. Furthermore, (2) passes the correction test to vet contrastive focus (Gryllia, 2009). Since *onul* 'today' in the question and *ecey* 'yesterday' in the reply constitute an alternative set, *ecey* in (2A) simultaneously has a contrastive meaning. As a consequence, the information structure role of *ecey* in (2A) is contrastive focus, even though the so-called topic marker (*n*)*un* is attached to it.

This (*n*)*un*-marked constituent associated with contrastive focus is realized differently from that with contrastive topic. In (3A), the (*n*)*un*-marked element in the first position can be dropped as the parentheses imply. Comparing two sentences (with or without *ku chack-un*), when it overtly appears, the sentence signals contrast to the fronted constituent. This has something in common with Choi (1999)'s argument. She claims that only elements with contrastive meaning can be scrambled in Korean, which means *ku chack-un* 'the book-NUN' in (3A) gives contrastive meaning.

(3) Q: nwuka ku chayk-ul ilk-ess-ni?
who the book-ACC read-PST-INT
'Who read the book?' [kor]

A: (ku chayk-un) Kim-i ilk-ess-e.
the book-NUN Kim-NOM read-PST-DECL
'(As for the book,) Kim read it.' [kor]

In fact, Choi does not concede existence of contrastive topic in Korean, and the scrambled and (*n*)*un*-marked constituents are analyzed as only contrastive focus in her proposal. However, this notion is contradictory to the definition that focus cannot be elided. Given that *ku chayk-un* in (3A) can felicitously disappear, we cannot say that it is associated with focus. Since contrast should be realized as either contrastive focus or contrastive topic, *ku chayk-un* in (3A) must be evaluated as a contrastive topic.

In short, (*n*)*un* in Korean can assign three meanings to the adjoining NP: aboutness topic, contrastive topic, and contrastive focus. In other words, (*n*)*un* provides constraints, but only partial ones, which cause discrepancies between form and meaning. Because this marker can be combined with constituents that are not topics, it is my position that 'topic-marker' is not an appropriate label. The same goes for *wa* in Japanese (Song and Bender, 2011). Case markers in these languages (e.g. *i/ka* and *ga* for nominatives) also convey an ambiguous interpretation, such as either focus or background (i.e. non-topic).

In some languages, a lexical marker known for marking topic coincides with cleft constructions which clearly carry a focus meaning. (4) in Ilonggo[1] exemplifies such a mismatch (Schachter, 1973). In Ilonggo, the topic marker *ang* is in complementary distribution with case markers similarly to *wa* in Japanese and (*n*)*un* in Korean. One difference is that the case relation is marked by an affix attached to the verb (e.g. the agentive marker *nag-* in (4)).

(4) a. nag- dala ang babayi sang bata
AG.TOP- bring TOP woman NONTOP child
'The woman brought a child.' [hil]

[1]a.k.a. Hiligaynon, an Austronesian language spoken in the Philippines

b. ang babayi ang nag- dala sang bata
TOP woman TOP AG.TOP- bring NONTOP child
'It was the woman who brought a child.' [hil] (Croft, 2002, p. 108)

(4a) is a topicalized construction in which the topic marker *ang* is combined with *babayi* 'woman'. (4b) is a focused construction, in which the topic marker *ang* is still combined with the focused constituent *babayi*, and one more topic marker appears at the beginning of the cleft clause *nag- dala sang bata*. That implies that the so-called topic marker does not necessarily express topic meaning.

5.2 Focus/Topic Fronting

Using the methodology presented in Chapter 4 (§4.1), my cross-linguistic survey collects distributional information about focus/topic fronting and draws a tentative conclusion: If focus and topic compete for the sentence-initial position, topic always wins. To take an example, in Ingush both topic and focus can precede the rest of the clause, but a focused constituent must follow a constituent conveying topic, as exemplified in (5).

(5) Jurta jistie joaqqa sag ull cymogazh jolazh.
town.GEN nearby J.old person lie.PRS sick.CVsim J.PROG.CVsim
(topic) (focus)
'In the next town an old woman is sick (is lying sick).' [inh]

Mista xudar myshta duora?
sour porridge how D.make.IMPF
(topic) (focus)
'How did they make sour porridge? (How was sour porridge made?)' [inh] (Nichols, 2011, p. 683)

That means that if topic and (narrow) focus co-occur, topic should be followed by focus even in languages which place focused constituents in the clause-initial position. The same phenomenon can be found in many other languages. For example, in Nishnaabemwin,[2] if both the subject and the object of a transitive verb appear preverbally, the first is marked for topic and the second for focus (Valentine, 2001). No counterexamples to this generalization have been observed, at least among the languages that I have examined hitherto (§4.1).

[2] an Algic language (a.k.a. Odawa and Eastern Ojibwe [ojg/otw]), spoken in the region surrounding the Great Lakes, in Ontario, Minnesota, Wisconsin, and Michigan

Yet, there are some cases in which it is unclear which role (i.e. focus or topic) the fronted constituent is assigned to. I would like to call the constructions in which this kind of ambiguity takes place 'focus/topic-fronting' (a.k.a. Topicalization). Prince (1984) provides two types of OSV constructions in English, and argues that the change in word order is motivated by marking information status, such as new and old information.

(6) a. John saw Mary yesterday.

b. Mary, John saw yesterday.

c. Mary, John saw her yesterday. (Prince, 1984, p. 213)

Both (6b-c) relate to (6a), but (6b) is devoid of the resumptive pronoun in the main clause, whereas (6c) has *her* referring to *Mary*. These are called Topicalization and Left-Dislocation by Prince,[3] but I use the label focus/topic-fronting for the first type of syntactic operation.

The focus/topic fronting constructions have two potential meanings, as exemplified in (7). That is, (7a) can be paraphrased into either (7b) or (7c), whose information structures differ.

(7) a. The book Kim read.
b. It was the book that Kim read.
c. As for the book, Kim read it.

If the fronted NP is focused, its configuration is the same as cleft constructions (7b). If it behaves as the topic within the context, the sentence can share the same information structure as (7c). That means (7a) in itself would sound ambiguous, if it were not for any contextual information. Gundel (1983), in order to distinguish the different structures, makes use of two terms Focus Topicalization and Topic Topicalization, suggesting that OSV constructions like (7a) are ambiguous. Gussenhoven (2007) also takes notice of such an ambiguity, and regards the constructions like (7a) as containing 'reactivating focus'.

---

[3]Prince (1984) argues that the choice of one over another is not random but is influenced by the information status of what the speaker is talking about. According to Prince, Topicalization has two characteristics; one is that it is used to mark information status of the entity itself, and the other is that it involves an open proposition. In short, in Prince's analysis, information status factors (e.g. new *vs*. given), have an effect on the composition in the OSV order, which removes the fronted NP referring to a discourse-new entity from a syntactic position that disfavours it. As indicated in Chapter 3 (§3.1), since this dissertation is not concerned with information status, such a distinction based on new information *vs*. given information is not used in the present work.

(8) Q: Does she know JOHN?

A: JOHN she DISLIKES. (Gussenhoven, 2007, p. 96)

Nevertheless, it is my position that the names that Gundel and Gussenhoven make use of still lead to confusion.[4]

Other languages also have the focus/topic-fronting constructions. In the following Cantonese example, the fronted constituent *nei1 bun2 syu1* 'this book' can play the role of either focus or topic of the sentence, and the choice between the two readings hinges on the context.

(9) Nei1 bun2 syu1 ngo5 zung1ji3

DEF CLF book 1.SG like

(a) 'It is this book that I like.' or

(b) 'As for this book, I like it.' [yue] (Man, 2007, p. 16)

The same phenomenon can be observed in Nishnaabemwin. Information structure in Nishnaabemwin, whose basic word order is VOS, is also accomplished via syntactic means. If a verbal argument appears before the verb, then it is marked for information structure. Its meaning, just as the previous examples in English and Cantonese, becomes ambiguous if only one argument is preverbal (Valentine, 2001). In fact, this kind of ambiguity frequently happens in languages in which focus shows up clause-initially. Ingush, for instance, assigns non-contrastive focus to the clause-initial position as mentioned earlier (Nichols, 2011). Notably, topic also has a strong tendency to be preposed.

In brief, one form has two different information structure meanings; the construction often refered to as Topicalization (Prince, 1984) almost invariably and cross-linguistically sounds ambiguous unless the given context is ascertained. Regarding the selection of terminology, this dissertation calls such a construction focus/topic-fronting, because (i) this explicitly displays the ambiguous meaning, and (ii) the previous terminology (i.e. topicalization) confuses syntactic and pragmatic notions.

---

[4]From a different point of view, Dan Flickinger (p.c.) says that (7c) does not look like a proper paraphrasing of (7a). In his intuition, the fronted item *The book* conveys only focus meaning. If his thought holds true, the focus/topic fronting constructions are actually equivalent to cleft constructions, like a pair of (7a-b). In fact, other native speakers who read focus/topic constructions in other languages have similar thought. For instance, Ka Yee Lun (p.c.) says that (9) in Cantonese can convey both meanings, but the first reading is predominant. For now, I do not make a conclusion about which one is a sound interpretation, but what is important is that the name 'Topicalization' is not appropriate in any cases.

5.3 Competition between Prosody and Syntax

Some languages have a fixed focus position, and there are potentially three subclasses of the position. First, some languages have a system with very weak or no interaction between prosody and syntax with respect to focus. These include Catalan (Engdahl and Vallduví, 1996), Akan (Drubig, 2003), and Yucatec Maya (Kügler et al., 2007). In those languages, displacing constituents is the only way to identify focused elements. The second subclass assigns focus to a particular position. Constraints on this position necessarily correlate with phonological marking in the second type of languages. Hungarian belongs to this type, in which the focused and accented item appears immediately prior to the verb (É. Kiss, 1998; Szendrői, 2001). The third type, which occasionally brings about a mismatch between form and meaning, includes languages in which prosody and syntax compete in expressing focus. That is, in this type of language, either prosodic or syntactic structure can be used to mark focus, depending on the construction.

Büring (2010) calls the third type 'Mixed Languages' and creates the following generalization about them.

(10) MARKED WORD ORDER → UNMARKED PROSODY: Marked constituent order may only be used for focusing X if the resulting prosodic structure is less marked than that necessary to focus X in the unmarked constituent order. (Büring, 2010, p. 197)

Büring argues that mixed languages include Korean, Japanese, Finnish, German, European Portuguese, and especially most of the Slavic languages. According to my survey, Russian and Bosnian Croatian Serbian (i.e. the Slavic languages) clearly fall under this type: which either (i) employ a specific accent to signal focus or (ii) assign the focused constituent to the clause-final position. Those languages differ from the first type in that prosodic markings are productively used for conveying focus meaning, and also differ from the second type in that prosody can independently work without any syntactic positioning. For instance, the subject *sobaka* 'dog' in (11a) can have focus meaning if and only if it bears the accent for focus, which means (11a) is informatively ambiguous in the absence of information about accent. In contrast, (11b) where the subject is in the final position sounds unambiguous, and *sobaka* is evaluated as focused.

(11) a. Sobaka laet.
dog bark
'The dog bark.' [rus]

b. Laet sobaka.
bark dog
'The DOG bark.' [rus]

The distinction between (11a-b) is more clearly shown with the *wh*-test. If the question is *Who barks?* as given in (12Q1), both sentences can be used as the reply. If the reply is (11a) in the neutral word order, the verb *laet* bears an accent. In contrast, if the question is (12Q2) which requires the predicate to be focused, (11b) cannot be an appropriate answer and also there should be no sentential stress on the verb *laet*.[5]

(12)Q1: Kto laet?
who barks
'Who barks?' [rus]

A1: Sobaka laet. / Laet sobaka. [rus]

Q2: Čto delaet sobaka?
what doing dog
'What does the dog do?' [rus]

A2: Sobaka laet. / #Laet sobaka. [rus]

The same holds true for Bosnian Croatian Serbian.[6] When the question is given as (13Q2), the sentence in which the subject is not *in situ* sounds infelicitous, and the verb *laje* is not allowed to bear a sentential stress.

(13)Q1: Ko laje?
who barks
'Who barks?' [hbs]

A1: Pas laje. / Laje pas.
dog barks. / barks dog.
'The dog barks.' [hbs]

---

[5]Marina Oganyan and Varya Gracheva, p.c.

[6]Bojan Belić, p.c.

Q2: Šta(Što) radi pas?
what doing dog
'What does the dog do?' [hbs]

A2: Pas laje. / #Laje pas. [hbs]

In summary, in the third type of language, prosody takes priority over syntax in the neutral word order with respect to expressing focus (i.e., the prosodic marking wins). In contrast, when the sentence is not in the default word order, syntactic structure wins. Since sentences in an unmarked word order are normally ambiguous along these lines, focus position is not defined for sentences with unmarked word order, only for those with other word orders.

5.4 Multiple Positions of Focus

Even if a language employs a specific position for expressing focus, the focused constituent does not necessarily take that position, as exemplified in Russian in the previous section. That is, focus can be assigned to multiple positions. For instance, the focus in Russian may not be clause-final (as presented in (11)), if the accent falls on another constituent. In this case, clause-final focus does not seem to be the same as cleft constructions in Russian, and the accented constituent *in situ* is not also necessarily equivalent to informational focus. A more complex phenomenon with respect to syntactic operations on focus is exemplified by Greek (Gryllia, 2009). In Greek, whose basic word order is VSO or SVO, focus can be both preverbal and postverbal and there is no informative difference between them.

(14) Q: Thelis kafe i tsai?
want.2SG coffee.ACC or tea.ACC
'Would you like coffee or tea?' [ell]

A1: Thelo $[\text{kafe}]_{\text{C-Foc}}$.
want.1SG coffee.ACC
'I would like coffee.' [ell]

A2: $[\text{Kafe}]_{\text{C-Foc}}$ thelo.
coffee.ACC want.1SG
'Coffee I would like.' [ell] (Gryllia, 2009, p. 44)

The preverbal focus, such as *kafe* in (14A2), is not *in situ*, because verbs precede objects in the neutral word order in Greek. Yet, there is no evidence that preverbal focus plays the role of identification and this sentential form is informatively the same as cleft constructions in Greek. Gryllia, moreover, argues that the focus in both positions can receive the interpretation of contrastive focus as well as non-contrastive focus. That is, there are options for focus realization in Greek; (i) preverbal non-contrastive focus, (ii) preverbal contrastive focus,(iii) postverbal non-contrastive focus, and (iv) postverbal contrastive focus. The multiple focus positions in Greek demonstrate convincingly that forms which express information structure are not in a one-to-one relation with information structure components and thereby cannot unambiguously mark a specific information structure meaning.

Another important phenomenon related to focus positions can be found in Hausa. According to Hartmann and Zimmermann (2007), Hausa employs two strategies for marking focus. One is called *ex situ* focus, and the other is *in situ* focus. They are exemplified in (15A1-A2), respectively.

(15) Q: Mèe sukà kaamàa?
what 3PL.REL.PERF catch
'What did they catch?' [hau]

A1: **Kiifii** (nèe) sukà kaamàa.
fish PRT 3PL.REL.PERF catch
'They caught FISH.' [hau]

A2: Sun kaamàa **kiifii**.
3PL.ABS.PERF catch fish
'They caught FISH.' [hau] (Hartmann and Zimmermann, 2007, p. 242–243)

In situ focus in Hausa does not require any special marking, whereas *ex situ* focus in the first position is prosodically prominent. Moreover, Hausa employs two focus particles *nèe* and *cèe*, but they can co-occur with only *ex situ* focus as shown in (15A1).[7] For this reason, Büring (2010) regards Hausa as a language without a specific marking system for focus. This analysis of focus realization in Hausa implies that some languages can assign focus to a constituent *in situ* without the help of pitch accents.

[7]This is an intriguing phenomenon, because in other languages *in situ* foci in the unmarked word order normally require an additional constraint, such as pitch accents. In other words, as shown in the examples of the Slavic languages (presented in the previous section), it is common that focused constituents in the default position need to be accented if the language uses multiple strategies for marking focus or topic.

The examples presented in this section motivate flexible representation of information structure, particularly for sentences in unmarked word order. That is to say, in some circumstances, we cannot exactly say where focus is signaled.

5.5 Summary

Just as with other grammatical phenomena, there are discrepancies between forms and meanings with respect to information structure. This chapter has looked over several cases in which there are mismatches in mapping between information structure markings and meanings. First, lexical markers of expressing information structure occasionally cause such a mismatch. For example, *wa* and (*n*)*un* in Japanese and Korean respectively have been known as topic markers in these languages, but they can sometimes be used for expressing contrastive focus. Second, topic and focus appear sentence-initially in quite a few languages, but there are some cases in which we cannot decisively say whether the fronted item is associated with topic or focus. Such a construction has often been called 'Topicalization' in previous literature, but I would like to use different terminology in order to be more accurate; these constructions are examples of focus/topic-fronting. Third, if prosody and syntax compete for expressing information structure, prosody takes priority in most cases. Finally, many languages place a focused constituent in a specific position, but this placement is optional in some languages. The last two properties are related to expressing focus in sentences in default word order. Information structure in unmarked sentences is addressed in Part IV in detail and also taken into consideration in terms of implementation presented in Part V.

Part III

A CORPUS STUDY

This part offers a corpus-based analysis of information structure, exploiting a multilingual parallel text. As with other linguistic investigations, a deep analysis of information structure requires the creation of language resources in which linguistic features related to the phenomena in question are annotated in a fine-grained way. As discussed thus far, different languages use different phonological, morphological, and syntactic means to express information structure. Moreover, for many languages, including even well-known and widely-studied languages, the full range of possibilities for information structure markings remains unknown. Thus, the most comprehensive way of delving into cross-linguistic structuring of information is to analyze multilingual texts. Multilingual texts facilitate the creation of systematic methods for identifying foci and topics in monolingual texts, and investigating how strategies of information structure in different languages are related to each other from a multilingual perspective.

Chapter 6

A CORPUS STUDY: ANNOTATION

The multilingual text this corpus study exploits is *The Adventure of the Speckled Band* out of the *Sherlock Holmes* series. The text was examined in four languages: English [eng], Spanish [spa], Russian [rus], and Korean [kor]. The original text was written in English by Arthur Conan Doyle.

This corpus study is designed to accomplish the following purposes: First, the current work aims to produce corpus-based generalizations about information structure. These distributional findings might support the previous theoretical work on information structure, or they might show where the theory is wrong and/or not yet complete. Second, the current work aids in the development of computational models that incorporate information structure, enriching grammar libraries, machine translation systems, and so forth. Finally, the whole data set will be readily distributed in order for other researchers to utilize it, facilitating further development along this line of study.

This chapter is structured as follows: First, §6.1 briefly surveys previous corpus work on information structure. Second, §6.2 gives an explanation of the methodology of the current corpus study. Third, §6.3 provides the annotation schema used in handling the multilingual parallel texts. Next, §6.4 reports on an evaluation annotation reliability, and then §6.5 explains how this corpus study can contribute to language processing applications.

6.1 Previous Corpus Studies

There have been several corpus studies for information structure so far, but my corpus study is different from previous work in two ways. First, my work exploits a (a-i) multilingual and (a-ii) parallel text, in which a sentence in one language is aligned with the corresponding sentence in the other languages. Second, another advantage of the current annotation schema is its coverage of (b) dropped elements. Analyzing *pro*-drop (e.g. subject-drop, topic-drop) is crucial to the study of information structure, because pronominal correspondence tends to be influenced by topicality and focality in many human languages.

It is my belief that utilizing fully parallel texts allows us to learn how information structure is differently expressed in different languages in a more comprehensive way. There are several guidelines for annotating information structure in multilingual texts as well as monolingual texts (Calhoun et al., 2005; Dipper et al., 2007; Brunetti et al., 2011). However, these previous studies differ from the current work in that their multilingual corpora are sets of monolingual texts written in several languages, rather than parallel corpora. There have also been previous studies which conduct a parallel analysis of bitexts (e.g. Japanese to English (Komagata, 1999), Swedish to English (Johansson, 2001), and Norwegian to English (Bouma et al., 2010)). While these studies provide information about cross-linguistic differences, a truly robust understanding of distributional differences in information structure in different languages necessitates the use of multilingual texts rather than just bitexts.

The annotation schema used in this study is adapted from Dipper et al. (2007), but my version additionally deals with dropped elements. As discussed in Chapter 3 (§3.2.3), inomissibility is the most important criterion for distinguishing focus from other components of information structure. Whether or not an element is dropped is the most important factor for distinguishing focus from topic and background, because focus cannot be elided (Lambrecht, 1996; Erteschik-Shir, 2007). For this reason, a corpus study of information structure needs to run parallel with an analysis of dropped elements (Kaiser, 2009). Nevertheless, previous corpus studies on information structure have paid less attention to information structure phenomena that involve dropped elements. In order to provide a more systematic explanation of multilingual processing, it is necessary to include *pro*-drop in the annotation schema.

In spite of the two differences outlined above, the previous corpus studies still have significance for the current study. First, the annotation schema constructed by these studies is a theoretically sound and based on cross-linguistic findings (Calhoun et al., 2005; Dipper et al., 2007; Brunetti et al., 2011). As noted above, my annotation schema makes particular use of the guidelines developed by Dipper et al. Their schema deals with various layers, including phonology, morphology, syntax, semantics, and information structure itself. As is well-known, information structure often interacts with various linguistic phenomena. Thus, in order to draw a bigger picture of strategies of information structure, it is appropriate to see how information structure is related to other grammatical structures. Furthermore, their schema has already been applied to the annotation of less

commonly studied languages, extending beyond Indo-European languages to other human language families. Thus, I believe that this annotation schema has been substantially verified from a cross-linguistic stance.

There are several other corpus studies worthy of note. Komagata (1999) utilizes a parallel expository text in English and Japanese for exploring a binomial partition between theme and rheme and the discourse status that theme requires. The data is originally downloaded from an online journal (*The Physician and Sportsmedicine*). Similarly to the present study, Komagata's main purpose of using a parallel text is to support a grammatical theory-based model of information structure. Komagata's model is based on the CCG (Combinatory Categorial Grammar, Steedman 2001) framework, providing further evidence that exploiting parallel corpora is a productive method for modelling information structure across grammatical frameworks. Johansson (2001) and Bouma et al. (2010) also employ parallel texts, though their studies are exclusively concerned with cleft constructions. What is of significance is that one of the main purposes in their studies is to present a fine-grained methodology for establishing differences in information structure across languages. In other words, just with other corpus-based studies, designing methodology is as important as what findings are obtained for the study of information structure. In particular, Bouma et al. show how state-of-the-art NLP tools can be used for more comprehensive construction and data analysis. Their use of a phenomenon-specific, semi-automatic corpus annotation methodology was partially and analogously applied to the current work. Finally, although no parallel text is used, the work of Gracheva (2013) is very significant to the current work. She explores the Russian National Corpus (Grishina, 2006), and establishes data-based constraints on markers of contrast in Russian. Building upon the findings from the data exploitation, she provides a variant version of the type hierarchy that I present in this dissertation (Figure 9.1 (p. 174)). Her method for identifying a specific component of information structure in a running text is significant to the current work, because the text of this study needed to be annotated in the same way.

Finally, handling *pro*-drop is very crucial in multilingual machine translation. Although corpora are not substantially exploited, Mitkov et al. (1995) and Mitkov (1999) lay emphasis on handling *pro*-drop in multilingual processing. Mitkov et al. give an explanation of the difficulty of translating dropped anaphora. Pronouns in English-like languages sometimes have a strong tendency to be translated into non-anaphoric words in other languages such as Malay. They argue that the machine

translation system has to rely on multilingual anaphora resolution in order to provide more felicitous translations. In line with this claim, Mitkov proposes a model of multilingual anaphora resolution that covers English, Polish, and Arabic. The model takes empty pronouns into special consideration; zero anaphora are very common in Polish, and pronouns can be realized only as suffixes of verbs in Arabic. These are significant to the present work in that I create a model of information structure for better performance in multilingual machine translation.

6.2 Data Compilation

The current work focuses on a subset of *The Adventure of the Speckled Band*. The first 100 sentences in each language (400 sentences, in total) are annotated. While the corpus size is small, 400 sentences proved enough to achieve the goals of the current work: providing a fine-grained qualitative analysis of information structure that captures multilingual generalizations across languages, a substantially different task to that of calculating quantitative properties by means of statistical measures. In a similar study, Hasegawa and Koenig (2011) investigate two exclusive focus particles in Japanese; *shika* 'except' and *dake* 'only', in a corpus of 100 examples gathered from websites of Japanese newspapers. The current work, replicates their approach of using a small corpus of running texts. In this case, 100 translation sets across the four languages were examined.

Using *The Adventure of the Speckled Band* has several merits in the study of information structure. First, since it is a naturally occurring text, it is possible to identify how sentences are informatively structured with respect to the contextual information provided in the previous sentences. In many cases, a decontextualized single sentence contains insufficient information for revealing which information structure component is associated with which constituent in the sentence. Moreover, information status, such as new and given information, can be detected only within context by referring to preceding sentences. Thus, it is preferable for a corpus study focusing on information structure to use a running text. A second merit of the text is that, since the text is a detective novel, it naturally includes quite a few Q/A pairs. As surveyed previously, Q/A pairs are crucially useful for identifying focus. Third, since the novel has been translated into many languages, it would be relatively easy to include other languages in the future. In fact, the text has already been annotated in several languages by other DELPH-IN groups (Francis Bond and Tim Baldwin, p.c.). The current

study can make a contribution to the work. Fourth, there exists a previous study that makes use of *The Adventure of the Speckled Band* for the study of information structure from the viewpoint of comparative linguistics (von Prince, 2012). In the future, it would be good to compare the two data-based analyses. Last but not least, copyright infringement problems were avoided by choosing the source text from among texts published more than 70 years ago. Rather than using existing translations which not yet out of copyright, the translations into the other three languages were separately provided by three native speakers. These translations are also released into the public domain for comparative studies.

6.2.1 Languages

The languages in this corpus study show some major cross-linguistic differences in information structure. (i) English employs prosody such as A/B accents for expressing information structure and several syntactic operations such as clefts, focus/topic fronting, dislocation, etc. (ii) Spanish is a subject-drop language, in which subjects not associated with focus can be missing (conditioned by the rich morphological paradigm in verbal items). Plus, Spanish employs clitic left dislocation constructions which also play a role in configuring information structure (Bildhauer, 2007, among many others). (iii) Russian takes advantage of its relatively free word order for marking information structure (Rodionova, 2001; Neeleman and Titov, 2009), and also has several clitics to convey information structure meanings (Gracheva, 2013). Plus, Russian also has cleft constructions (King, 1995). (iv) Korean is a topic-drop language (Sohn, 2001) in which any argument can disappear. Korean has lexical markers in complementary distribution, such as *i*/*ka* for nominative, (*l*)*ul* for accusative, and (*n*)*un* for contrast or topic. The lexical markers can be omitted, a phenomenon sometimes referred to as case-ellipsis (Yatabe, 1999; Sato and Tam, 2012). The use of this marking system is closely linked to information structure. Korean also makes use of scrambling (i.e. OSV word order), which although regarded as a dummy operation in syntax and semantics, does play a role in information structure marking (Choi, 1999; Song and Bender, 2011). In sum, the four languages annotated in the current work encompass many of the diverse phenomena related to information structure.

6.2.2 Software

This corpus study made use of EXMARaLDA (`http://www.exmaralda.org`) as an annotation tool.[1] EXMARaLDA, implemented in JAVA, facilitates the annotation using tiers which serve to align units of analysis. Dealing with data sets in a XML format, this software allows annotation of linguistic features at various layers (varying from phonology to discourse), using multiple tiers consisting of cell(s) for each word or phrase. The basic unit of each XML file is a sentence. Each word (or minimal unit of analysis) has cells across various tiers in rows. A screenshot is presented in Figure 6.1. A sample format that shows an annotated sentence is presented in the next section.

6.2.3 Annotators

Seven annotators, including me as an annotator for Korean, participated in this task, all native or near native speakers of the language they worked on; two annotators for English and Spanish, one annotator for Spanish, two annotators for Russian, and two annotators for Korean. Every annotator was aware of basics in corpus-based studies, and had experience in the creation of language resources before their work on this project. There were three annotator tasks; (i) translation, (ii) annotation for generalization, and (iii) dual annotation for kappa testing.

Translation

(i-a) The rough translation for Spanish, Russian, and Korean was created by one of the annotators. (i-b) After that, other annotators who speak the language as the mother tongue proofread the translations for each language. The proofreading was required due to the fact that the annotators in the current study are not professionally trained in human translation, and the construction of a bilingual or multilingual corpus, translation is as important as annotation. Given this, one vulnerable point of the current corpus study is that translated texts were provided by those who do not have a professional experience in human translation. Nonetheless, using texts translated and proofread by non-experts prevents the current work from infringing on copy-right, which can be problematic in translations as well as source texts.

---

[1] EXMARaLDA has also been used in the Project SFB632 (`http://www.sfb632.uni-potsdam.de`) in Germany (Dipper et al., 2007).

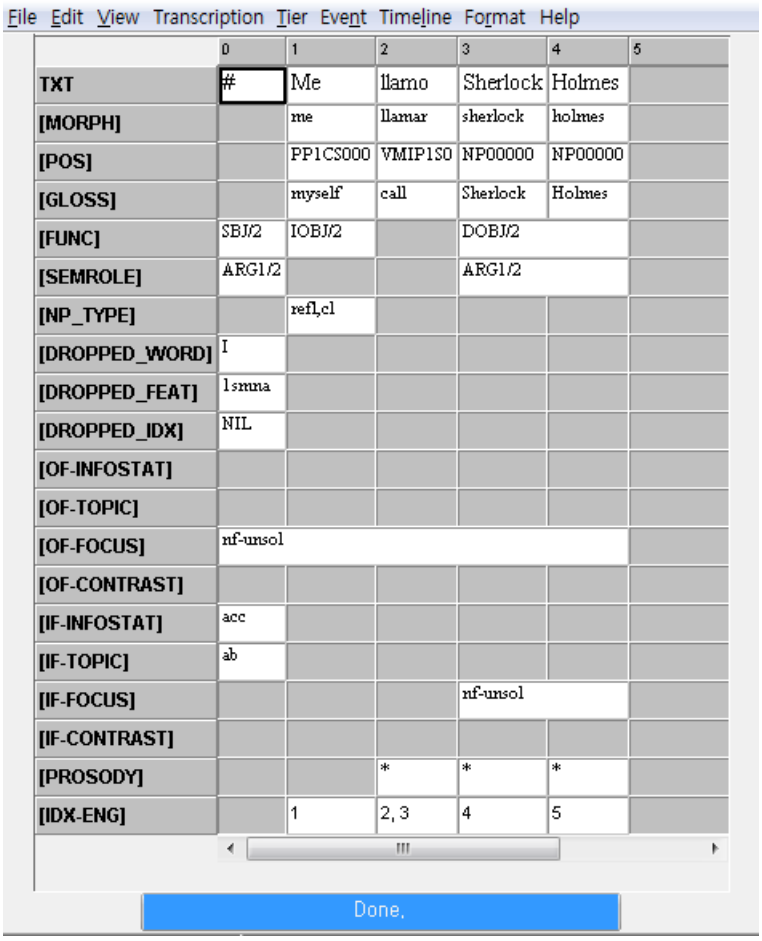


Figure 6.1: Screenshot of EXMARaLDA

Annotating #1 to #100

Sets of sentences from #1 to #100 were examined for each language. Each set in English, Spanish, Russian, and Korean was annotated three times. (ii-a) First, annotators taking charge of the three languages worked with me. We annotated the 100 sentences in the languages together. Whenever a new phenomenon was found in at least one language, we came up with a consistent way for annotating it across the languages under consideration. (ii-b) After rough versions were created, annotators revised their own annotations from the start to the end. (ii-c) Finally, other annotators, who knew the annotation schema, revised the entire set once again. The annotators who took charge of English

and Spanish cross-checked each others' annotation. For Russian and Korean, other annotators who took part in drawing the annotation schema from the beginning reexamined the annotation in the languages.

Annotating #101 to #110

In addition, 10 sentences in Spanish, Russian, and Korean were annotated for evaluation (§6.4). Two annotators for each language separately tagged sentences from #101 to #110. (iii) Consequently, sixty sentences in total were included as an evaluation set.

6.2.4 Preliminary Work

The preliminary steps for annotating the texts in four languages include the following: First, the original text of *The Adventure of the Speckled Band*, written in English, was obtained from an online resource (`http://www.gutenberg.org/files/1661/1661-h/1661-h.htm`). Second, annotators translated and proofread the first 100 sentences into the other three languages (Spanish, Russian, and Korean) in the way presented in the previous subsection. Third, the 400 sentences were aligned in a semi-automatic way, using a Python script which aligned the 100 sentences in each language, and then the translators revised the alignment between the English text and each translated text. The sentence alignment script that I created for this purpose is quite simple, when contrasted to machine learning approach commonly used when dealing with larger data sets. Because the texts to be annotated consist of only 100 sentences in each language, it was not necessary to use a machine learning tool for the sentence-alignment. Fourth, the text in four languages was cleaned-up, which included removing excess white spaces, etc. Next, tokenization was carried out, because segmentation is based on Dipper et al. (2007); all punctuation marks are present in separate cells in the [TXT] tier. This substep was automatically carried out using a Perl script. Finally, another script in Python converted the files in the line-by-line format into XML files in the schema of EXMARaLDA.

6.3 Annotation

As is well-known, information structure is relevant to various phenomena in human language. The annotation schema of the current work, thus, covers not merely information structure itself but

also relevant linguistic domains, such as prosody, morphology, semantics, syntax, etc. The tiers in the present schema reference (i) phonology, (ii) morphology, (iii) syntax, (iv) semantics, (v) dropped elements, (vi) information structure, and (vii) word-alignment across languages. They are summarized as follows.

(1) a. Phonology: accents

b. Morphology: morphemes, Part-Of-Speech, and glosses

c. Syntax: syntactic functions and semantic roles

d. Semantics: NP types (e.g. definiteness, possessives, demonstratives, etc.)

e. Dropping: dropped word, features, and index

f. Information Structure: information status, topic, focus, and contrast

g. Word Alignment: English↔Spanish, English↔Russian, and English↔Korean

The annotation schema of this corpus study is largely adapted from Dipper et al. (2007). Yet, the schema proposed in the guideline was altered in several ways in order to restrict attention to the information structure phenomena which this dissertation has a direct interest in.

First, the phonological layer has been eliminated except for a tier for marking stress, as there are currently no spoken data of *The Adventure of the Speckled Band*. Second, the morphological layer has been modified to reflect tags used by the POS taggers deployed by the current study. Third, most of the syntactic and semantic layers (with the exception of functional categories) and NP types have been removed. This is because these layers can be semi-automatically generated in future work by using the DELPH-IN grammars (e.g. ERG for English (Flickinger 2000), SRG for Spanish (Marimon 2012), RRG for Russian (Avgustinova and Zhang 2010), and KRG for Korean (Kim et al. 2011)). Fourth, information-structure related layers have not been significantly modified. These consist of INFOSTAT (information status, such as given, new, accessible), TOPIC (aboutness or frame-setting topic), FOCUS (new (un)solicited focus), and CONTRAST (contrastive topic or contrastive focus). The discourse-related information is bisected into two frames; inner frame *vs*. outer frame. Fifth, three additional tiers have been added for dealing with dropped items. Finally, IDX (InDeX) tiers were added to track word-alignment across the languages.

Table 6.1: Sample annotation in English (#24)

	0	**1**	**2**	**3**	**4**	**5**	**6**
TXT	“	My	name	is	Sherlock	Holmes	.
PROSODY					*	*	
MORPH	“	my	name	be	Sherlock	Holmes	.
POS	“	PP$	NN	VBZ	NP	NP	SENT
GLOSS							
FUNC		SBJ/3					
SEMROLE		ARG1/3					
NP_TYPE		poss					
DROPPED_WORD							
DROPPED_FEAT							
DROPPED_IDX							
OF-INFOSTAT							
OF-TOPIC							
OF-FOCUS	nf-unsol						
OF-CONTRAST							
IF-INFOSTAT					new		
IF-TOPIC		ab					
IF-FOCUS					nf-unsol		
IF-CONTRAST							
IDX-SPA		1	2		3	4	
IDX-RUS	0	1	2		3	4	5
IDX-KOR	0	1	2	5	4	5	6

Sample annotations of an English sentence and its corresponding Spanish sentence are presented in Table 6.1 and 6.2, respectively. Though the original sentence is quite simple, and both sentences convey almost the same interpretation, the constructions are quite different. (a) The English sentence is a typical identification copula construction, while the corresponding sentence in Spanish is realized with a ditransitive verb. That is, the literal meaning of the Spanish sentence is something like *I call myself Sherlock Holmes*. (b) The genitive pronoun *my* in English corresponds to the reflexive clitic *me* in Spanish. (c) Finally, the subject is dropped in Spanish, which is conditioned by the subject agreement morpheme attached to the verb *llamo* ‘(I) call’. The following subsections delve into the details of of each layer presented in (1), taking Table 6.1 and Table 6.2 as illustrative examples.

Table 6.2: Sample annotation in Spanish (#24)

	0	**1**	**2**	**3**	**4**
TXT	#	Me	llamo	Sherlock	Holmes
PROSODY			*	*	*
MORPH		me	llamar	sherlock	holmes
POS		PP1CS000	VMIP1S0	NP00000	NP00000
GLOSS		myself	call	Sherlock	Holmes
FUNC	SBJ/2	IOBJ/2		DOBJ/2	
SEMROLE	ARG1/2			ARG2/2	
NP_TYPE		refl,cl			
DROPPED_WORD	I				
DROPPED_FEAT	1smna				
DROPPED_IDX	NIL				
OF-INFOSTAT					
OF-TOPIC					
OF-FOCUS	nf-unsol				
OF-CONTRAST					
IF-INFOSTAT	acc				
IF-TOPIC	ab				
IF-FOCUS				nf-unsol	
IF-CONTRAST					
IDX-ENG		1	2, 3	4	5

6.3.1 Prosody

With respect to phonological factors, this corpus annotation is exclusively concerned with whether or not phonological accent falls on a word. The annotators read the sentence aloud considering the context, if a word was judged as bearing an accent, an asterisk ('*') was inserted into the corresponding cell in the [PROSODY] tier. A distinction between accents, such as distinguishing between A and B accents in English, was not considered, because distinctions at this level can be more difficult to accurately perceive. That is, patterns of accents may be differently detected by different annotators, regardless of their status as native speakers.

An example of accent annotation can be seen in Table 6.1. The name *Sherlock Holmes* in the fourth and fifth columns is pronounced with accents, while the verb *is* in the third column is not accented. The name is also accented in Spanish as indicated at the second last row of Table 6.2, and the verb *llamo* bears an accent, unlike *is* in English.

The main purpose of annotating the [PROSODY] tier is that the asterisks representing accents

can be used to observe how focus projection is related to prosodic realization. The focus of an accented item can spread into the larger constituent(s) (a.k.a. focus projection). That implies that the accent triggers the spreading of the focus domain.[2]

The [PROSODY] tier is marked in English, Spanish, and Russian, but not in Korean. This is because Korean is a language less sensitive to accents. As a preliminary observation, I selected ten sentences in Korean, and two Korean annotators tagged the [PROSODY] tiers for marking phonological prominence. Even within this small sample, there was annotator disagreement, with a quite different distribution of asterisk-marked cells across the two sets of annotation.[3] Thus, it appears that prosody in Korean cannot be reliably marked, using the methodology of the current study.

6.3.2 Morphology

The morphological tiers in this annotation are not based on traditional glossing formats, such as the IGT (Interlinear Glossed Texts) format. Converting words in a surface form into the IGT format can be time-consuming. In order to build up annotation in four languages in a relatively short time, this corpus study deploys automatic POS taggers for each language, rather than annotating these layers by hand. That is, the two morphological tiers ([MORPH] and [POS]) in EXMARaLDA are directly filled out using the results of these taggers. These automatically filled tiers can be converted again into the IGT format in an automatic way, if we can draw a mapping table between each tagger's tagset and the glosses in the IGT format. Because this conversion is not directly necessary in analyzing the multilingual text, that job is left to future work. Although they are not in the IGT format, the morphological information still needs to be looked at for two reasons. First, some morphemes in the languages (e.g. *i/ka* and (*n*)*un* in Korean) have an influence on articulation of information structure. Thus, information-structure related morphemes should be observed specifically. Second, I want this annotated corpus to be used in further studies. The availability of tagged information helps other researchers who can utilize and alter this data set for their own purposes.

This corpus study employed three POS taggers: For English and Russian, Treetagger (`http://`

[2]Additionally, I argue that focus can spread by means of other kinds of marking (e.g. lexical markers of expressing focus). Chapter 13 addresses F(ocus)-marking with respect to focus projection.

[3]In fact, the Russian annotators did not show such an agreement in the [PROSODY] tiers, either. This poor agreement eventually had an adverse effect on the Kappa testing (§6.4) in Russian. The degree of disagreement in Korean was worse than that in Russian.

Table 6.3: Tagset in syntactic tiers

tier	tag	type
FUNC	SBJ	subject in surface form
	OBJ	object of transitives in surface form
	IOBJ	indirect object of ditransitives in surface form
	DOBJ	direct object of ditransitives in surface form
SEMROLE	ARG1	the first argument in ARG-ST
	ARG2	the second argument in ARG-ST
	ARG3	the third argument in ARG-ST

`www.cis.uni-muenchen.de/~schmid/tools/TreeTagger`, Schmid 1994, 1995) was used. There is also a Spanish version of treetagger, but the Freeling tagger (`http://nlp.lsi.upc.edu/freeling/index.php`, Carreras et al. 2004) was used. This is mainly because SRG employs the Freeling tagger system-internally, so using it will allow the annotated text in Spanish to more easily be used in future work by the DELPH-IN members. Finally, the Espresso Tagger (`http://air.changwon.ac.kr`, Hong and Cha 2008) was deployed for Korean.

The [MORPH] tier includes the lemma of each word in a sentence. For example, the verb *llamo* '(I) call' in Spanish in the second column of Table 6.2 stems from *llamar* in the second row. The [POS] tier directly comes from each tagger. For example, 'VMIP1S0' in the [POS] tier of Table 6.2 stands for 'Verb (V) - Main (M) - Indicative (I) - Present (P) - 1st (1) - Singular (S) - No Gender (0)'.[4] The [GLOSS] tier was provided by annotators, and serves as an aid for revising word-alignment tables in the current work and converting the morphological tiers into the IGT format in the future work.

6.3.3 Syntax

There are only two tiers related to syntactic operations. One tier deals with functional categories (e.g. SBJ, OBJ, IOBJ, and DOBJ) in the surface form, and the other is concerned with semantic roles. In the context of HPSG/MRS-based grammar engineering, the functional categories specified in the [FUNC] tier are identical to what verbal items have in VAL, while those in the [SEMROLE] tier are represented as ARG#n in RELS of verbal items. The information of ARG#n comes from ARG-

[4] `http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html`

ST (ARGument-Structure) verbal items lexically involve. In the [SEMROLE] tier, as represented in the indexing number, ARG1 corresponds to the first element in the list of arguments, ARG2 corresponds to the second one, and ARG3 corresponds to the third on, respectively. The tagset is given in Table 6.3.

The point to be considered is that [FUNC] and [SEMROLE] are not necessarily the same as each other. For instance, in passive constructions, a syntactic subject (i.e. a promoted argument) plays a role of ARG2 or ARG3. This way of annotations is different from the guideline Dipper et al. (2007) offer. Their annotation does not care of argument structure, and instead thematic roles (e.g. agent, theme, experience, etc.) are marked. However, because this corpus work is supposed to be used for creation of an MRS-based model, information about ARG#n is crucial.

The number after slash in each cell of the [FUNC] and [SEMROLE] tiers represents the index of the verb that dominates the argument. For example, SBJ/2 in the [FUNC] tier of Table 6.2 means that the dropped element syntactically functions as the subject of *llamo*, whose index is 2. One constituent can sometimes play two or more functions at the same time. For example, the head noun of relative clauses has two different functions with respect to the matrix clause and the relative clauses; all functions the constituent has are annotated, delimited by a comma.

6.3.4 Semantics

Semantics in this corpus annotation was only tagged with respect to NP types. In many previous studies on information structure, NP types are assumed to be relevant to configuring information structure. However, it is my argument that definiteness is neither a necessary nor a sufficient condition for focus or topic. NP types were tagged as indicated in Table 6.4.

The tags in Table 6.4 are attached to NPs with specific determiners (e.g. *the* for definiteness in English) or special types of NPs that make a contribution to semantics (e.g. bare plurals in English (Diesing, 1992)). The annotation for NP types was done in a greedy way as far as possible just as regular expressions ordinarily do. For example, 'the ... of the ...' is wholly marked as [def], though they consist of two different definite NPs. If an NP has two or more determiners at the same time, multiples tags are inserted into the corresponding cell(s), using commas as a delimiter. For instance, *only the two ...* is tagged as [uni,def,num].

Table 6.4: Tagset of NP types

tag	**type**	**example(s)**
all	universal quantifiers	'all men'
any	NPI	'any sound
bare	bare NPs	'grown-ups'
cl	clitics	'me' in Spanish
def	definiteness	'the little prince'
dem	demonstratives	'this flower'
dist	distributives	'each day'
ind	indefiniteness	'a sheep'
kind	kindness	'such power'
mul	multal	'many years ago'
neg	negative determiners	'no reply'
num	numeral expressions	'six years'
ord	ordinal	'first', 'tenth'
pau	paucal	'little', 'few'
poss	possessives	'my cold'
refl	reflexive	'self'
sup	superlative	'most'
uni	uniquitive	'the only'
wh	wh-words	'what'

The tagset presented in Table 6.4, to my knowledge, cover almost all types of NPs that have been studied in the field of formal semantics. As I mentioned in Chapter 3, I argue that definiteness is not a necessary or sufficient condition for focus or topic. That is, NP types may show a tendency in correspondence with respect to information structure, but there is no bidirectional restriction. NP types were annotated in this corpus study in order to explore the relationship and prove my claim.

6.3.5 Pro-drop

One of the main goals of this corpus study is to offer corroborative evidence about the interaction between *pro*-drop and focus/topic, using a multilingual parallel text that covers four types of pronominal distribution: English does not allow *pro*-drop, Spanish allows subject-drop, Russian allows *pro*-drop in certain contexts, and Korean allows topic-drop (Sohn, 2001).

If an annotator finds a dropped element in a sentence, '#' is inserted into the position in the [TXT] tier. For example, the first column in Table 6.2 is assumed as a dropped subject in Spanish. Tiers prefixed by 'DROPPED' point to the missing expression (DROPPED_WORD), the properties

of the dropped element (DROPPED_FEAT), and index of its antecedent (DROPPED_IDX). For example, in the first column of Table 6.2, 'I' is the English translation corresponding to the dropped subject in Spanish, '1smna' in the next row indicates the agreement features of the dropped element, such as '1st - singular - masculine - nominative - animacy', and 'NIL' in the next row means there is no referent corresponding to the dropped subject in the previous sentences. If a referent appears previously, the cell is tagged as the index of the referent; for instance, 12.5 means the dropped element refers to the fifth word in the twelfth sentence. Among them, tiers of [DROPPED_WORD] have an English word corresponding to the word in each language. This tier has two purposes. First, the tier is a note when annotators proofread their own annotation and cross-check other annotator's annotation. They are not directly related to creating linguistic generalization, but used for ease of annotation. Second, when the morphological layers need to be converted into the IGT format, the dropped element should be selectively handled. If the IGT conversion requires dropped elements to be included in the analysis, the English words in [DROPPED_WORD] should be separately used.

6.3.6 Inner Frame vs. Outer Frame

The current schema makes use of two different frames, unlike Dipper et al. (2007). The IF (Inner Frame) and OF (Outer Frame) layers have the same set of tiers and differentiate two types of discourses: dialogues between characters within the story (IF) and Dr. Watson's narration (OF). The OF relates to the author's representation of the reader's understanding of the common ground, while the IF (ostensibly) reflects the characters' representation of the common ground. For example, in (2), the parenthetical sentence within quotation marks *The name is familiar to me,* is analyzed as inner frame, while the whole sentence including the quotation is analyzed as outer frame.

(2) " The name is familiar to me , " said he .
INNER FRAME
OUTER FRAME

The sentence that has both inner frame and outer frame was annotated as shown in Table 6.5. The [OF-INFOSTAT] tier has an 'active' tag on *he* in the outer frame, while the [IF-INFOSTAT] tier has an empty cell for the word. Instead, the [IF-INFOSTAT] tier has an 'active' tag on *The name* and an 'inactive' tag on *me*. The whole quoted sentence is annotated as 'nf-unsol' in the outer frame, but in the inner frame only *is familiar to me* is annotated as focus.

Table 6.5: Sample annotation in English (IF/OF) (#73)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
TXT	“	The	name	is	familiar	to	me	,	”	said	he	.
PROSODY			*		*							
MORPH	“	The	name	be	familiar	to	me	,	”	say	he	.
POS	“	DT	NN	VBZ	JJ	TO	PP	,	”	VBD	PP	SENT
GLOSS												
FUNC		SBJ/3									SBJ/9	
SEMROLE		ARG1/4									ARG1/9	
NP_TYPE		def										
DROPPED_WORD												
DROPPED_FEAT												
DROPPED_IDX												
OF-INFOSTAT											active	
OF-TOPIC												
OF-FOCUS	nf-unsol											
OF-CONTRAST												
IF-INFOSTAT		active					inactive					
IF-TOPIC		ab										
IF-FOCUS				nf-unsol								
IF-CONTRAST												
IDX-SPA	0	1	2		5		3	6	7	8		10
IDX-RUS	0	1	2		4		3		6	7	8	9
IDX-KOR	0	1	2	5	3, 4			6	7			

Table 6.6: Tagset of information status and information structure

tier	**index**	**tag**	**type**	**note**
INFOSTAT	a-i	new	new	newly introduced
	a-ii	acc	accessible	inferable from the context
	a-iii	inactive	inactive	not recently mentioned
	a-iv	active	active	mentioned immediately before
TOPIC	b-i	ab	aboutness topic	what the speaker is talking about
	b-ii	fs	frame-setting topic	what restricts the domain of what is spoken
FOCUS	c-i	nf-unsol	new focus unsolicited	inomissible but not replying to a question
	c-ii	nf-sol	new focus solicited	what answers to a question
CONTRAST	d-i	cf	contrastive focus	occurring with alternative(s) and inomissible
	d-ii	ct	contrastive topic	occurring with alternative(s) but omissible

6.3.7 Information Status and Information Structure

This subsection describes the most critical annotation in the current study: information status and information structure. The tagset for information-structure related categories, largely taken from Dipper et al. (2007), is presented in Table 6.6.

Annotating the multilingual parallel text in this corpus study, the most important part is identifying components of information status and information structure in an objective and coherent way. The procedure for diagnosing related components is comprised of three steps.

Information Status

In §3.1, I argue that information status is neither a necessary nor a sufficient condition for information structure. Nonetheless, two tiers of [INFOSTAT] (one in the inner frame (IF), one in the outer frame (OF)) were annotated in this corpus construction to substantiate the relationship later. Information status is tagged only in each language, not seeing how it is realized in other languages. Note that not all translations necessarily exhibit word-to-word correspondance in other languages. For example, subjects are seldom dropped in English, but they are often missing in Spanish (a subject-drop language) and Korean (a topic-drop language).

(a) There are four components of information status in this annotation: namely (a-i) new, (a-ii) accessible, (a-iii) inactive, and (a-iv) active. These are largely adapted from Dipper et al. (2007) who classify topics into three subtypes: active (i.e., the antecedent can be determined within the last or the current sentence), inactive (i.e., the antecedent exists, but is not active), and accessible (i.e., there is no antecedent, but it can be inferred by the situational context or our world knowledge). In line with the classification, information status is annotated as follows: Event-denoting expressions, such as verbs, adverbs, are not tagged with respect to information status. If a referring expression is newly appears without any referents in the text, the item is tagged as (a-i) 'new'. If there is no referent to an item, but the item can be inferred by the situational context or our world knowledge, the item is tagged as (a-ii) 'acc' (i.e. accessible). If a referent is already mentioned previously, but is not active, the referring expression is tagged as (a-iii) 'inactive'. Finally, if a referent is mentioned in the immediate preceding sentence or in the same sentence, the referring expression is tagged as (a-iv) 'active'.

Information Structure in a Monolingual Domain

Information structure components are identified using several diagnostic tests preferentially in each language. These are the same tests to vet information structure markings and meanings discussed in

Chapter 3. The tagset is also adapted from Dipper et al. (2007).

As mentioned earlier (§3.2.4), there are two diagnoses for focus: namely the deletion test and *wh*-questions. Regarding the corpus annotation, as pointed out in Chapter 3 (§3.2.4), the former is conducted first, and then the latter is done subsidiarily. It is true that the *wh*-questions test has been more widely used for identifying focus, but for the purpose of annotating naturally occurring texts the deletion test needs to be carried out first. This is because applying the *wh*-test to corpus annotation often causes confusion, as Gracheva (2013) indicates. In other words, it is almost impossible to separate a single sentence from the context and test it independently, when we look at a running text.

(b) Focus can be identified with reference to whether or not the constituent is omissible, because inomissibility is regarded as the key notion in focus realization in this dissertation. If an annotator judges a constituent incapable of being dropped in the context, it is evaluated as conveying focus meaning. In other words, if the sentence without a particular constituent sounds infelicitous to annotators (when a constituent that does not carry contrastiveness is), the constituent is labeled as 'nf-', which stands for 'new focus'. More specific type of 'nf-' is identified with reference to *wh*-questions. If the constituent with 'nf-' is an answer to a *wh*-question in its previous sentence(s), it is tagged as (b-i) 'nf-sol' (new focus - solicited). Otherwise, every constituent with 'nf-' is tagged as (b-ii) 'nf-unsol' (new focus - unsolicited).

(c) Since it is my position that topics are not directly associated with information status, topic is tagged without reference to the [INFOSTAT] tiers. For example, although a constituent is tagged as 'active' in [INFOSTAT] or whatever, annotators do not refer to the tag in annotating the [TOPIC] tiers. Instead, annotators are required to employ the two tests for topic: namely the *tell-me-about* test (Choi, 1999) and the paraphrasing test presented in Oshima (2009) and Roberts (2011). The *tell-me-about* test verifies that constituents associated with topic in Korean should be marked by (*n*)*un* though the opposite way is not necessarily true (i.e., not all (*n*)*un*-marked constituents are evaluated as containing topic meaning.). Hence, only (*n*)*un*-marked constituents can be analyzed as (c-i) 'ab' (aboutness topic) in Korean. For other languages (e.g. English, Spanish, and Russian), if a constituent pass one of the paraphrasing tests, the constituent is tagged as 'ab'. The *as-for* test Oshima proposes is repeated in (3).

(3) The *as for* test: If an utterance of the form: [$_{S_1}$ ... X ...] can be felicitously paraphrased as [*As for* X, S_2] where S_2 is identical to S_1 except that X is replaced by a pronominal or empty form anaphoric to X, X in S_1 is a topic. (Oshima, 2009, p. 140)

When an annotator found one constituent serving as topic, (s)he tried to paraphrase the constituent into one of the expressions presented by Oshima and Roberts (e.g. *about ...*, *what about ...*, *as for ...*, and *speaking of ...*). If the paraphrased sentence sounds completely natural to the annotator, 'ab' is inserted into the corresponding cell of the [TOPIC] tiers.

Frame-setting topics are independently annotated. When an expression that entails a temporal, spatial, manner, or conditional meaning turns up clause-initially, the constituent is tagged as (c-ii) 'fs' (frame-setting), following the notion in previous literature (Li and Thompson, 1976; Chafe, 1976; Lambrecht, 1996; Féry and Krifka, 2008). In fact, there is one potential problem in this way of annotation, because I used a syntactic criterion for frame-setting topics. It could be a problematic point that tagging 'fs' hinges on a theoretic assumption, rather than a reliable linguistic test. Therefore, I cannot firmly say that frame-setting topics are always sentence-initial with reference to the corpus data. Nonetheless, it is my belief that using the annotated text I can establish a generalization about distributional properties of frame-setting topics. That is to say, with reference to the corpus annotation, I can verify whether frame-setting topics can occur multiple times in a single clause and whether they are followed by aboutness topic.

(d) Contrast entails an alternative set in the context. If a constituent is labeled as 'c' in the [CONTRAST] tiers, there must be an alternative in the context. The distinction between (d-i) 'cf' (contrastive focus) and (d-ii) 'ct' (contrastive topic) depends on whether the constituent can be felicitously elided in the context. When a constituent with a contrastive meaning disappears, if the sentence without it sounds awkward to annotators, 'cf' is inserted into the corresponding cell. Otherwise, 'ct' is added.

Information Structure in a Multilingual Domain

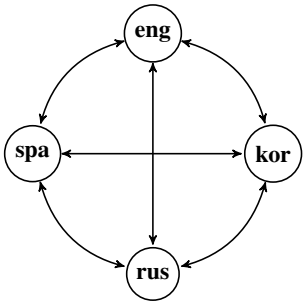
There are some cases to which the diagnostic tests presented thus far are not straightforwardly applicable. In those cases, it might be better to take advantage of a multilingual parallel text referring to annotation in other languages. If a constituent is evaluated as a specific information structure com-

ponent (e.g. ‘nf-sol’, ‘ab’, ‘ct’, etc.) in some or all other languages, the corresponding constituent in the current language is preferentially tagged as the same, other things being equal. This task was done on a second pass when the information was available in other languages. For example, since Korean employs lexical markers such as *i/ka* and (*n*)*un* for expressing information structure, the distinction between non-topic and topic is relatively clear in Korean annotation. To take another instance, we can surely identify which element is focused in cleft constructions in English. These kinds of information can be used when annotators sometimes have trouble in tagging information structure components in other languages.

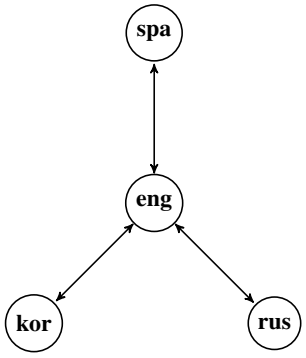
6.3.8 Word Alignment

Ideally, word alignment of this corpus annotation would be created in the way represented in (4a), in which two-way arrows connect all languages. Each language would have three alignment tables to the other languages, yielding twelve tables (3×4) in total. Yet, it would be a difficult task to create so many tables in a relatively short time and with restricted resources. In particular, constructing the word alignment tables as in (4a) would require the annotator(s) to have linguistic knowledge in the four languages. This is almost impossible for this corpus study.

(4) a.



b.



Alternatively, word alignment of the current work has been created as schematized in (4b), in which English functions as the pivot language between the other three languages. That it, each text in Spanish, Korean, and Russian has only two-way alignments (i.e. English to the language, and vice versa.). This indirectly facilitates word alignments among the texts in non-English languages. As a result, six alignment tables (3×2) are created.[5]

[5]During the refinement of the word alignment tables, grammatical diversity across four languages naturally raised

Word alignments along these six pairs were made in a semi-automatic way using GIZA++ (Och and Ney, 2003). Machine learning toolkits including GIZA++ normally require a large amount of data. The sentence pairs of the current work consist of only 400 sentences (100 in each language). This number is too small to run GIZA++ with. In order to make up for the shortage of data, more sentence pairs were added into the training sets. For Spanish↔English, 45,000 sentence pairs taken from Europarl Parallel Corpus '96, '97, '98, and '99 (`http://www.statmt.org/europarl`, Koehn 2005) were added after the 100 sentences of the current work. For Korean↔English, I made use of 30,000 sentences randomly chosen from the *Sejong* bilingual corpora,[6] in which translations from English to Korean account for approximately 50%, and vice versa. These four alignment tables, then, were manually refined by native speakers. For Russian↔English, as I could not obtain a freely available bilingual corpus,[7] I had no choice but to run GIZA++ with only 100 sentence pairs. In order to compensate for the potential shortcomings in the Russian↔English word alignment tables, two different annotators refined the automatically constructed alignment tables by hand.

6.4 Evaluation

In order to verify whether the annotation was made reliably, I compute Kappa (κ) measurement to quantify how much two annotators agree with each other (Cohen, 1960; Carletta, 1996). One advantage of Kappa testing is it can subtract agreement by chance from the result, which ensures more reliable measures. Kappa is defined as follows. $Pr(o)$ means the relative observed agreement between annotators, and $Pr(e)$ stands for the hypothetical probability of chance agreement.

$$\kappa = \frac{Pr(o) - Pr(e)}{1 - Pr(e)} \tag{6.1}$$

Dipper et al. (2007) present how to use Kappa testing in the corpus construction for the study of information structure. They claim that $\kappa > 0.8$ is indicative of fairly reliable agreement, and 0.67

several issues. The linguistic phenomena that involve variation across four languages include expletives, relatives, copulae, phrasal verbs, progressive forms, possessive verbs, and so forth. Because such a variation in languages is out of the scope of this dissertation, they are left to the future work.

[6]There seems to be no canonical reference that gives an explanation in English about the *Sejong* bilingual corpora, to my knowledge. Instead, the overall explanation of configuration of the corpora is presented in Song and Bond (2009).

[7]As a monolingual corpus in Russian, Russian National Corpus (Grishina, 2006) can be of use to creating a data-based generalization about information structure in Russian (Gracheva, 2013).

$< \kappa <$ 0.8 allows for tentative conclusions to be drawn.

The English sentences from #101 to #110 are given in (5). As explained in §6.2.3, the evaluation was conducted using the corresponding sentences in the other three languages.

(5) <eng:#101> Julia went there at Christmas two years ago,
and met there a half-pay major of marines, to whom she became engaged.
<eng:#102> My stepfather learned of the engagement
when my sister returned and offered no objection to the marriage;
<eng:#103> but within a fortnight of the day which had been fixed for the wedding,
the terrible event occurred which has deprived me of my only companion."
<eng:#104> Sherlock Holmes had been leaning back in his chair with his eyes closed
and his head sunk in a cushion,
but he half opened his lids now and glanced across at his visitor.
<eng:#105> "Pray be precise as to details," said he.
<eng:#106> "It is easy for me to be so,
for every event of that dreadful time is seared into my memory.
<eng:#107> The manor-house is, as I have already said, very old,
and only one wing is now inhabited.
<eng:#108> The bedrooms in this wing are on the ground floor,
the sitting-rooms being in the central block of the buildings.
<eng:#109> Of these bedrooms the first is Dr. Roylott's, the second my sister's,
and the third my own.
<eng:#109> There is no communication between them,
but they all open out into the same corridor.

The measurements in each sentence from #101 to #110 are also indicated in Table 6.7. Additionally, Kappa measures in each sentence are visualized in Figure 6.2 for ease of comparison. In total, Kappa is calculated at 84.79% in Spanish, 81.74% in Russian, and 84.88% in Korean. The average of them is 83.80%. However, this average is not so suggestive, because the word length in each language is not the same. The average word length of sentences #101 to #110 in Spanish is 22.9, that in Russian is 19.9, and that in Korean is 14.4. That is, sentences in Spanish are longer than those in the other languages, and sentences in Korean are relatively short. In order to draw a more trustworthy conclusion, I calculated Kappa once again pretending the three test sets in Spanish, Russian, and Korean as a single test set consisting of 30 sentences. This Kappa value is 83.75%. Given that both

Table 6.7: Kappa measures (κ) – All Tiers

	Spanish			Russian			Korean		
# of sentences	$Pr(o)$	$Pr(e)$	κ	$Pr(o)$	$Pr(e)$	κ	$Pr(o)$	$Pr(e)$	κ
101	94.75%	63.46%	85.63%	92.22%	66.89%	76.51%	98.41%	68.19%	95.01%
102	95.33%	63.16%	87.33%	93.73%	64.24%	82.46%	95.92%	70.58%	86.13%
103	98.97%	64.22%	97.13%	96.52%	64.96%	90.07%	97.62%	71.44%	91.66%
104	93.73%	75.19%	74.71%	96.25%	75.38%	84.77%	99.29%	78.12%	96.74%
105	95.56%	67.44%	86.35%	96.67%	72.75%	87.77%	95.54%	74.78%	82.30%
106	98.33%	64.51%	95.30%	90.18%	64.37%	72.43%	95.05%	72.84%	81.80%
107	95.83%	64.34%	88.32%	97.33%	71.33%	90.70%	92.86%	64.97%	79.61%
108	91.21%	63.96%	75.62%	97.92%	61.86%	94.54%	97.25%	67.24%	91.61%
109	91.11%	60.61%	77.44%	87.78%	61.06%	68.61%	88.66%	61.20%	70.76%
110	90.56%	66.25%	72.02%	88.57%	63.11%	69.02%	92.86%	70.01%	76.18%
Total	94.73%	65.35%	84.79%	93.97%	66.98%	81.74%	95.39%	69.50%	84.88%

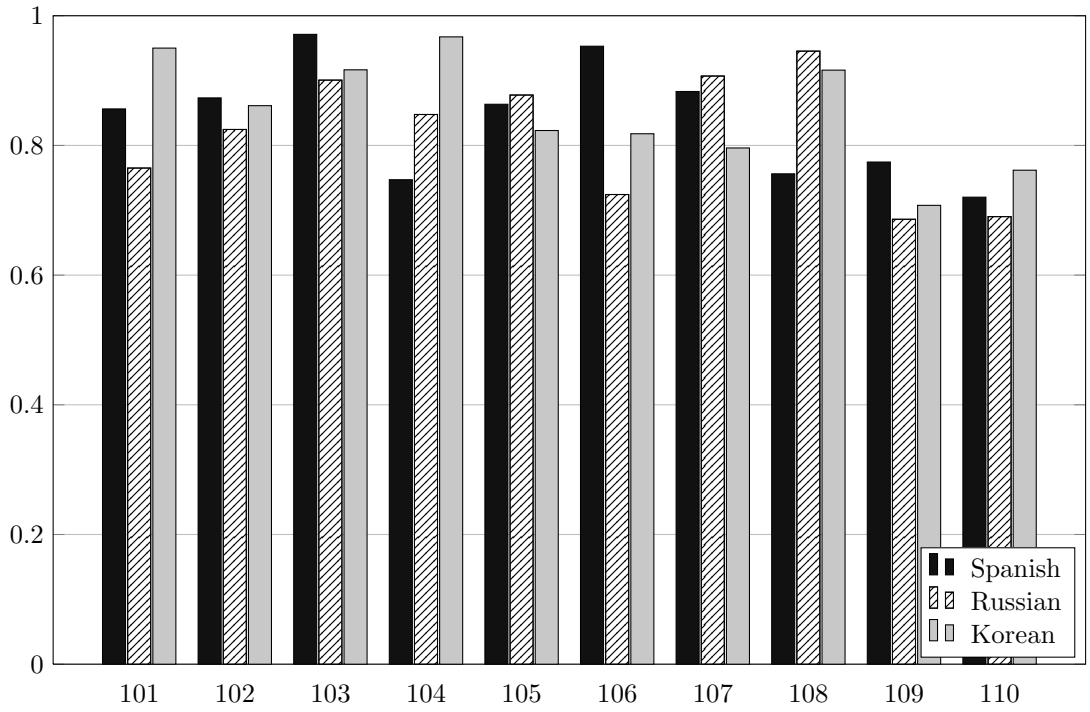


Figure 6.2: κ from #101 to #110 – All Tiers

the values (83.80% and 83.75%) are greater than 80%, we can say that the annotation was reliably created.

Table 6.8: Kappa measures (κ) – OF/IF Tiers

	Spanish			Russian			Korean		
# of sentences	$Pr(o)$	$Pr(e)$	κ	$Pr(o)$	$Pr(e)$	κ	$Pr(o)$	$Pr(e)$	κ
101	93.56%	57.02%	85.02%	91.67%	60.94%	78.67%	99.31%	58.56%	98.34%
102	93.79%	57.46%	85.40%	97.79%	61.42%	94.28%	100.00%	59.96%	100.00%
103	97.60%	60.42%	93.93%	97.83%	57.21%	94.92%	98.96%	63.02%	97.18%
104	93.43%	75.00%	73.72%	94.56%	75.29%	77.99%	100.00%	84.26%	100.00%
105	90.09%	64.31%	72.22%	97.66%	69.04%	92.43%	95.31%	66.60%	85.96%
106	97.92%	60.96%	94.66%	88.24%	62.66%	68.50%	96.15%	64.42%	89.19%
107	94.79%	62.89%	85.96%	98.11%	65.30%	94.56%	94.55%	61.17%	85.95%
108	89.77%	60.83%	73.89%	96.09%	59.23%	90.42%	96.30%	64.15%	89.67%
109	89.29%	54.79%	76.30%	90.97%	60.13%	77.36%	94.89%	58.22%	87.77%
110	85.11%	60.75%	62.05%	94.64%	61.56%	86.06%	88.39%	62.02%	69.44%
Total	93.19%	61.35%	82.38%	94.67%	63.32%	85.46%	96.88%	65.25%	91.02%

I also calculated Kappa using only information-structure related tiers (OF/IF Tiers). The numbers are given in Table 6.8. All numbers in the κ columns are more than 0.67 with one exception in <spa:#110>, and the total measures are more than 0.8 (82.38% in Spanish, 85.46% in Russian, and 91.02% in Korean). These numbers also indicate that the annotation was reliably created. Thus, we can say that the annotated data provides a trustworthy conclusion.

It is noteworthy that Kappa values for #109 are relatively low in both Table 6.7 and Table 6.8. There are two reasons for the low values. The first reason is that the sentence is an ellipsis construction as shown in (5). The syntactic layer for the sentence is differently analyzed by the two annotators. The second reason is that the sentence involves contrast. This implies that the annotation in [CONTRAST] tiers has a room for refinement in the future.

For ease of comparison, Figure 6.4 includes four charts. The upper left chart is illustrative of overall differences between Kappa using all tiers (i.e. Table 6.7) and Kappa only using information-structure related tiers (i.e. Table 6.8). This chart indicates a slight decrease in Spanish, a slight increase in Russian, and a significant increase in Korean (84.88%→91.02%, in total). The factors influencing on them are as follows: (i) In Spanish, I noticed that dropped subjects were sometimes differently annotated. For this reason, several bars whose sentences include dropped subjects are rather short. Notably, the bars for Spanish on #104, #105, and especially #110 in Figure 6.3 are relatively short. (ii) In Russian, I noticed that accents were variably annotated. The bottom left

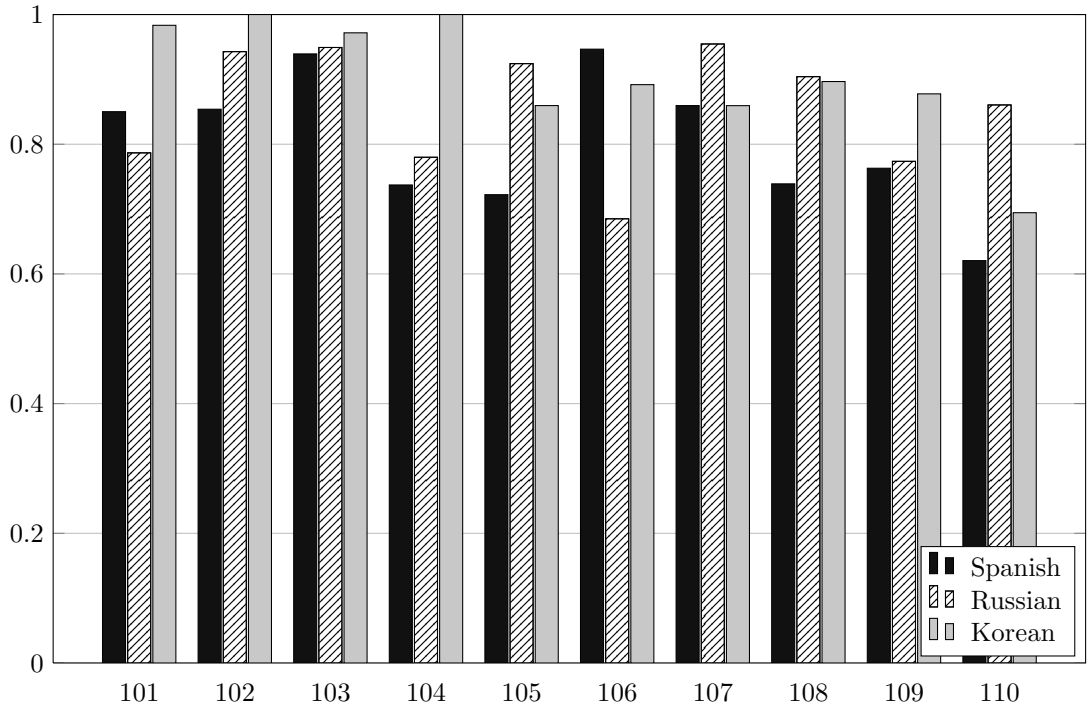


Figure 6.3: κ from #101 to #110 – OF/IF Tiers

chart of Figure 6.4 shows that values excluding the prosody tiers usually increase. (iii) In Korean, there are two factors adversely impacting Kappa. First, light verb constructions consisting of a verbal noun plus a light verb were differently tagged by the two annotators. For instance, one annotator regarded the two words *yakhon-ul ha-* 'engagement-ACC LV-' in (6a) as a single entry, but the other took only the verbal noun *yakhon* as the verb not considering *ha-* as a light verb. Second, the syntactic tiers of relative clauses were variably annotated in the syntactic layer. One annotator annotated them quite strictly, and the other did less strictly. For example, *saken* 'event' in (6b) may or may not be analyzed as the dependent of *mwusimwusiha* 'terrible'.[8]

(6) a. ... yakhon-ul ha-yss-eyo.
... engagement-ACC LV-PAST-DECL
'... was engaged.' <kor:#101>

[8]In fact, this is one of the hot topics in Korean syntax. Since this corpus study does not care about them, I would not like to provide further discussion. For more information, see Kim and Park (2000) and Sohn (2001).

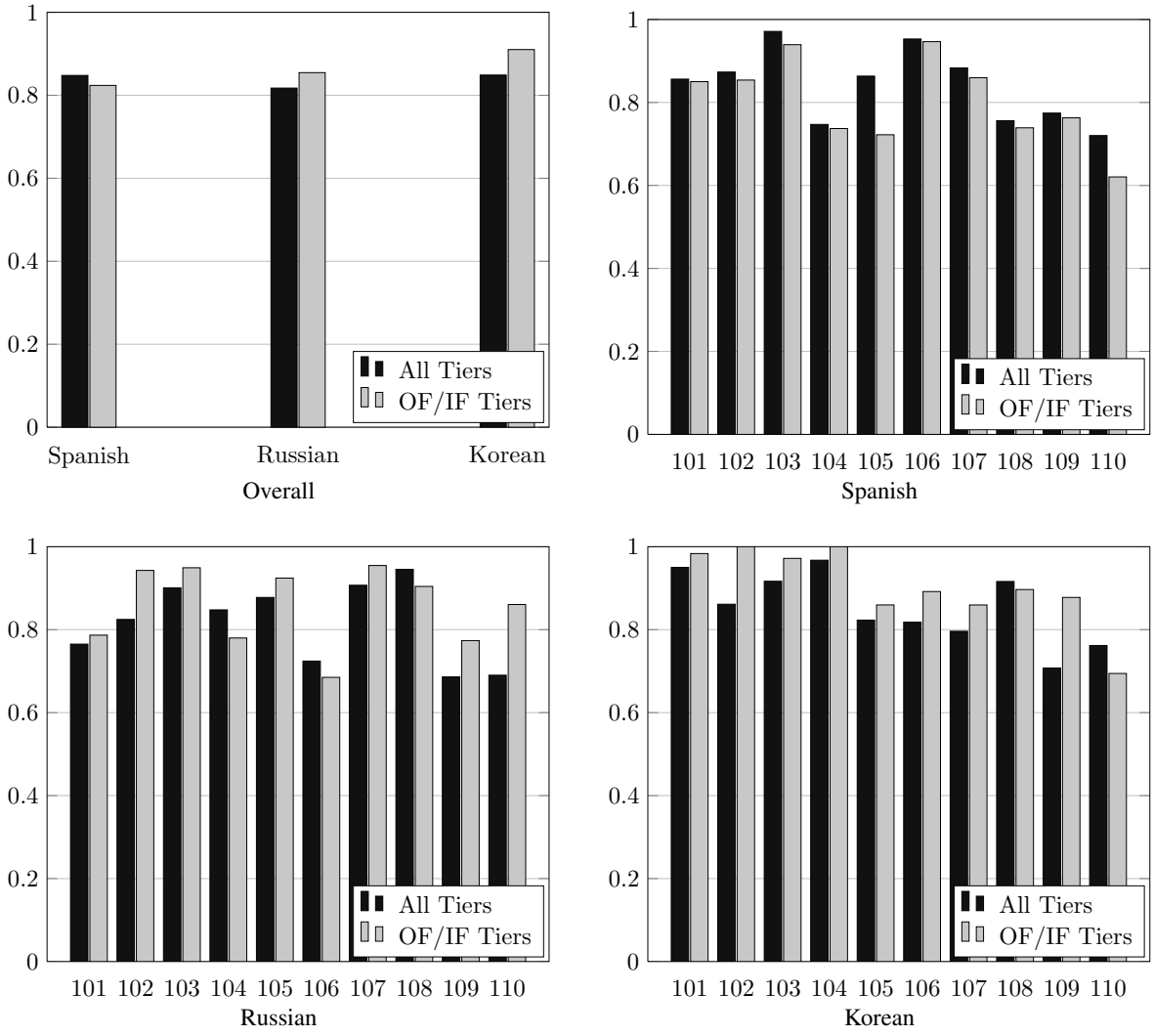


Figure 6.4: Comparison between All Tiers and OF/IF Tiers

b. ... mwusimwusiha-n saken-ul ...
... terrible-REL event-ACC ...
'... the terrible event ...' <kor:#103>

Despite the differences, both of them are not directly related to information structure. Accordingly, information-structure related tiers in Korean are quite similarly annotated by the two annotators.

The measurement presented thus far implies that the annotation of sentences #1 to #100 was also

reliably constructed. Building upon them, the following chapter explores the results of the corpus study.

6.5 Contributions

The data set constructed hitherto can be accessed in `svn://lemur.ling.washington.edu/shared/bebo`.[9] This resource is provided as an open source (i.e. free of charge) within the MIT license (`http://opensource.org/licenses/MIT`).

This corpus data can make a significant contribution to further work. First, the corpus data the current work establishes can be used to improve the stored libraries in the LinGO Grammar Matrix system (Bender et al., 2010b). Other library developers can utilize the data set to verify whether their grammar library works feasibly from a cross-linguistic and data-based viewpoint. Second, this resource can be used to the benefit of the study of *pro*-drop, given that this corpus annotation includes the major types of *pro*-drop: Spanish is a subject-drop language, Korean is a topic-drop language, and English and Russian rarely employ *pro*-drop. From a point of theoretical linguistics, this data set can be used for better analysis of the relationship between pronouns and discourse salience (Kaiser, 2009; Arnold, 2010). From a point of natural language processing, this data set has the potential to be helpful in producing higher performance in (multilingual) anaphora resolution and machine translation (Mitkov et al., 1995; Mitkov, 1999).

[9]This address may be changed in the future for one reason or another. If you want to download it and have any trouble, feel free to contact to the author.

Chapter 7

A CORPUS STUDY: ANALYSIS

Building upon the annotated text, this chapter creates linguistic generalizations about information structure from a multilingual perspective. §7.1 observes the marking systems of expressing information structure in the four languages (i.e. English, Spanish, Russian, and Korean), and captures cross-linguistic findings. §7.2 looks into how information structure was annotated in specific constructions (e.g. clefts and focus/topic fronting) in a comparative way. §7.3 looks at relative clauses, complement clauses, and adverbial clauses with respect to information structure.

7.1 Markings

7.1.1 Prosody of Expressing Information Structure

Regarding the correlation between prosodic patterns and information structure meanings, it turns out that every part tagged as 'nf-(un)sol' includes at least one accented word in English, Spanish, and Russian. On the other hand, all accented words are not necessarily inside a focus domain labeled as 'nf-(un)sol'. For example, in Table 6.2 (presented in the previous chapter), the constituent annotated as 'nf-unsol' in the sentence is an NP, *Sherlock Holmes*, and is accented and marked with asterisks in the [PROSODY] domain. The verb *llamo* 'called' bears an accent, but it is not in the scope of 'nf-unsol'. The accented words in the scope of 'nf-unsol' are assumed to function as the core that spreads focus meaning to its larger constituent(s). That is, in languages in which prosody is mainly responsible for expressing information structure (e.g. English, Spanish, and Russian), it seems true that F(ocus)-marking is licensed by an accent (Zubizarreta, 1998), and an extended focus domain should contain at least one F(ocus)-marked word therein. In a nutshell, the annotated data in this corpus study implies that the following rules offered by Büring (2006) are correct for languages in which prosody is crucially used for expressing focus.

(1) a. Basic Focus Rule: An accented word is F-marked.

b. Focus of a sentence: An F-marked constituent not dominated by any other F-marked constituent.

c. Focus Projection: either (i) F-marking of the head of a phrase licenses F-marking of the phrase, or (or both) (ii) F-marking of an internal argument of a head licenses the F-marking of the head. (Büring, 2006, p. 322–323)

7.1.2 Lexical Markers of Expressing Information Structure

Lexical Markers in Korean

This subsection reviews some Korean sentences which exemplify distinction between (*n*)*un* and the ordinary case markers, such as *i*/*ka* for nominatives and (*l*)*ul* for accusatives. There are a lot of previous studies about the topic marker in Korean, but most of them investigate quite simple sentences created by linguists, without reference to naturally occurring texts.[1] The data of this corpus study consists of only 100 sentences, but there are several sentences showing peculiar properties in usage of (*n*)*un* as compared to *i*/*ka* and (*l*)*ul*.

First, *pwuin* 'madam', *na*(*y*) 'I', and *ne* 'you' in (2) convey contrastive meanings in that they form an alternative set with each other. All of them are annotated as just 'cf', because they are not omissible though (*n*)*un* is not attached to the words. If the nominative markers attached to subjects are replaced with (*n*)*un*, the sentence still sounds good. This means that subjects associated with contrast are not necessarily (*n*)*un*-marked. On the other hand, the accusative markers attached to objects cannot be replaced with (*n*)*un*. This implies there must be an asymmetry in using (*n*)*un* with subjects *vs*. objects. It seems clear that contrastive foci are not always realized with (*n*)*un* in Korean.

(2) nwuka hetusun pwuin-ul kkayw-ess-ko,
somebody Hudson madam-ACC wake-PST-and,
pwuin-**i** na-**lul**,
madam-**NOM** me-**ACC**
nay-**ka** tto ne-**l** kkaywu-n ke-y-a.
I-**NOM** again you-**ACC** wake-REL thing-COP-DECL
'Mrs. Hudson has been knocked up, she retorted upon me, and I on you.' <kor:#10>

However, subjects associated with contrastive topic should be marked with (*n*)*un*. In (3), if *elkwul-kwa cheyhyeng* 'face and figure' is combined with *i*/*ka*, the sentence sounds less natural

[1] Yoo et al. (2007) explore the *Sejong* bilingual corpora, and conclude that there is no significant correlationship between markers and definiteness.

in context. On the other hand, *meli* 'hair' and *phyoceng* 'expression' in the following lines are annotated as 'cf' and can be (*n*)*un*-marked.

(3) elkwul-kwa cheyhyeng-**un** selun-ccum-ulo poy-ess-nuntey,
face-and figure-**NUN** 30-about-as be.seen-PST-but,
imi huyn meli-**ka** na iss-ess-ko
already white hair-**NOM** grow PROG-PST-and
phyoceng-**i** chochwey-ha-ko cichy-e iss-ess-ta.
expression-**NOM** weary-LV-and haggard PROG-PST-DECL
'Her features and figure were those of a woman of thirty,
but her hair was shot with premature grey, and her expression was weary and haggard.' <kor:#34>

In (4), there are three (*n*)*un*-marked constituents. The first (*n*)*un*-marked NP *nunglyek-un* 'ability-NUN' is labeled as an aboutness topic because it is one of the arguments of *eps* 'non-existent', whose ARG-ST is <NP(NOM), NP(DAT)>. It was annotated as 'ab'. The second (*n*)*un*-marked constituent *twiey-(n)un* 'after-NUN' is an adjunct. All (*n*)*un*-marked adjuncts in Korean are annotated as 'ct', 'cf', or sometimes just 'c' when choice is unclear. In particular, *twiey-(n)un* has an alternative pair in the immediately preceding clause (i.e. *cikum tangcang* 'right now'). In this case, it is annotated as just 'c' because two annotators analyzed it differently. Additionally, it was also tagged as 'fs' in the [TOPIC] tier. That implies that it conveys a meaning of both frame-setting topic and contrast (i.e. contrastive frame-setting topic). The third (*n*)*un*-marked constituent *sayngkak-ha-ci-nun* 'think-LV-COMP-NUN' is non-nominal. From the annotation, it is borne out that if (*n*)*un* is attached to non-nominal categories, the (*n*)*un*-marked constituents are interpreted as containing a contrastive meaning. This is in line with what Sohn (2001) claims. Because this (*n*)*un*-marked word cannot be eliminated, it is in the domain of contrastive focus, tagged as 'cf'.

(4) cikum tangcang salyeykum-ul tuli-l nunglyek-**un** eps-supni-taman,
now right reward-ACC give-REL ability-**NUN** non.existent-HON-but,
han tal-ina 6-cwuil twiey-**nun** ce-to kyelhon-ul ha-yse
one month-or 6-week after-**NUN** I-also marriage-ACC do.and
cey swuip-ul kwanli-ha-l swu isskey toy-ko kulemyen
my income-ACC control-LV-REL possibility exist become-and if.so

ceketo sensayngnim-kkeyse cey-ka unhyey-to molu-ntako sayngkak-ha-ci-**nun**
at.least sir-NOM(HON) I-NOM mercy-also not.know-COMP think-LV-COMP-**NUN**
anh-usi-l kes-i-pni-ta.
not-HON-REL thing-COP-HON-DECL
'At present it is out of my power to reward you for your services,
but in a month or six weeks I shall be married, with the control of my own income
and then at least you shall not find me ungrateful.' <kor:#56>

In brief, if (*n*)*un* is attached to adjunctive NPs or non-nominal phrases, it gives a contrastive meaning. This is also verified in the following example. There are two (*n*)*un*-marked adjuncts in (5), which were annotated as 'ct' because they are alternatives to each other.

(5) ... yengci-ka pwukccok-ulo-**nun** pekhusye-kkaci
estate-NOM north-DIR-**NUN** Berkshire-till
seccok-ulo-**nun** haymphusye-ey-kkaci ilul-ess-supni-ta.
west-DIR-**NUN** Hampshire-LOC-till extend.over-PST-HON-DECL
'... the estates extended over the borders into Berkshire in the north,
and Hampshire in the west.' <kor:#74>

At the same time, they were also analyzed as 'fs'. Hence, they are instances of contrastive frame-setting topic like the second (*n*)*un*-marked element in (4) above. These examples give me an insight with respect to frame-setting topics. They are often realized with contrast, but not always. This implies that a cross-classification between contrast and frame-setting topic is required. Accordingly, two subtypes of frame-setting topics are additionally necessary; one is contrastive frame-setting topic, and the other is non-contrastive frame-setting topic.

Focus Particles

Although only a small amount of data is annotated in the current work, quite a variety of types of focus particles appear in the text. The well-known focus particle *only* (sometimes called an exclusive operator (Beaver and Clark, 2008)) and its corresponding translations are found in all four languages. Other focus particles turn up as well, which include *even* and *also* in English, and their translations in the other three languages. What is of importance with respect to the appearance of the focus particles is that the constituents to which the particles are attached are always annotated

as ‘nf’ (new-focus) in the [FOCUS] tiers. That is, the constituents occurring with the particles (e.g. *only*, *even*, and *also*) are inomissible all the time. There is no exception to this in all four languages. That means the particles always assign a focus to the modificands.[2]

In addition to these, there are several expressions that always assign a contrastive meaning to their adjacent constituents. These expressions include *at least*, *save*, *but*, and *yet* in English and their translations in other languages. Besides, *ppwun* ‘except’, and *kkaci* ‘even’ or ‘till’ in Korean also play the same function; any constituents they are attached to are annotated as ‘cf’ in Korean. In particular, *ppwun* should occur only in negative sentences, as with *shika* ‘except’ in Japanese (Hasegawa and Koenig, 2011). This co-occurrence constraint needs to be studied in detail in future work.

Interestingly, focus picked out by *only* may or may not spread into larger phrases. In other words, the focus meaning given by *only* and its translated words in other languages (i) sometimes can be restricted to the single constituent they attach to, and (ii) sometimes that focus can be projected. Thus, some focus particles, such as *only*, can function as a trigger to invoke focus projection. From a theoretic viewpoint, Choe (2002) also gives almost the same insight in Korean. He argues that in Korean a focus particle *man* ‘only’ sometimes extends focus to the larger phrases which it belongs to. This argument is supported by my corpus annotation in Korean. This implies that focus projection can happen by means of focus particles as well as prosody. Chapter 13 deals with this matter again.

7.1.3 Topic Positioning

There are two types of topics in the [TOPIC] tiers; one is (a) ‘fs’ for frame-setting topic, and the other is (b) ‘ab’ for aboutness topic. Frame-setting topic can appear one or more times in a single clause, when the constituent (a-i) is not regarded as an argument of the main verb, (a-ii) occurs sentence-initially, and (a-iii) plays a role in restricting the spatial, temporal, or manner domain of interpretation. Aboutness topic, if it appears, can show up only once. Aboutness topic should (b-i) be one of the syntactic arguments of a main verb and (b-ii) pass the ‘tell-me-about’ test (Reinhart, 1981; Choi, 1999) and the paraphrasing test (Oshima, 2009; Roberts, 2011). For instance, in the

[2]Regarding the grammatical status of focus-sensitive operators, Beaver and Clark (2008) argue that there are two subtypes: namely conventional ones and non-conventional ones. The current work does not probe this difference, because the data seems too small to create a formal pragmatic generalization.

following example, *as to reward* in the sentence-initial position is annotated as a frame-setting topic ('fs'). In contrast the second constituent *my profession* is not annotated as aboutness-topic, because it cannot pass the *speaking of ...* test nor the other tests.[3]

(6) a. As to reward, my profession is its own reward; <eng:#63>

b. [As to reward], #speaking of [my profession], it is its own reward;

The constituent in the corresponding sentence in Russian is realized differently. In the corresponding Russian sentence, it is realized as a left-dislocated NP: *moya profyessiya* 'my profession' in (7) corresponds to a resumptive pronoun *eto* in the rest of sentence. Notably, left-dislocation constructions appear several times in Russian, but the corresponding sentences in other languages are not realized as left dislocation.

(7) Chto kasayetsya voznagrazhdyeniya, **moya** **profyessiya** –
what concerns reward **my** **profession**
eto oozhye samo po syebye voznagrazhdyeniya;
it already self in self reward <rus:#63>

In spite of the differences in the surface form, the information structure components that the left-most constituents involve are identical: *as to reward* in English and *moya profyessiya* are annotated as 'fs' in that they function to restrict the domain of what the speakers speak of. The same goes for the other languages. For example, the Korean expression corresponding to *moya profyessiya* in (7) appears with the nominative marker as presented below. In the Korean annotation of (8), the first part *poswu-ey kwanhayse-nun* is tagged as 'fs', and the remaining part is entirely tagged as 'nf-unsol'.

(8) poswu-ey kwanhayse-nun, cey il cachey-ka poswu-i-pni-ta.
reward-DAT about-NUN, my job itself-NOM reward-COP-HON-DECL. <kor:#63>

Aboutness topic sometimes occurs in a non-initial position, which means that the topic-first constraint must be language-specific. Even in Korean, there is one sentence in which the aboutness topic does not turn up in the first position; this may be a counterexample to the argument that Korean

[3]As mentioned in the previous section, they may not be analyzed in the same way because we constituents tagged as 'fs' are annotated mainly by definition of frame-setting topics (Féry and Krifka, 2008).

is a topic-first language. Because only one sentence is annotated like this, I would like to defer the decision about whether Korean is a genuinely topic-first language or not. This needs to be explored in future research by exploiting a larger running text in Korean. Suffice it to say that topics in Korean may be constrained in a less strict way with respect to sentence positioning.

(9) ..., ku ttetoli-tul-eykey **apeci-nun** elma toyci anh-nun myech
DET vagabond-PL-DAT **father-NUN** a few
eyikhe-uy ttang cwungeyse kasitempwul-i mwuseng-ha-n kos-ey
acre-GEN land among bramble-NOM overgrow-LV-REL place-LOC
chenmak-ul chi-tolok helak-ul hay-cwu-ess-supni-ta.
camp-ACC encamp-COMP permission-ACC do-give-PST-HON-DECL.
'... and he (my father) would give these vagabonds leave to encamp upon the few acres of bramble-covered land which represent the family estate, ...' <kor:#92>

7.1.4 All Focus

There are several expressions all annotators analyzed as focus (i.e. 'nf-unsol'), which include vocatives, exclamations, and some idiomatic expressions. For instance, in the following examples, the underlined expressions are separately tagged as 'nf-unsol'; an idiomatic expression *Good-morning*, a vocative expression *madam*, and exclamations *Ah* and *yes*.

(10) a. "**Good-morning**, **madam**!" said Holmes cheerily. <eng:#23>

b. "**Ah** **yes**, I recall the case; <eng:#59>

These kinds of expressions are assumed to be assigned focus. Another constructional type regarded as involving focus is a special type of questions, as exemplified below.

(11) Sherlock Holmes: You have come in by train this morning, I see."
Helen Stoner: "You know me, then?" <eng:#38–#39>

(12) Helen Stoner: She was but thirty at the time of her death, and yet her hair had already begun to whiten, even as mine has."
Sherlock Holmes: "Your sister is dead, then?" <eng:#96–#97>

Lambrecht (1996, p. 232) argues that all-focus constructions are different from other types of focused constructions in that all-focus constructions pragmatically serve to express "an absence of the

relevant presuppositions". It is my understanding that using questions like (11b) and (12b) performs such a pragmatic function in conversation. The questions in (11b) and (12b) expose the speakers' astonishment about what happened. These constructions help convey the pragmatic information that the speakers did not have any presupposition about a particular revelation until their conversation partners brought it up.

7.2 Constructions of Expressing Information Structure

In looking into the translation set of constructions licensed by special syntactic operations, such as clefts, focus/topic-fronting, and dislocation, one finding is that these constructions are translated into the other languages in different surface forms. In other words, there can be constructional differences, even though translations involve a similar (or the same) information structure. For example, a fronted construction presented in (6) is translated into a different type of construction in Russian as shown in (7). The corresponding translation in Korean is constructed in a similar way to that in English (not that in Russian), as previously shown in (8). They are repeated below for the sake of convenience.

(6) As to reward, my profession is its own reward; <eng:#63>

(7) Chto kasayetsya voznagrazhdyeniya, **moya** **profyessiya** –
what concerns reward **my** **profession**
eto oozhye samo po syebye voznagrazhdyeniya;
it already self in self reward <rus:#63>

(8) poswu-ey kwanhayse-nun, cey il cachey-ka poswu-i-pni-ta.
reward-DAT about-NUN, my job itself-NOM reward-COP-HON-DECL. <kor:#63>

There seem to be two reasons for this variation in forms of expressing information structure. First, acceptability of the literal translations is varied across the four languages. For example, a literal translation of dislocated constructions in English and Spanish to Korean sometimes sounds clumsy. Because the annotators were required to translate English sentences into the most natural and most productive form in the target language, those kinds of mismatches are completely normal in my understanding. Second, the productivity of different means for marking information structure differs in different languages. For example, clefts in English and Korean show different

distributional properties from each other as indicated in previous corpus studies (Kim, 2007; Kim and Yang, 2009). In sum, in the surface form, information structure is not expected to be articulated in a one-to-one correspondence in translation.

Clefts

There are three factors influencing cleft constructions, which this corpus study provides some evidence about.

First, the categorical choice of focused XPs in cleft constructions has been one of the topics in the study of clefts (Kim, 2007). Though the annotated text in this corpus study is not so big, several types of cleft constructions show up, in which the categories of the focused XP vary. Particularly, (13c) takes quite long constituents as the focused XP.

(13) a. It was **early in April in the year '83** that I woke one morning to find Sherlock Holmes standing, fully dressed, by the side of my bed. <eng:#7>

b. It was **from her** that I had your address. <eng:#54>

c. ... and it was **only by paying over all the money which I could gather together** that I was able to avert another public exposure. <eng:#91>

Second, *it*-clefts are sometimes translated into *wh*-clefts (a.k.a. pseudo clefts) or non-cleft constructions. For example, an *it*-cleft sentence (13b) in English (*It was from her that I had your address.*) is translated into a (near) *it*-cleft sentence in Russian, but a plain sentence in Korean. They are transliterated in (14a) and (14b), respectively. That implies that cleft constructions are not necessarily translated into the same type of constructions in other languages, as mentioned before.

(14) a. Jeto ot nejo ja uznala vash adres.
This from her I found.out your address <rus:#54>

b. ku pwun-hanthey-se sensayngnim cwuso-lul al-ass-supni-ta.
the person-DAT-from sir address-ACC know-PST-HON-DECL <kor:#54>

Third, cleft constructions are expected to exhibit an exhaustive (a.k.a. contrastive) effect (É. Kiss, 1998; Kim, 2012b). This means that the focused XPs in clefts deliver a contrastive focus meaning. Previous literature has argued that cleft constructions are used to express contrastiveness. The question is whether these XPs are always or only possibly contrastively focused. Li (2009) argues that

the *shi ... de* cleft constructions in Chinese are responsible for contrastive meanings. In a similar but slightly different vein, Gracheva (2013) argues that the ordinary cleft constructions in Russian only possibly carry a contrastive meaning. The annotated text of the current work shows that focused XPs in clefts (at least in English, Spanish, Russian, and Korean) do not always yield a contrastive meaning. If a contrast is established, the contrastive constituent must evoke an alternative set by definition. However, the cleft sentences in (13) do not entail such an alternative element in the context at all. Of course, some of them can evoke imagined contrast sets, but since the main criterion for establishing contrast in this analysis is the explicit existence of an alternative partner in the text, we cannot assert that focused XPs in clefts are always associated with contrast. My tentative conclusion is that the focused XPs in clefts may or may not be contrastive.

Focus/Topic Fronting

In the English annotation, there is only one focus/topic-fronting construction in the annotated data as presented in (15). The underlined **this** is annotated as 'ab' in English, whose translations in Spanish and Russian are also fronted and annotated as 'ab' as well. The corresponding word in Korean is dropped and annotated as nothing (i.e. an object-drop but not a topic-drop).[4] That means that it is associated with background, which can be elided as well.

(15) She had a considerable sum of money - not less than £1000 a year - and **this** **she** bequeathed to Dr. Roylott entirely while we resided with him, with a provision that a certain annual sum should be allowed to each of us in the event of our marriage. <eng:#83>

(16)

...	**emeni-nun**	wuli-ka	kyelhon-ul	ha-myen	maynyen	ilcenghan
...	**mother-NUN**	we-NOM	marriage-ACC	do-if	every.year	certain
aykswu-ka		wuli-eykey-to	tuleo-ntanun	coken-ulo,	wuli-ka	hamkkey
sum.of.money-NOM		us-DAT-also	come-REL	condition-as	we-NOM	together
sa-nun	tongan	apeci-eykey	motwu	yangto-ha-yss-supni-ta.		
live-REL	during	father-DAT	all	transfer-LV-PST-HON-DECL. <kor:#83>		

Emeni-nun 'mother-NUN' corresponding to *she* next to *this* is annotated as 'ab'. When the missing word is restored with *(n)un*, the sentence sounds rather awkward. Thus, we cannot say that the

[4]This analysis runs counter to Huang (1984)'s claim. Huang argues that object-drop happens only after the element moves forward to signal topic, which implies that object-drop is a subcase of topic-drop.

missing word carries a topic meaning in the Korean sentence. This is another example in which correspondence of information structure components is mismatched. We can say that (a) a (*n*)*un*-marked NP is not always translated into topic-marked forms in other languages, and (b) dropped elements in Korean are not always evaluated as containing a topic meaning. We cannot say that only constituents realized as topic can be dropped in Korean. These findings can be of use to machine translation; in particular, translating the untranslatable (Bond, 2005). The details are addressed in §12.2 with respect to argument optionality.

7.3 Multiclausal Constructions

One of the main purposes for this process of text annotation is to look into how information structure is configured in multiclausal utterances. As discussed in Part IV later, this dissertation takes notice of various clausal types of human languages, which previous literature has less interest in. Because this dissertation aims to create a computational model to deal with information structure in practical applications, the model should cover a variety of sentential types. The subordinate clauses that this subsection looks at include complement clauses, relative clauses, and adverbial clauses.

7.3.1 Complement Clauses

Clausal subordinates largely have to do with so-called embedded root phenomena (Heycock, 2007). Previous literature on embedded root phenomena pays attention to two factors; one is assertion, and the other is topicalized constituents in embedded clauses. Asserted clausal subordinates exhibit root effect, and thereby act like independent clauses. Since topic exhibits a root effect on syntactic operations (Büring, 1997; Portner and Yabushita, 1998; Erteschik-Shir, 1999, 2007; Bianchi and Frascarelli, 2010), it has been reported that asserted clauses can relatively freely involve a topicalized constituent within them. In short, assertion is the key to identify information structure properties of complement clauses.

Assertion, as a kind of illocutionary speech act, refers to what commits a speaker to the truth of the expressed proposition (Searle, 1976). This is naturally in line with the definition of focus (i.e. what is important and/or new in a sentence). For this reason, I assume that assertion is a sufficient condition for focus; asserted expressions presumably involve a focus interpretation in

information structure. Whether the clausal subordinates are asserted or not is largely dependent upon the lexical properties of the main predicate (Maki et al., 1999; Haegeman, 2004; Heycock, 2007; Bianchi and Frascarelli, 2010; Roberts, 2011; Lim, 2012). In other words, the whole subordinate clause is focused if it is asserted.

Previous studies have one argument in common: *Say* verbs normally render complement clauses asserted, which means topicalized NPs can occur within the complement clauses as exemplified in Korean (p. 44).[5] Heycock (2007) classifies *say* verbs with the following definition.

(17) Class A predicates (e.g. "say", "report", "be true", "be obvious"). The verbs in this group are all verbs of saying. Both the verbs and the adjectives in this group can function parenthetically, in which case the subordinate clause constitutes the main assertion of the sentence. It is claimed however that if the subordinate clause occurs in subject position (as in, e.g. "*That German beer is better than American beer is true*") it is not asserted. (Heycock, 2007, p. 189)

As the annotated text is originally a novel, quite a few sentences include *say* verbs. The analysis of complement clauses begins with the sentences with *say* verbs in order to substantiate the assumption that asserted expressions are sufficiently focused. This assumption is corroborated by several factors. First, if verbs of saying are used, the main clause does not bear any accent (invariably). Instead, all accented words are inside of the complement clauses. Since non-accented words in accent languages cannot be interpreted as containing a focus meaning (Lambrecht, 1996) and all sentences should have at least one focus, the focused part must be in the complement clause(s). Second, the complement clauses are mostly inverted within the double-quotation marks, such as ["...", verb subject]. In fact, this is a focus/topic-fronting construction. Since the remaining part (e.g. verb plus subject) does not bear any accent, the sentential form is articulated as a bipartite structure of focus and background (i.e. *focus-bg*). Finally, our annotators tried to conduct the deletion test, and learned that the main clause, when consisting of subjects and verbs of saying, can be freely elided without a meaning change. That implies that what is important and/or new in the utterances is only the complement clauses.

[5]Previous studies have claimed that (semi-)factive verbs and quasi-evidential verbs also have the same properties as *say* verbs.

7.3.2 Relative Clauses

A single constituent can sometimes depend upon two or more clauses simultaneously; in particular, relativized NPs have syntactic and semantic functions within relative clauses as well as the matrix clause. The same goes for information structure; relativized NPs can have two or more information structure relations to different clauses at the same time.

In previous literature, there are three arguments on information structure properties of relativized NPs. The first claims that the relativized NPs (or relative pronouns) have a topic feature in the relative clauses (Kuno, 1976; Bresnan and Mchombo, 1987; Jiang, 1991; Bjerre, 2011). The second claims that relative clauses show a striking similarity to constructions of expressing focus, such as clefts (Schachter, 1973; Schafer et al., 1996). That is, the former regards relativized NPs as topic, whereas the latter regards them as focus. The third claim is that it is true that relativized NPs have a tendency to be interpreted as the topic of relatives, but quite a few counterexamples exist (Ning, 1993; Huang et al., 2009). Thus, the third position means that relativized NPs do not necessarily serve as the topic of relative clauses. This corpus study probes which of these positions is most tenable from a data-based perspective. This corpus study is basically in line with the third argument, but provides a more detailed explanation of the information structure properties of relative clauses

Annotators, including me, spent quite a bit of time discussing the realization of information structure in relative clauses across four languages. The discussion led us to a distinction between restrictive relatives and non-restrictive relatives. As is well-documented by theories of grammar, restrictive and non-restrictive relatives show different behaviours in syntactic, semantic, and information structure. Especially, non-restrictive relatives exhibit nearly all root phenomena, while restrictive ones do not (Heycock, 2007). We learned that there is no one-to-one relation between head NPs in restrictive relatives and their information structure tags, while the head NPs of non-restrictive relatives are invariably annotated as aboutness topic (at least in English, Spanish, and Russian). This clear-cut distinction can be further supported by tests for topics, such as *about ...*, *what about ...*, *as for ...*, and *speaking of ...* (Oshima, 2009; Roberts, 2011). All relativized NPs in non-restrictive relatives pass the tests successfully, while those in restrictive relatives do not.

(18) a. ... he unravelled **the problems** which were submitted to him. <eng:#20>

b. I have heard of you from **Mrs. Farintosh**, whom you helped in the hour of her sore need. <eng:#53>

c. This is **my intimate friend and associate**, **Dr. Watson**, before whom you can speak as freely as before myself. <eng:#25>

(19) a. #... he unravelled the problems, and speaking of them, they were submitted to him.

b. I have heard of you from Mrs. Farintosh, and speaking of her, you helped her in the hour of her sore need.

c. This is my intimate friend and associate, and speaking of him, he is Dr. Watson, ...

The same goes for appositions as presented in (18c) and (19c). The left-hand NPs in appositions (e.g. *my intimate friend and associate*) bear the topic role with respect to the right-most NP (e.g. *Dr. Watson*). Thus, the first claim above seems true at least with respect to relativized NPs in non-restrictive relatives and apposition constructions. However, relativized NPs of restrictive relatives cannot be consistently annotated with a single tag. This means that the third claim seems true for restrictive relatives. In addition, since Korean does not employ relative pronouns and does not have two different types of relatives, relativized head NPs are differently annotated on a case by case basis depending on the context.

7.3.3 Adverbial Clauses

Adverbial clauses are annotated largely depending on the types of modification. English, Spanish, and Russian employ specific words as conjunctions, and these serve as the clue for identifying information structure relations that adverbial clauses have to the main clauses. For example, *when*, and *if* make the clause function as a frame-setter.

(20) a. There is no vehicle save a dog-cart which throws up mud in that way, and then only **when** you sit on the left-hand side of the driver. <eng:#46>

b. I shall go mad **if** it continues. <eng:#50>

According to the definition of frame-setting topic, temporal, spatial, and conditional clauses that appear sentence-initially are annotated as 'fs', following the guideline. Other adverbial clauses are not annotated at all, because there is no clue to vet the meaning and function.

Adverbial clauses also have their own information structure relations inside the matrix clause. For example, (21) is a Spanish translation of (20b). In (21), there are two dropped subjects: The

first in the matrix clause is interpreted as 'I', and the second in the adverbial clause is interpreted as 'it'. The conditional clause *si continúa* 'if continues' was entirely annotated as 'fs'. At the same time, the dropped subject which hypothetically occurs between *si* and *continúa* was tagged as 'ab', while the verb *continúa* was tagged as 'nf-unsol'. The 'ab' tag on the dropped subject was put in parentheses to differentiate it from the 'fs' tag on the whole adverbial clause.

(21) Ø Me volveré loca si Ø continúa.
(I) me go mad if (it) continues <spa:#50>

The 'fs' tag that an adverbial clause has is determined by the relation to its matrix clause. Adverbial clauses are also a kind of clauses and are capable of configuring information structure internally.

7.4 Summary

From this corpus annotation using a multilingual parallel text, I have made several interesting discoveries about properties of information structure across languages. The formal and computational model presented in this dissertation are bolstered by these findings: First, focused elements should contain at least one accented word (at least in English, Spanish, and Russian). There was no exception to this relationship in the annotated data. That is, extended foci require the appearance of a prosodically prominent word. This finding supports previous studies about focus projection. Second, in Korean, if (*n*)*un* is attached to non-nominal categories, it always invokes an interpretation of contrastiveness. This is also in line with the previous studies about the (*n*)*un*-marking system in Korean (Sohn, 2001). Third, felicitous use of (*n*)*un* shows an asymmetry depending on whether it is attached to subjects or objects. Contrastive foci in Korean are not necessarily realized with (*n*)*un*, but contrastive topics on subject are (*n*)*un*-marked. This finding also supports theoretic claims in several previous studies (Choi, 1999; Song and Bender, 2011). Fourth, some focus particles that assign focus to their head can function like a trigger of extension of focus meaning. This is relevant to the claim of Choe (2002), and reexamined in Chapter 13. Fifth, focused XPs in clefts may or may not be associated with contrast in that they sometimes show up out of the blue (i.e. without an alternative in the text). Sixth, it is borne out that *say* verbs assign focus to their complement clauses, as several previous studies claim. Seventh, in relative clauses, the head NPs have an aboutness topic relation to non-restrictive relative clauses, but not necessarily to restrictive ones.

Part IV

INFORMATION STRUCTURE IN HPSG/MRS

This dissertation represents information structure using the MRS formalism via ICONS (Individual CONStraint). The core concept of using ICONS includes three factors. First, information structure markings should be represented distinctively from information structure meanings in order to solve discrepancies between them. Second, information structure values should be underspecified by default unless there is a clue to identify a specific meaning. Third, the relationship between an expression and the clause(s) that the expression appears in conveys the essence of information structure that ICONS captures. Chapter 8 surveys previous literature that studies information structure within a theory of grammar. Chapter 9 proposes fundamentals of ICONS, and Chapter 10 applies ICONS to HPSG/MRS-based formalism. Chapter 11 delves into how ICONS operates in multiclausal utterances. Building upon the proposals, Chapter 12 enters into the details of various forms of expressing information structure. Finally, Chapter 13 addresses how focus projection is calculated within the current work.

Chapter 8

LITERATURE REVIEW

This chapter surveys previous literature based on the HPSG (Head-driven Phrase Structure Grammar, Pollard and Sag 1994), MRS (Minimal Recursion Semantics, Copestake et al. 2005), and other frameworks. First, §8.1 investigates HPSG-based studies on information structure, which are largely couched upon a pioneering study offered by Engdahl and Vallduví (1996). §8.2 looks into how several previous studies represent information structure using the MRS formalism, and how they differ from the current presentation. Most previous studies pay attention to how phonological structure interacts with information structure, §8.3 surveys related works. §8.4 offers an explanation of how other frameworks treat information structure within their formalism, and which implications they show to the current work.

8.1 Information Structure in HPSG

It is my understanding that Engdahl and Vallduví (1996) is the first endeavor to study information structure within the HPSG framework. This pioneering work has had a great effect on most HPSG-based studies of information structure. The main constraints Engdahl and Vallduví (1996) propose are conceptualized in (1) and (2). Many HPSG-based studies on information structure, irrespective of whether they use MRS, present a variant version of (1) and (2) as a means of encoding information structure. For this reason, the previous studies have several points in common in the way that they represent information structure and calculate information structure values.

(1)
$$\begin{bmatrix} \text{PHON} \mid \text{ACCENT} & \textit{accent} & & \\ \text{CONTEXT} & \begin{bmatrix} \text{C-INDICES} & [\,] & & \\ \text{BACKGROUND} & [\,] & & \\ \text{INFO-STRUCT} & \begin{bmatrix} \text{FOCUS} & \textit{sign} & \\ \text{GROUND} & \begin{bmatrix} \text{LINK} & \textit{sign} \\ \text{TAIL} & \textit{sign} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(2) a. $\boxed{1}\begin{bmatrix}\text{PHON}\,|\,\text{ACCENT} & a\\ \text{INFO-STRUCT}\,|\,\text{FOCUS} & \boxed{1}\end{bmatrix}$

b. $\boxed{1}\begin{bmatrix}\text{PHON}\,|\,\text{ACCENT} & b\\ \text{INFO-STRUCT}\,|\,\text{GROUND}\,|\,\text{LINK} & \boxed{1}\end{bmatrix}$

c. $\begin{bmatrix}\text{PHON}\,|\,\text{ACCENT} & u\end{bmatrix}$

Engdahl and Vallduví (1996) regard information structure as an interface across different layers in human language. The interface-based studies across layers can be more precisely explained within the HPSG framework, because HPSG accounts for various structural layers (e.g. phonology, morphosyntax, semantics, and pragmatics) in an interactive way. Engdahl and Vallduví pay particular attention to co-operation between phonological behaviours and contextual information. In their proposal, *accent* has three subtypes in English. They use the traditional distinction between the A and B accents (Bolinger, 1958; Jackendoff, 1972); *a* for A-accented words, *b* for B-accented ones, and *u* for unaccented ones. Thus, words in English can have one and only one of the following structures.

There is a difference between this approach and mine. Their value of features like FOCUS and LINK in (1) are whole signs, whereas I take information structure to operate over parts of the semantic representation.

In order to determine if their constraint works analogously cross-linguistically, they also analyze sentences in Catalan, in which information structure is expressed without reference to prosodic patterns. Unlike English, Catalan does not place a constraint on PHON to instantiate information structure. INFO-STRUCT in Catalan, instead, has to do with SUBCAT (SUBCATegorization) and phrasal types of daughters. Although their analysis particularly restricts attention to left/right dislocation constructions in Catalan, their approach has had a strong influence on following HPSG-based studies, including De Kuthy (2000) for German, Bildhauer (2007) for Spanish, Chang (2002) and Chung et al. (2003) for Korean, Ohtani and Matsumoto (2004) and Yoshimoto et al. (2006) for Japanese, and many others. Those previous HPSG-based studies on information structure show both similarity and difference between each other.

On one hand, these studies share a common proposal that information structure is an independent module within a grammatical framework; (i) either under SYNSEM|CONTEXT (Engdahl and Vallduví, 1996; Chang, 2002; Ohtani and Matsumoto, 2004; Yoshimoto et al., 2006; Paggio, 2009)

or (ii) outside of SYNSEM (De Kuthy, 2000; Chung et al., 2003; Bildhauer, 2007). The common notion is that information structure should be separately represented from CAT (CATegory) and CONT (CONTent). This dissertation, however, merges information structure into CONT (i.e. MRS).

On the other hand, previous studies are differentiated from each other in the values the relevant types utilize in formalizing components of information structure. In other words, it is necessary to see whether the whole sign is the value of the information-structure related features or whether that value is something semantic (i.e. MRS). The traditional means of formalizing information structure values is using coreferences between the whole sign and the value in the list of FOC(US) and TOP(IC). Engdahl and Vallduví (1996) make use of this method, and Chung et al. (2003) and Ohtani and Matsumoto (2004) utilize the same method for handling information structure in Korean and Japanese, respectively. Recently, several studies co-index something inside of MRS with a value in the list of FOCUS, TOPIC, and others. In Yoshimoto et al. (2006), Bildhauer (2007), and Sato and Tam (2012), the RELS itself has a coreference with a value in the lists of components of information structure. Paggio (2009) also makes use of MRS, but the values in the lists of components of information structure are co-indexed with the value of INDEX (e.g. *x1*, *e2*, etc.). This way of using MRS looks better and feasible. These two methods represent just two of many methods for representing information structure in HPSG and MRS. There are some other approaches. Chang (2002) represents information structure using just a string (as presented in (7) (p. 152)). Kim (2007) and Kim (2012b) take just a boolean feature as the value of FOCUS and TOPIC, and these features are under an independent structure called INFO-ST (as shown in (10) (p. 155)). Sometimes, a specific feature structure is newly introduced, and stands for logical forms (Webelhuth, 2007; De Kuthy and Meurers, 2011).

8.1.1 Sentential Forms

Engdahl and Vallduví (1996) argue that information structure is an integral part of grammar, which is similar to the basic argument of Lambrecht (1996) in that information structure is regarded as a subtype of sentential grammar.

There exist various suggestions on how information structure affects forms at the sentence level, such as topic-comment and focus-ground (i.e. bipartite structures). There are two basic components

in Engdahl and Vallduví (1996)'s proposal; focus and ground. While ground acts as an usher for focus, focus is defined in their analysis as the actual information or update potential of a sentence. Ground, consisting of link and tail, is viewed as something already subsumed by the input information state.[1] This definition implies a sentence can have a ground if and only if the informative content guarantees its use. For example, sentences with sentential focus (a.k.a. *all-focus* in Paggio (2009) and this dissertation), such as a reply to questions like *What happened?*, do not have any necessity to include ground. Since, in their analysis, ground is divided into link and tail, in line with Vallduví (1990), they make use of a tripartite structure[2] which depends on different combinations of focus, link, and tail. Building upon some extra constraints such that focus cannot precede link (i.e. linear order in instantiating information structure (e.g. link $>$ focus $>$ tail)), they propose four types of sentential forms; (i) link-focus, (ii) link-focus-tail, (iii) focus-tail, and (iv) all-focus. For example, (3A1) is a link-focus instruction, while (3A2) is a link-focus-tail instruction.

(3) Q1: So tell me about the people in the White House. Anything I should know?
A1: Yes. The **president** [$_f$ hates the Delft CHINE SET]. Don't use it.
Q2: In the Netherlands I got a big Delft china tray that matches the set in the living room. Was that a good idea?
A2: Maybe. The **president** [$_f$ HATES] the Delft chine set. (but the **first lady** LIKES it.) (Engdahl and Vallduví, 1996, p. 5)

This classification is similarly implemented as a hierarchy in Paggio (2009), though the terms are different (i.e. *topic* for link, and *bg* (background) for tail). The type hierarchy Paggio proposes for Danish is shown in Figure 8.1 (Ibid. p. 140), and the lowest subtypes are exemplified in (4) respectively.

(4) a. (Hvad lavede børnene?) [$_T$ De] [$_F$ spiste is].
(what did children.DEF) they ate icecream
'What did the children do? They ate icecream.' (*topic-focus*)

b. (Hvad spiste børnene?) [$_{BG}$ [$_T$ De] spiste] [$_F$ is].
(what ate children.DEF) they ate icecream
'What did the children eat? They ate icecream.' (*topic-focus-bg*)

---

[1]Note that ground is not the same as background. Ground is thought of as opposite to focus, while background is neither focus nor topic. Tail, so to speak, has almost the same way in realization as background.

[2]Büring (2003) suggests another tripartite structure such as topic-focus-background.

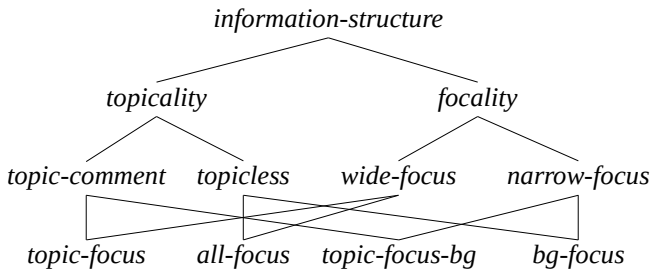


Figure 8.1: Type hierarchy of Paggio (2009)

c. (Hvem har spist isen?) [$_{BG}$ Det har] [$_{F}$ børnene].
(who has eaten icecream.DEF) that have children.DEF
'Who has eaten the icecream? The children did.' (*bg-focus*)

d. (Hvad skete der?) [$_{F}$ Børnene spiste] is].
(what happened there) children.DEF ate icecream
'What happened? The children ate icecream.' (*all-focus*) [dan] (Paggio, 2009, p. 139)

This dissertation agrees that information structure needs to be investigated as a subtype of sentential grammar (Lambrecht, 1996; Engdahl and Vallduví, 1996; Paggio, 2009). However, the type hierarchy given in Figure 8.1 is altered in this dissertation to accommodate a cross-linguistic perspective. In particular, it is necessary to delve into whether or not the hierarchy for sentential forms has to deal with the linear order of components of information structure. At first glance, *bg-focus* in Figure 8.1 might look inconsistent with the focus-tail instruction presented by Engdahl and Vallduví. As exemplified in (4c), a constituent associated with *bg* can precede other constituents associated with *focus* in Danish, which means the linear ordering constraint (i.e. link $>$ focus $>$ tail) is language-specific. The different linear orders notwithstanding, this dissertation claims that *bg-focus* in Figure 8.1 is actually the same as focus-tail. Paggio (2009) calls the identificational focus of (4c) *bg-focus* that serves to identify a referent as the missing argument of an open proposition (Lambrecht, 1996, p. 122). This is similar to cleft constructions (É. Kiss, 1998). For this reason, the type hierarchy for sentential forms in this dissertation is built up without an ordering constraint, exclusively considering which components participate in forming information structure.

8.1.2 Location within the Feature Geometry

Previous literature commonly introduces an independent typed feature structure for information structure into *sign*. The independent structure is either CXT (ConteXT) dealing with pragmatic (i.e. contextual) information or just INFO-ST; Chang (2002) employs PRA|DF|TFA in (6), Ohtani and Matsumoto (2004) uses CONX|INFO-ST, and Kim (2007) uses just INFO-ST immediately under *sign*. Similar structures are used in other papers: SYNSEM|LOC|CONTEXT|INF-ST (Yoshimoto et al., 2006), SYNSEM|IS (Bildhauer and Cook, 2010), CTXT|IS (Bjerre, 2011), INFO-STRUC (De Kuthy and Meurers, 2011), etc. The difference among them is not significant, because it is merely a matter of choosing among different formats of *sign*. The functionality of these features is almost the same in that information structure is separately represented from both morphosyntactic structure (i.e. CAT) and semantic structure (i.e. CONT).

Information structure as presented here is an independent module in grammar, however, that does not necessarily mean that information structure should be separately represented on AVMs. Unless there is a necessity to separate components of information structure from CONT(ent), the independent structure is redundant. In the formalism described here, information structure is an independent module of grammar. However, in seeking a minimal solution, is it possible to represent information structure without introducing additional structure? Partee (1991) addresses this, too: information structure is not independent of truth-conditions. If information structure is truth-conditionally relevant, it should be represented in semantics. Engdahl and Vallduví (1996), nevertheless, make use of the separate representation in the belief that information structure and logical semantics have to be represented in the grammar in a modular manner. They leave final resolution of the question to future work. My understanding is that other HPSG-based studies seem to employ the same approach, not seriously investigating the question.[3] The next chapter shows information structure can be fully represented without using CTXT or introducing an independent structure.

Representing information structure only within CONT (i.e. MRS) has another important merit in the context of multilingual machine translation. As stated earlier, this dissertation argues translation means reshaping the packaging of the information of a sentence. Thus, one of the most important considerations in representing information structure is its availability in multilingual machine

[3]One exceptional study is Webelhuth (2007), in which information structure components are dealt with under CONT.

translation as a computational model. Because all ingredients relevant to translation must be accessible in MRS within our transfer-based system (Oepen et al., 2007), information structure should be accessible in MRS.

8.1.3 Underspecification

One of the main motivations and advantages in the HPSG/MRS formalism is using underspecification. The value of a particular typed feature structure is preferentially left underspecified unless a constraint identifies the value with a more specific type. This makes grammatical operation more flexible and more economical. For example, (2c) means that unaccented words leave their information structure value underspecified, facilitating varied meanings of an unmarked expression. However, underspecification has been scarcely used in previous HPSG-based studies on information structure. Kuhn (1996) argues that using underspecification is a more effective way to represent information structure; especially, for the purpose of implementing HPSG-based NLP applications (e.g. machine translation, TTS (Text-To-Speech) systems, etc.). This dissertation relies on underspecified values of components of information structure. More specific reasons why underspecification is crucial for representation of information structure are discussed in the following subsections.

Prosody

In most HPSG-based studies of information structure, a typed feature structure for representing prosody is commonly introduced. The interface between information structure and prosody has also been studied in non-Indo-European languages such as Korean and Japanese. (5), taken from Chang (2002), and stands for a typed feature structure for prosody in Korean which has two attributes such as TC (Terminal Contour) and STR (STRess). The values of the former include falling (HL%), neutral (H%), and rising (LH%), and those of the latter stand for four levels of stress.

(5)
$$\begin{bmatrix} \textit{ppm} & \\ \text{PROS} & \begin{bmatrix} \textit{pros} & \\ \text{TC} & \langle \searrow, \rightarrow, \nearrow \rangle \\ \text{STR} & \langle 0, 1, 2, 3 \rangle \end{bmatrix} \end{bmatrix}$$

In his formalism, this structure has a correlation with another typed feature structure: namely PRA

(PRAgmatics). Information structure values, such as topics and foci, are gathered into the lists under PRA|DF|TFA as presented in (6)[4].

(6)
$$\begin{bmatrix} \textit{pra} & \\ \text{SA} & \textit{sa} \\ \text{DF} & \begin{bmatrix} \text{TFA} & \begin{bmatrix} \text{TOP} & \text{list}(\textit{phon}) \\ \text{FOC} & \text{list}(\textit{phon}) \end{bmatrix} \\ \text{POV} & \text{list}(\textit{ref}) \\ \text{CTR} & \text{list}(\textit{ref}) \end{bmatrix} \\ \text{BKG} & \textit{bkg} \end{bmatrix}$$

For example, *(n)un* and *i/ka* in Korean have one of the feature structures presented in (7a) and (7b), respectively. In (7), the PHON structure is the same as the STEM structure in `matrix.tdl` of the LinGO Grammar Matrix system. That is, the value type of PHON is just string. Despite the name, it is not directly related to any phonological information.

(7) a. i. Zero Topic
$$\begin{bmatrix} \text{STR} & \langle 0 \rangle \\ \text{PHON} & \langle \boxed{1} \rangle \\ \text{TFA} & \begin{bmatrix} \text{TOP} & \langle \boxed{1} \rangle \end{bmatrix} \end{bmatrix}$$

ii. (Thematic) Topic
$$\begin{bmatrix} \text{STR} & \langle 1 \rangle \\ \text{PHON} & \langle \boxed{1} \rangle \\ \text{TFA} & \begin{bmatrix} \text{TOP} & \langle \boxed{1} \rangle \end{bmatrix} \end{bmatrix}$$

iii. Contrastive Topic
$$\begin{bmatrix} \text{STR} & \langle 3 \rangle \\ \text{PHON} & \langle \boxed{1} \rangle \\ \text{TFA} & \begin{bmatrix} \text{TOP} & \langle \boxed{1} \rangle \end{bmatrix} \end{bmatrix}$$

b. i. (Narrow) Focus
$$\begin{bmatrix} \text{STR} & \langle 2 \rangle \\ \text{PHON} & \langle \boxed{1} \rangle \\ \text{TFA} & \begin{bmatrix} \text{FOC} & \langle \boxed{1} \rangle \end{bmatrix} \end{bmatrix}$$

ii. Contrastive Focus
$$\begin{bmatrix} \text{STR} & \langle 3 \rangle \\ \text{PHON} & \langle \boxed{1} \rangle \\ \text{TFA} & \begin{bmatrix} \text{FOC} & \langle \boxed{1} \rangle \end{bmatrix} \end{bmatrix}$$

Ohtani and Matsumoto (2004), similarly, analyzed *wa*-marked and *ga*-marked NPs in Japanese:

[4]DF stands for Discourse Function, and TFA means Topic-Focus Articulation. Additionally, SA is short for Speech Act, BKG is for BacKGround, POV is for Point-Of-View, and CTR is for CenTeR.

Wa-marked NPs are interpreted as either topic, restrictive focus or non-restrictive focus,[5] whereas *ga*-marked NPs are interpreted as either restrictive focus or all focus. In the same way as (7) in Korean, in the formalism Ohtani and Matsumoto propose *wa* and *ga* can have one of the feature structures in (8a) and (8b), respectively.

(8) a. i.

$$\boxed{1}\begin{bmatrix}\text{MORPHON} & \begin{bmatrix}\text{MORPH} & \langle X, wa\rangle\\ \text{PHON} & [\text{ACCENT}\ U]\end{bmatrix}\\ \text{INFO-ST} & [\text{LINK}\{\boxed{1}\}]\end{bmatrix}$$

ii.

$$\boxed{1}\begin{bmatrix}\text{MORPHON} & \begin{bmatrix}\text{MORPH} & \langle X, wa\rangle\\ \text{PHON} & [\text{ACCENT}\ A]\end{bmatrix}\\ \text{INFO-ST} & [\text{FOC}\{\boxed{1}\}]\end{bmatrix}$$

b. i.

$$\begin{bmatrix}\text{ACCENT} & U\\ \text{HEAD} & nom\\ \text{INFO-ST} & [\ ]\end{bmatrix}$$

ii.

$$\boxed{1}\begin{bmatrix}\text{ACCENT} & A \vee U\\ \text{MARKING} & ga\\ \text{SPEC} & \langle[\text{TOPIC}\ X]\rangle\\ \text{FOC} & \{\boxed{1}\}\end{bmatrix}$$

iii.

$$\boxed{1}\begin{bmatrix}\text{ACCENT} & A\\ \text{FOC} & \{\boxed{1}\}\end{bmatrix}$$

(7) and (8), though their formats slightly differ, are actually the Korean and Japanese variants of (2) for English. Bildhauer (2007) argues that it is rather unclear where the information about accents comes from. This criticism might be partially right when we think of the current computational environments in sentence processing. Because our applications are mostly text-based, for now it would be quite difficult to resolve the type of accents within the text domain. Nonetheless, the criticism seems rather unfair when we consider the future direction of language application. Even if there is no implementation yet that hooks up the HPSG grammar to ASR (Automatic Speech Recognition) systems with prosody extraction or TTS (Text-To-Speech) systems with prosody generation, if there is a robust correlation between information structure and prosodic accents, the grammar can deploy the level of stress for yielding higher performance on the basis of information structure. Hence, it is necessary to allow the formalism of grammar to model prosodic information using underspecifi-

[5] In Ohtani and Matsumoto (2004, p. 95), restrictive focus means wide focus. This terminology is not used in this dissertation.

cation (Kuhn, 1996). I believe that this strategy contributes to refining meaning representation via prosodic information in the long term. For this reason, classifying lexical rules by means of prosodic patterns is not a totally wrong choice.

However, there still remains another controversial point in (7) and (8). In fact, they are tantamount to introducing different lexical items of *i/ka* and (*n*)*un* in Korean, and *ga* and *wa* in Japanese. In other words, (7a) implies there could be three different morphemes (i.e. homonyms) for a zero topic (*n*)*un*, a thematic topic (*n*)*un*, and a contrastive topic (*n*)*un*. Realistically, using multiple rules for (*n*)*un* and *wa* would not be a good choice if it were not for an inevitable reason. Since Korean and Japanese very productively employ (*n*)*un* and *wa*, respectively.[6], if there are two or more *wa* and (*n*)*un* items in the dictionary of the respective grammars, it causes problematic spurious ambiguity. In other words, if we include all rules in (7a) into the lexical items, every (*n*)*un*-marked constituent produces spurious parse trees. As a result, the number of parse trees can sometimes grow too massively to handle.[7] If there is something that the multiple-entry approach captures that the single-entry approach does not, then we should use the former, because there could be a loss in information processing. Yet, as discussed hitherto, the lexical markers (e.g. (*n*)*un* and *wa*) and the prosodic patterns each contribute only partial information. In other words, neither of them can be be a decisive clue to identify which meaning of information structure is assigned to the constituent in the present work. If we can employ an acoustic system for resolving prosodic patterns in future work, the multiple-entry approach will be worthy of being used for better performance.

To sum up, my alternative way of constraining lexical markers (especially in Japanese and Korean) is as follows: First, there is one and only one entry for each marker. Second, the lexical rules include prosodic structures in principle, but they are preferentially underspecified. Third, the meaning that each marker potentially conveys is flexibly and tractably represented to cover all the partial information.

---

[6]In particular, when exploring the *Sejong* Korean Treebank, this corpus search reveals that subjects in Korean are combined with (*n*)*un* more than twice than the ordinary nominative marker *i/ka*.

[7]In fact, this is one of the major problems that cause a bottleneck in parsing and generation in the old version of KRG (a.k.a. KRG1). It had two types of (*n*)*un*; one for topic, and the other for contrast. These two (*n*)*un* sometimes had an adverse effect on system performance. Occasionally, even not a long sentence could have a large number of parse trees if (*n*)*un* occurs multiple times in the sentence. Accordingly, the sentence could not be generated in most cases because of memory overflow. For more information, see Song et al. (2010).

Ambiguity

In many previous studies across theories of grammar, so-called F(ocus)-marking is represented as a boolean feature (i.e. [FOCUS *bool*]) as proposed in Zubizarreta (1998). Handling information structure by a boolean feature is also used in other unification-based frameworks. For instance, Choi (1999), within the framework of LFG (Lexical-Functional Grammar (Bresnan 2001)), makes use of [± New] and [± Prom] as presented later in (17). Besides, other components of information structure are also marked in the same manner. These include [TOPIC *bool*], [CONTRAST *bool*], [HIGHLIGHT *bool*], and so on. For instance, Kim (2007) claims that *beer* in (9A) is constrained as (10): since *beer* in (9A) is contrastively focused (i.e. an answer to an alternative question (Gryllia, 2009)), it has both [HIGHLIGHT +][8] and [FOCUS +] in his analysis.

(9) Q: Did John drink beer or coke?
A: John drank beer. (Kim, 2007, p. 229)

(10)
$$\begin{bmatrix} \text{PHON} & \langle \text{beer} \rangle \\ \text{SYN} \mid \text{HEAD} \mid \text{POS} & \textit{noun} \\ \text{SEM} & \begin{bmatrix} \text{INDEX} & i \\ \text{RELS} & \left\langle \begin{bmatrix} \text{PRED} & \textit{beer-rel} \\ \text{ARG1} & i \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{INFO-ST} & \begin{bmatrix} \text{HIGHLIGHT} & + \\ \text{FOCUS} & + \end{bmatrix} \end{bmatrix}$$

In contrast to the approach the previous studies make use of, the present work does not use boolean features for representing information structure meaning. This is mainly because using boolean features does not allow us to represent information structure as a relationship between an entity and a clause. This dissertation represents information structure into semantic representation via ICONS (Individual CONStraints). The main motivation of ICONS is that information structure values should be filled out as a relationship with the clause an information structure-marked constituent belongs to, rather than as a property of the constituent itself. Chapter 9 provides fundamentals of ICONS in detail.

[8] [HIGHLIGHT *bool*] in Kim (2007) stands for whether or not the constituent conveys a contrastive meaning (i.e. the same as [CONTRAST *bool*] in an actual sense).

8.1.4 Marking vs. Meaning

Most of the previous studies are exclusively concerned with markings, as the name F(ocus)-marking implies. Hence, they are rather inappropriate to deal with the discrepancies between the forms expressing information structure and meanings of information structure. The lexical markers *wa* and (*n*)*un* in Japanese and Korean are typical cases showing this kind of mismatch; they can be used to express even contrastive focus. (10) is illustrative of this kind of flaw. If (*n*)*un* in Korean is used contrastively and an NP with it is focused, then the NP in Kim (2007)'s AVMs would be constrained as either (i) [HIGHLIGHT +, TOPIC +] focusing on the NP-marking system or (ii) [HIGHLIGHT +, FOCUS +] putting more weight on the meaning. Another potential constraint on the NP would be [HIGHLIGHT +, FOCUS +, TOPIC +] within his proposal, but it looks bizarre because topic and focus are mutually exclusive. Eventually, neither of them fully exhibits the mismatch.

As an alternative method, this dissertation proposes two strategies. First, information structure markings should be separately specified from information structure meanings. The former should be constrained using a morphosyntactic feature, and can be language-specific. The latter should be attributed within the semantics (i.e. under CONT), and rely on a cross-linguistically valid type hierarchy. Second, both of them have to be constrained in a flexible way. There are more than a few cases in which we cannot convincingly say which element is associated with which information structure meaning. From this point of view, it is necessary to specify information structure values as flexibly as possible. This is particularly important when creating a computational model of information structure.

8.2 Information Structure in MRS

This dissertation, not introducing another structure, represents information structure in the MRS semantic representations. There are two motivations for doing so. The first one is that information structure impacts semantic properties. As discussed previously, information structure (especially, semantic focus (Gundel, 1999)) is sometimes relevant to truth-conditions, and topic has to do with scopal interpretation (i.e., topics take the widest scope) (Büring, 1997; Portner and Yabushita, 1998; Erteschik-Shir, 1999, 2007; Bianchi and Frascarelli, 2010). Hence, it is right to incorporate information structure into the meaning representation in a direct manner. The second one is primarily

practical: The infrastructure for machine translation does MRS-based transfer (Oepen et al., 2007).

Previous HPSG-based studies can be divided into two subgroups: One represents information structure without reference to MRS (De Kuthy, 2000; Chang, 2002; Chung et al., 2003; Ohtani and Matsumoto, 2004; Webelhuth, 2007; Kim, 2007, 2012b), and the other links information structure values in the independent typed feature structure to MRS (e.g. INDEX or RELS) (Wilcock, 2005; Yoshimoto et al., 2006; Paggio, 2009; Bildhauer and Cook, 2010; Sato and Tam, 2012). The approach of this dissertation has both similarities and differences to earlier work in representing information structure in MRS.

Wilcock (2005), to my knowledge, is the first attempt to use MRS for representing information structure, modelling the scope of focus analogously to quantifier scope (i.e. HCONS).

(11) a. The president [$_f$ hates the china set].

b. 1:the(x,2), 2:president(x), 3:the(y,4), 4:china(y), 4:set(y), 5:hate(e,x,y)
TOP-HANDLE:5, LINK:1, FOCUS:3,5 (wide focus)

This is similar to the basic idea of this dissertation, in that information structure can be represented as a list of binary relations in the same way as HCONS. The difference between Wilcock's proposal and that of this dissertation is that information structure in his model is represented as handles, whereas the model of this dissertation represents the relationships between individuals and clauses as binary relations. This facilitates scaling to multiclausal constructions. For instance, (11b) taken from Wilcock (2005, p. 275) represents the wide focus reading of (11a) (i.e. from 3 to 5). Note that in this representation, LINK (*topic* in this paper) and FOCUS have no relation to the clause or its head (*hate*).

Yoshimoto et al. (2006) uses MRS, too. In their model, information structure values are unified with whole MRS predications rather than just indices. Based on this assumption, they apply the information structure values to analyzing floating quantifiers[9] in Japanese. However, their AVM does not look like a standard MRS representation, and it is rather unclear how their model can be used for practical purposes.

Paggio (2009) also models information structure referring to the MRS formalism, but the com-

[9]The relationship between floating quantifiers and focus is investigated in Kim (2011a), too. This intriguing topic needs to be more researched in the future.

ponents of information structure in Paggio's proposal are represented as a part of the context, not the semantics. Though each component under CTXT|INFOSTR involves co-indexation with individuals in MRS, her approach cannot be directly applied to the LOGON MT infrastructure which requires all transfer-related ingredients accessible in MRS (Oepen et al., 2007).

Bildhauer and Cook (2010) offer another type of MRS-based architecture: Information structure in their proposal is represented directly under SYNSEM (i.e. SYNSEM|IS) and each component (e.g. TOPIC, FOCUS) has a list of indices identified with ones that appear in EPs in RELS, which is not applicable to the LOGON infrastructure for the same reason.[10] Thus, I have the same reaction.

Amongst the various ways presented so fat, the way of this dissertation is mostly close to that of Paggio (2009) in that individuals (the value type of INDEX) are constrained for representation of information structure (i.e. Individual CONStraints). The main differences between Paggio's approach and mine are as follows: First, I put the feature whose value represents information structure inside of CONT. Second, I represent information structure values using a type hierarchy of *info-str*. Third, the feature to represent information structure involves a binary relation between individuals and clauses. The next chapter dwells on the details.

8.3 Phonological Information in HPSG

Quite a few HPSG-based studies pay attention to how phonological behaviours have an effect on structuring information in a sentence. Amongst them, this subsection briefly surveys only Bildhauer's proposal.

Though this dissertation does not give much attention to phonological constraints on information structure, it is still necessary to formalize some prosodic information in relation to information structure markings for at least two reasons. First, focus projection has been considered to be triggered by prosody. Second, as Kuhn (1996) and Traat and Bos (2004) point out, TTS (Text-To-Speech) synthesizers and automatic speech recognizers can be improved by using information structure. Thus, it is my expectation that including prosodic information in the HPSG formalism facilitates the use of HPSG-based grammars for those kinds of systems in the long term.

---

[10]This, of course, does not mean that every grammar should be compatible with the LOGON infrastructure. The ultimate goals of this dissertation is creating a computational library within the Grammar Matrix, which can be effectively used to enhance performance of HPSG/MRS-based MT systems. Given that LOGON, for now, is the readily available infrastructure for the purpose, this dissertation follows the requirements as far as possible.

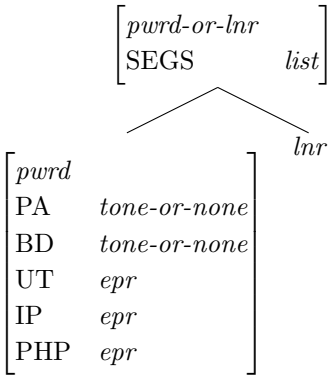
According to Bildhauer (2007)'s account, there are three HPSG-based approaches to phonology; (i) metrical tree-based approach (Klein, 2000; Haji-Abdolhosseini, 2003), (ii) grid-only approach (Bonami and Delais-Roussarie, 2006), and (iii) a hybrid approach that takes advantage of the two former approaches (Bildhauer's own). Bildhauer (2007, p. 160) says the metrical tree-based approach provides a representation of prosodic consistency, but deploys only nested structure. This has a drawback when it comes to handling intonational tunes. Bildhauer also argues that the grid-only approach of Bonami and Delais-Roussarie basically involves a flat representation, but seems too language-specific to be straightforwardly applied to other languages. The three approaches each yield their own explanation about how phonological information can be calculated within the HPSG framework in a general sense, and how the information co-operates with information structure. Another approach to the HPSG-based interface between prosody and syntax is provided in Yoshimoto (2000). Its basic assumption is that P(rosodic)-structure and C(onstituent)-structure forms a bistratal phase with each other. The bistratal approach is not considered in this dissertation for two reasons. First, Yoshimoto's proposal is not directly concerned with information structure. Second, although the interaction between prosodic and syntactic structures is examined, the analysis is rather language-specific (i.e. for Japanese) as implied by the name of the typed feature that plays the key role (e.g. MORA).

This dissertation, largely accepting the hybrid approach, keeps an eye towards being compatible with the HPSG-based formalism (Bildhauer, 2007) proposes. The account of Bildhauer can be divided into two layers. One is the PHON list as an immediate feature of *sign*, whose value hierarchy is made up of four levels; (i) prosodic word, (ii) phonological phrase, (iii) intonational phrase, and (iv) phonological utterance. The other layer is intonation, which takes charge of (v) pitch accents, and (vi) boundary tones. Building upon their operation, a schema of focus prominence rules is suggested, mainly concentrating on the top level of prosodic hierarchy (i.e. phonological utterance). Bildhauer's formalism basically develops from Klein (2000)'s proposal in which the level of syllables does not matter, and instead the prosodic hierarchy is represented by prosodic words (*pwrd*) and leaners (*lnr*). The elementary unit of PHON is *pwrd*, whose skeleton is sketched out in (12) (Bildhauer, 2007, p. 161).[11]

---

[11] In (12), the value of SEGS is a list of segments.

(12)



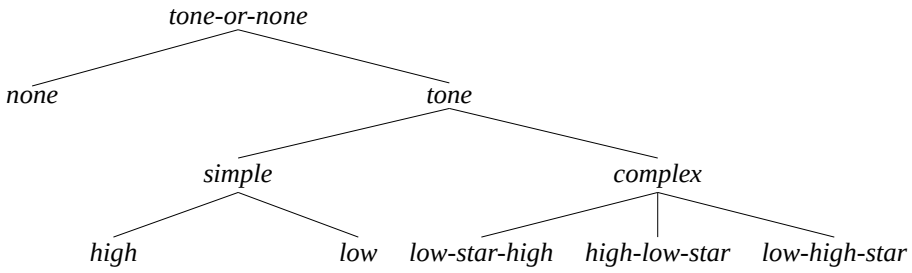
First, the lowest three features within *pwrd* in (12) represent prosodic hierarchical levels above prosodic word; PHP stands for PHonological Phrase, IP is for Intonational Phrase, and UT is the abbreviation of phonological UTterance. Each of them has *epr* meaning Edges and Prominence as its value type, whose typed feature structure is provided in (13); LE stands for Left Edge, RE for Right Edge, and most importantly DTE for Designated Terminal Element. Grounded upon the prosodic rules that Bildhauer (2007) creates, PHP, IP, and UT are defined by a relational constraint, which places a restriction on LE, RE, and DTE values of *pwrd* objects and thereby specifies the relation that a prosodic word has to higher prosodic constituents (Ibid. p. 181).

(13)

$$epr \Rightarrow \begin{bmatrix} \text{LE} & bool \\ \text{RE} & bool \\ \text{DTE} & bool \end{bmatrix}$$

Second, pitch accents (PA) and boundary tones (BD), which carry intonational information, take *tone-or-none* as their value type. Bildhauer (2007, p. 183) provides the hierarchy of *tone-or-none* in Spanish as follows, which is to be further revised for better cross-linguistic coverage in this dissertation. Each type name on the bottom line is an element in the ToBI format. For example, *high* means H, *low* means L, and *low-high-star* means L+H* (i.e. the B-accent in English (Bolinger, 1961; Jackendoff, 1972)).

(14)



Those pitch accents and boundary tones have to do with *pwrd*, whose relationship is ruled as follows (Ibid. p. 183–184). Pitch accents are attached to phonological phrases, and boundary tones are connected to intonational phrases (Steedman, 2000).

(15) a. $\left[\mathrm{PHP} \mid \mathrm{DTE} \quad +\right] \rightarrow \left[\mathrm{PA} \quad \mathit{tone}\right]$

b. $\left[\mathrm{PHP} \mid \mathrm{DTE} \quad -\right] \rightarrow \left[\mathrm{PA} \quad \mathit{none}\right]$

c. $\left[\mathrm{IP} \mid \mathrm{RE} \quad +\right] \rightarrow \left[\mathrm{BD} \quad \mathit{tone}\right]$

d. $\left[\mathrm{IP} \mid \mathrm{RE} \quad -\right] \rightarrow \left[\mathrm{BD} \quad \mathit{none}\right]$

It is my position that Bildhauer (2007)'s proposal looks convincing and cross-linguistically valid. Therefore, the type hierarchy of *tone* and the typed feature structure for phonological structure Bildhauer presents are described in `matrix.tdl` though they are not implemented in a systemic way. Although the current work is not deeply concerned with prosodic realization of information structure, they are required to be included into the system as one of the language-common structures. This is highly motivated by the necessity to refer to prosodic patterns for refinement of meaning representation in the further studies.[12]

However, specific phonological rules given in (15) are only selectively implemented in the current work. For instance, in the following chapters, two hypothetical suffixes are used for indicating the A and B accents in English for ease of processing. The rules for them are in accordance with what Bildhauer proposes. However, no other rules use phonological information herein. There are two reasons for doing so. First, for many languages, the correlation between prosody and information structure is not fully tested and thereby remains still vague. Thus, I would like to leave creating specific rules for specific languages to those who describe each grammar using the current model. Second, since the current work does not make use of any acoustic system, it is almost impossible for me to implement and test Bildhauer's phonological rules in a comprehensive way.

[12] In Chapter 10, §10.3.3 gives the details.

8.4 Information Structure in Other Frameworks

8.4.1 CCG-based Studies

The CCG (Combinatory Categorial Grammar, (Steedman 2001)) framework[13], which has a high level of detail in the analysis of the relationship between intonation and other structures (e.g. syntax, semantics, and pragmatics), has addressed information structure since the early days of the theory (Steedman, 2000). Therefore, one of the main characteristics of CCG is that it is particularly and deeply oriented toward information structure. Moreover, several CCG-based studies have accounted for how categories of information structure in CCG can be of use for practical systems from the standpoint of computational linguistics.

The components of information structure that Steedman (2000) and Traat and Bos (2004) introduce include theme (i.e. topic), and rheme, and focus. There are three structures that coincide with each other, which include (a) surface structure, (b) information structure, and (c) intonation. Among these, only (c) has significance to combinatory prosody, consisting of (c-1) pitch accents and (c-2) boundary tones. Whereas pitch accents are viewed as properties of words, boundary tones behave like a boundary between theme and rheme categories. A sequence of one or more pitch accents followed by a boundary is referred to as an intonational phrasal tune.

Pitch accents and boundary tones in CCG are mostly represented in the ToBI format as follows. There are six pitch accents to mark theme and rheme such as L+H*, L*+H for theme and H*, L*, H*+L, and H+L* for rheme. Boundary tones are what make a clear difference between Steedman's analysis and others in that he considers them to be crucial to specifying the type of phrases and thereby configuring information information structure. Intermediate phrases consist of one or more pitch accents, followed by either the L or the H boundary, also known as the phrasal tone. Intona-

[13]CCG departing from CC (Categorial Grammar) has two versions of formalism, whose history of progress is also deeply related to incorporating information structure into the formalism. The first development of CG theories is called UCG (Unification Categorial Grammar, Zeevat (1987)), which employs an HPSG-style typed feature structures (i.e. *sign*). The HPSG-style formalism facilitates more efficient co-operation of interface across grammatical layers (e.g. syntax, semantics, etc.). The second development is UCCG (Unificational Combinatory Categorial Grammar, (Traat and Bos, 2004)), which integrates CCG and UCG, and then adds DRT (Discourse Representation Theory, (Kamp and Reyle, 1993)) into the formalism, in order to facilitate a compositional analysis of information structure. Roughly speaking, those categorial grammars replace phrasal structure rules by lexical categories and general combinatory rules. In other words, the CCG framework associates syntactically potent elements with a syntactic category that identifies them as functors. There are two major rules to combine with functional categories and their arguments, which specify directionality such as (i) forward application represented as '$>$' and (ii) backward application represented as '$<$'.

tional phrase, on the other hand, consists of one or more intermediate phrases followed by an L% of H% boundary tone. Therefore, in Steedman's analysis for information structure in English, the L+H* and LH% tune is associated with the theme, and the H* L and H* LL% tunes are associated with the rheme. For instance, a surface structure *Anna married Manny.* can be analyzed as follows (Traat and Bos, 2004, p. 302), if *Anna* plays the topic role, and *Manny* plays the focus role.

(16) a. **Anna** [$_f$ married [$_f$ MANNY]].

b. Anna L+H* LH% married Manny H* LL%

In (16a), **Anna** bears the B-accent (i.e. L+H*), MANNY bears the A-accent (i.e. H*), and the focus can be projected into either the NP MANNY itself or the VP *married **Manny***. In (16b), the topic meaning that ANNA conveys comes from a pitch accent L+H* after the word, and the focus meaning that ***Manny*** delivers comes from another pitch accent H*. A boundary tone LH% forms a border of theme. Finally, *married* without any boundary tone (i.e. an invisible boundary as an edge of an unmarked theme) is included in the rheme, but it creates an ambiguous meaning with respect to the focus domain. Traat and Bos (2004) represent (16b) into the CCG-based formalism, in which three information structure values θ, ρ, and ϕ are used for theme, theme, and phrase, respectively. Those values are used as the value types of INF (INFormation structure), and additionally focus is independently represented as a boolean type.

There are several implications that CCG-based studies on information structure have for my work. First, the CCG-based studies pay particular attention to creation of a computational model for information structure. That is to say, the CCG-based studies design their model with an eye toward implementing information structure-based applications from the beginning, while many previous studies based on other frameworks do not present how their analysis can be computationally implemented. In particular, Traat and Bos (2004) argue that information structure-based computational model should be used for both parsing and generation, and conduct an experiment to verify if their model works in both sides. The information structure-based model of the present work is created in the same way. This computational model, developed in the context of grammar engineering, can be used not only for parsing human sentences into semantic representation but also for generating sentences using the representation. Second, the CCG-based studies include prosodic information in their formalism in a fine-grained way and also create linguistic rules in which prosodic information

and information structure interact with each other in a systemic way. My model does not fully use prosodic information for the reasons discussed in §8.3, but further work along the lines of this dissertation will look at how the CCG-based work systematizes the interaction between prosody and information structure. Lastly, Traat and Bos (2004) make use of prosodically annotated strings as input for their experiment, because current automatic speech recognizers do not provide enriched prosodic information. In the experiment, I employ two suffixes (e.g. '-a' for the A-accent, '-b' for the B-accent) that hypothetically represent prosodic information. Though I am not working with naturally occurring speech, the '-a' and '-b' suffixes are inspired by prosodic annotation.

8.4.2 LFG-based Studies

While most HPSG/CCG-based studies on information structure tend to lay emphasis on the interaction between phonological factors and morphosyntactic structures, previous studies based on LFG[14] tend to be more concerned with morphosyntactic operation. Discourse-related information is largely represented in LFG either within an independent structure (i.e. i-structure) (King, 1997) or just inside of f-structure (Bresnan, 2001).

It is my understanding that the first endeavor to study linguistic phenomena related to information structure within the LFG framework is offered in Bresnan and Mchombo (1987). Grammatical functions in LFG can be roughly divided into discourse functions and non-discourse functions. In their analysis, grammaticalized discourse functions such as TOP(ic) and FOC(us) are captured within f-structure.

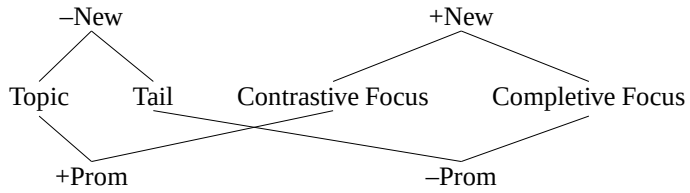
This way of putting information structure elements into f-structure, however, might be controversial, because information structure does not always coincide with grammatical functions such as OBJ(ect), COMPL(ement), and so forth (King and Zaenen, 2004). In order to overcome the potential problems, King (1997) introduces i-structure to represent how information structure units

[14]The LFG framework, as the name itself implies, has two motivations: (i) Lexical items are substantially structured, and (ii) grammatical functions (e.g. subject and object) behave importantly. LFG assumes several structural layers in the analysis of language phenomena, which include c-structure (constituent structure), and f-structure (functional structure). C-structure converts overt linear and hierarchical organization of words into phrases with non-configurationally structured information about grammatical functions, which plays a role to form f-structure. F-structure refers to abstract functional organization of the sentence, (e.g. syntactic predicate-argument structure and functional relations), which is of help to explaining universal phenomena in human language. In addition to the two basic structures, several other structures are also hypothesized such as a-structure (argument structure), s-structure (semantic structure), m-structure (morphological structure), p-structure (phonological structure), and i-structure (information structure).

(e.g. focus domain) are constructed. In other words, i-structure can be represented independently of morphosyntactic operation, thereby information structure forms and meanings can be disentangled. Several following LFG-based studies such as Choi (1999) and Man (2007) are also in line with King, indicating the necessity to solve discrepancies between c/f-structure and i-structure. While King is mainly concerned with Russian, the following studies adapt i-structure to other languages and substantiate its feasibility within the LFG framework. These include, Korean, German (Choi, 1999), and Cantonese (Man, 2007).

Another characteristic of LFG-based studies for information structure uses two types of boolean features which constrain information status such as new/given and prominent/non-prominent. This distinction is proposed in Choi (1999), who classifies (i) focus into (i-a) completive focus involving new information and (i-b) contrastive focus entailing alternatives in a set, and makes a clear-cut between (ii) topic and (iii) tail using [$\pm$ prominent]. Her cross-classification between them is sketched out in (17) (Ibid. p. 92).[15] Choi applies this classification to the representation of information structure in Korean, and Man (2007) applies it to Cantonese in almost the same way.

(17)



Those LFG-based studies have implications for this dissertation, though the underlying framework is different. First of all, Bresnan and Mchombo (1987) gives an explanation about information structure in multiclausal utterances. They delve into how topic relations in English and Chicheŵa can be captured in several types of multiclausal constructions such as embedded clauses, relative clauses, and cleft clauses. That means that it is necessary to capture an information structure relation between a subordinate clause and the main clause that the subordinate clause belongs to. Subordinate clauses constitute their own information structure within the clauses, but the relation to their main clauses additionally needs to be represented with respect to information structure. This is discussed in more detail in Chapter 9 (§9.1.3) and Chapter 11 (§11.2.1). Second, LFG-based studies deal with a variety of constructions in the study of information structure, whereas a large number of studies based on other frameworks treat only simple declarative sentences. The construction

[15] In (17), Prom is short for Prominence.

types that LFG-based studies address include interrogatives (*wh*-questions and *yes*/*no*-questions), negation, clefts (King, 1995), scrambling (Choi, 1999), and so-called topicalization (i.e. focus/topic fronting in this dissertation) (Man, 2007). Third, it is also noteworthy that LFG-based studies tend to apply their formalism directly within a specific language. Studies within other frameworks normally apply their formalisms to English first, and then project them analogously into other languages. As a consequence, the analyses tend to be rather dependent on English-like criteria. LFG-based work, on the other hand, straightforwardly looks into how a language configures information structure. LFG-based work on information structure has sometimes been criticized for not treating prosodic factors significantly, but to my understanding this is mainly because they do not start their work from English, and as we have seen, prosody is not heavily responsible for information structure markings in several languages (e.g. Chichewa, Korean, and Cantonese) unlike in English. Lastly, LFG-based studies, as mentioned before, take significant notice of the mismatches between meanings and markings of information structure and seek to reflect these discrepancies in their formalism. This dissertation has a similarity with Bresnan and Mchombo (1987) in that information structure is handled within SYNSEM, an independent structure is not needed.

8.5 Summary

Since a pioneering work of Engdahl and Vallduví (1996), information structure has received attention in quite a bit of HPSG-based research. Their main endeavor is to point out the necessity to see sentential form in relation to information structure. This motivation is also importantly applied to my work. Nevertheless, this dissertation has several differences from the previous studies. First, underspecification is not widely used in the previous studies, but this dissertation places emphasis on underspecification in representation of information structure. Second, while most previous studies do not differentiate information structure marking and information structure meaning, they are thoroughly kept distinct in this dissertation. Third, information structure is represented only under CONT(ent) (i.e. MRS) in this dissertation. Fourth, prosodic information is selectively incorporated into the formalism; the big picture is in accordance with what Bildhauer (2007) suggests, but the specific rules are not straightforwardly made use of in the current work. The following chapters address the details. In addition to them, several interesting points proposed in other frameworks

are taken into account, too. In particular, inspired by the LFG-based studies, Chapter 11 delves into information structure with special reference to various types of utterances (i.e. multiclausal constructions).

Chapter 9

INDIVIDUAL CONSTRAINTS: FUNDAMENTALS

The key notion that this dissertation suggests for representing information structure within the framework of HPSG (Pollard and Sag 1994) and MRS (Copestake et al. 2005) is to use ICONS (Individual CONStraints) along the lines of Song and Bender (2012).[1] §9.1 offers the basic necessities for using ICONS in processing information structure. §9.2, §9.3, and §9.4 propose three type hierarchies that place constraints on information structure semantically and morphosyntactically. §9.5 presents a simplified version of representation for ease of exposition.

9.1 Motivations

Using ICONS is motivated by three necessities: (i) resolving discrepancies between forms and meanings in information structure, (ii) facilitating underspecifiability for allowing flexible and partial constraints, and (iii) capturing the fact that information structure relations are between expressions and particular clauses. In addition to them, I establish one working hypothesis to facilitate (iv) informative emptiness in representing information structure.

9.1.1 Morphosyntactic Markings vs. Semantic Representation

First, the morphosyntactic markings for information structure needs to be kept distinct from semantic markings. This is analogous to the linguistic fact that morphological tense can sometimes differ from semantic tense (as exemplified before in the counterfactual constructions (p. 89)). Some forms of expressing information structure directly indicate specific information structure roles such as topic, focus, and contrast. For instance, the contrastive topic marker *thì* in Vietnamese directly

[1]The feature ICONS was originally proposed by Ann Copestake and Dan Flickinger, for the purpose of capturing semantically relevant connections between individuals which are nonetheless not well modeled as elementary predications, such as those found in intrasentential anaphora, apposition, and nonrestrictive relative clauses. Copestake and Flickinger suggested (p.c.) that the same mechanism can be used to anchor information structure constraints to particular clauses. In a more general system that uses ICONS, the value of ICONS would be a list of items of type *icons*, where *info-str* is a subtype of *icons*.

assigns contrastive topic meaning to the NP that the marker is attached to, as repeatedly exemplified below.

(1) Nam thì đi Hà Nội
Nam THI go Ha Noi
'Nam goes to Hanoi(, but nobody else).' [vie] (Nguyen, 2006, p. 1)

A specific sentence position can also play the same role. For example, if the word order is not neutral in Russian, the clause-final position assigns the non-contrastive focus meaning, while preposing is responsible for contrastive focus meaning (Neeleman and Titov, 2009). Yet, quite a few marking systems do not necessarily reveal which information structure meanings are conveyed. The typical case that involves the discrepancies between morphosyntactic markings and semantic representations is the information structure marker *wa* in Japanese and (*n*)*un* in Korean as discussed before. Even when a language has a relatively deterministic relation between forms and meanings, the forms do not always match with the information structure meanings. For example, the A-accent in English has been widely evaluated as containing focus meaning, but there are some counterexamples to this generalization about the relationship between the A-accent and focus as exemplified previously in §4.2.2 (i.e. Second Occurrence Focus (p. 65)). Moreover, there has been a debate on the function of the B-accent, which include (i) just topic (Jackendoff, 1972), (ii) contrastive topic (Kadmon, 2001; Büring, 2003), (iii) theme (Steedman, 2000), and (iv) contrast (Hedberg, 2006).

9.1.2 Underspecification

Second, unless there exists a decisive clue to identify the information structure meaning, the meaning is most parsimoniously represented as underspecified. This proposal is especially crucial for analyzing sentences which appear in an unmarked word order. Without clues to indicate a particular meaning (e.g. the contrastive topic marker *thì* in Vietnamese), any constituents in the unmarked order would are not specified for any meaning with respect to information structure. For instance, (2a) presented again below is in the neutral word order in Russian, and the orthography does not represent prosodic patterns related to information structure.

(2) a. Sobaka laet.
dog bark
'The dog barks.' [rus]

b. Laet sobaka.
bark dog
'The DOG barks.' [rus]

Thus, we do not always know which element plays which information structure role in text-based processing. For those kinds of relations, it would be better to leave the information structure values underspecified in order for them to cover all meanings that the constituents may potentially have. On the other hand, we can say the *sobaka* has focus meaning in (2b) in which the subject is not *in situ*, because the inversion behaves as the clue to determine focus.

As exemplified hitherto, we often cannot identify which element bears which information structure meaning without referring to contextual information. In particular, given that sentence-by-sentence processing usually lacks discourse-related information, it is not likely that we can precisely determine an information structure role of each constituent in many cases. One constituent, even though it is realized in a specific form, could have ambiguous meanings if it were not for contextual information. Hence, it is highly necessary to represent the information structure meanings in a flexible way. For instance, note the example in Greek ((p. 97)), which is provided again below for the sake of convenience.[2]

(3) a. Thelo kafe.
want.1SG coffee.ACC
'I would like coffee.' [ell]

b. Kafe thelo.
coffee.ACC want.1SG
'Coffee I would like.' [ell] (Gryllia, 2009, p. 44)

Because the postverbal focus *kafe* 'coffee' in (3a) takes the object position in the basic word order and there is no other clue to disclose the information structure role within the single sentence, it does not have to have any specific meanings *per se*. That means that *kafe* in (3a) can be evaluated as containing (i) non-contrastive focus, (ii) contrastive focus, or even (iii) background if the preceding verb *thelo* 'want' plays a focus role. Hence, the semantic representation of *kafe* in (3a) has to cover all those three meanings simultaneously (i.e. *non-topic* in this dissertation). On the other hand, the

[2]The subscript in the original example such as $[]_{\text{C-Foc}}$, which stands for contrastive focus, is removed in (3) in order to show the difference between the neutral sentence and the marked sentence.

preverbal focus in (3b) presents a clue to identify its information structure meaning in that it is not *in situ*. In other words, *kafe* in (3b) is constructionally marked and thereby conveys a more specific meaning than that in (3a) (i.e., it has no interpretation of background). Nonetheless, its meaning is still vague; it can be read as either non-contrastive focus or contrastive focus. Thus, the ideal representation would be able to include both meanings at the same time (i.e. *focus* as the supertype of both *semantic-focus* and *contrast-focus*).

9.1.3 Binary Relations

Third, using ICONS is motivated by the necessity of finding binary relations between a clause and an element (i.e. *individual* that ERG (English Resource Grammar, Flickinger 2000) uses in the construction of MRSs) that belongs to the clause. These binary relations are crucial in representing the information structure of various types of utterances. The typed feature structure of ICONS consists of three components to identify which element has which information structure value within which clause.

Information structure roles, including focus, topic, and contrast, can be represented not as a property of the constituent itself, but as a relationship that holds between the clause and the constituent it belongs to. For example, in the English sentence *The* DOG *barks.*, the subject *the* DOG with the A-accent should be viewed as the the focus of the clause headed by the predicate *barks*, rather than as simply focus. This approach is along line with Lambrecht (1996) and Engdahl and Vallduví (1996) who regard information structure as a subtype of sentential grammar. That is, whether a constituent is associated with focus or topic should be identified within the sentence that includes the constituent.

Furthermore, a constituent can have multiple relations with different clauses. As the corpus study in Part III showed, one element can have two (or more) information structure relations, if it belongs to different clauses simultaneously. The relations may or may not be the same with each other. This notion can be clearly understood if we consider multiclausal utterances such as those which contain relative and embedded clauses. Most previous studies on information structure treat only fairly simple and monoclausal constructions, which is naturally the basic step for modeling how information is packaged in a language. However, embedded clauses present another type of

construction, in which one element can have different information structure meanings. An embedded clause not only configures its own information structure, but also plays an information structure role in the domain of the main clause that takes the embedded clause as one of the arguments. The typical cases in which a single element has multiple relations at the same time with different predicates are relative clauses: The antecedent of relative clauses has relations with both (i) the verb in the relative clause and (ii) the other verb in main clause, whose values are not necessarily identical to each other.

Those kinds of relations have been already captured in an LFG-based study on information structure (Bresnan and Mchombo, 1987): They argue that relative pronouns function as the topic of the relative clauses in line with the theorem presented in Kuno (1976).[3] Thus, in their analyses the relative pronouns are assigned an information structure value within the relative clause, as shown in (4).

(4) a. The car [which you don't want ___] is a Renault.
TOPIC OBJ

b. I know [what you want ___].
FOCUS OBJ

c. [It is my car [that you don't want ___]]. (Bresnan and Mchombo, 1987, p. 757–758)
FOCUS TOPIC OBJ

The antecedent corresponding to the relative pronoun (e.g. *the car* in (4a)) has another information structure value within the main clause. Embedded constructions realized as free relative clauses (e.g. *what you want* in (4b)) play another information structure role within the main clause. The clefted NP (e.g. *my car* in (4c)) is assigned a focus meaning, but its relative pronoun (e.g. *that* in (4c)) plays the topic role in the relatives. Those analyses do not accord perfectly to the argument presented in this dissertation,[4] but they are still significant. They highlight the necessity to treat information structure as a matter of the relationship between an element and its clause.

---

[3]This dissertation does not entirely defer to this argument. §11.2 presents the details.

[4]I disagree with the idea that relative pronouns and relativized NPs serve as the topic of the relative clauses. Instead, I argue that they have to be evaluated as *focus-or-topic*. §11.2 dwells on the details.

9.1.4 Informative Emptiness

In addition to the motivations presented in the previous subsections, I provide one working hypothesis about informatively empty categories. Lambrecht (1996, p. 156) argues that expressions which cannot be stressed, such as expletives (e.g. *it* in *It is raining.* and *there* in *There is nobody in the room.*), unstressed determiners, and so on, cannot be used as topic in principle. What is to be noted is that they cannot be used for expressing any other information structure meanings, either. For this reason, this dissertation makes a working hypothesis that semantically empty categories (e.g. complementizers, expletives) and syncategorematic items[5] (e.g. relative pronouns) are informatively empty as well. That means no information structure category can be assigned to them, though they may be required by constructions which serve to mark information structure, such as the cleft construction in English. For example, in (5a), the expletive *it* and the copula *is* are semantically empty and the relative pronoun *that* is syncategorematic; thus, they are informatively vacuous. Likewise, since the copula *was* and the preposition *by* in passive sentences in English are semantically void, they cannot take part in information structure, as shown in (5b). ~~Strike~~ in (5) indicates that they are informatively meaningless.

(5) a. ~~It~~ ~~is~~ the book ~~that~~ was torn by Kim.

b. The book ~~was~~ torn ~~by~~ Kim.

Lexical markers to express information structure, such as case-marking adpositions (e.g. nominative *ga* in Japanese) are mostly semantically and informatively empty. Although they participate in forming information structure and behave as the clue to identify information structure meanings, they do not have their own predicate names, and do not exist in the semantic representation (i.e. MRS as presented here), either. In other words, they assign no information structure values to themselves, but instead perform a function to identify and assign the information structure values to the phrase that they are combined with. Since the information structure constraints in the representation of the current work are all relative to elements in the RELS list, what is not represented in the RELS list cannot bear any information structure value. In sum, semantically empty lexical items and

[5]Syncategorematic items refer to words that cannot serve as the subject or the predicate of a proposition. Lambrecht (1996) does not capture any generalization about them, but I argue that they cannot be used as topic, either.

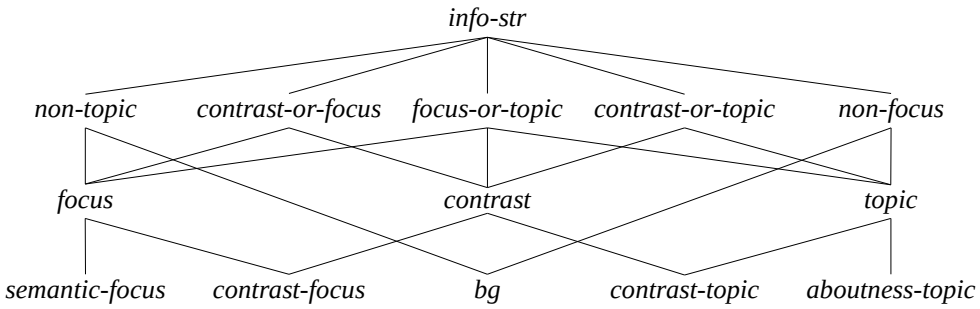


Figure 9.1: Type hierarchy of *Info-str*

syncategorematic items are incapable of bearing their own information structure value, but they can assign an information structure value to others.

9.1.5 Summary

The motivations and the working hypothesis presented in this section are rigorously applied within the remaining parts of this dissertation. They can be summarized as follows.

(6) a. The formal markings of information structure should be modelled separately from the semantic representation of information structure.

b. The information structure value should be specified so that it can cover all potential information structures that the sentence may have.

c. The semantic representation of information structure involves a binary relation that shows which element has which information structure relation to which clause.

d. Semantically empty and syncategorematic items are informatively empty.

These hypotheses are built upon in the following chapters into three type hierarchies: *info-str*, *mkg*, and *sform*.

9.2 Information Structure (info-str)

The type hierarchy of *info-str* is sketched out in Figure 9.1. The values of information structure are represented as node names (i.e. type names) within the *info-str* type hierarchy. For instance,

if a linguistic unit introducing an EP (Elementary Predicate) into RELS is computed as conveying meaning of non-contrastive focus (i.e. *semantic-focus*), it also introduces one *info-str* value whose type name is *semantic-focus* into ICONS. The nodes in the bottom line represent the most specific meaning, which cannot be further subdivided with respect to information structure. The nodes in the third line include the major components of information structure. *Focus* and *topic* are mutually exclusive, and *contrast* should be realized with either of them. The nodes in the second line are abstract. Each of them stands for a linguistic property that the major components of information structure exhibit: possibility of topicality or focality (*non-topic*, *non-focus*, and *focus-or-topic*), and possibility of contrastiveness (*contrast-or-topic* and *contrast-or-topic*). These are motivated by the need to capture via underspecification exactly the range of information structure meanings associated to particular information structure markings in certain languages, as detailed below.

This *info-str* hierarchy is based on Song and Bender (2011), but is extended with several additional nodes. *Non-topic* means the target cannot be read as topic (e.g. case-marked NPs in Japanese). *Focus-or-topic* is assigned to the fronted NPs in focus/topic fronting constructions. *Contrast-or-topic* is used for *wa* in Japanese and (*n*)*un* in Korean, because *wa* or (*n*)*un*-marked constituents in those languages can convey a meaning of non-contrastive topic, contrastive topic, or even contrastive focus. *Contrast-or-focus* likewise can be used for forms responsible for a meaning of non-contrastive focus, contrastive focus, or even contrastive topic.[6] *Non-focus* similarly indicates that the target cannot be the focus, and would be appropriate for dropped elements in *pro*-drop languages. As discussed thus far, *focus* and *topic* are mutually exclusive because they designate disjoint portions of a sentence. *Focus*, *contrast*, and *topic* multiply inherit from the components in the second row. The types in the bottom line represent the fully specified meaning of each component of information structure. *Semantic-focus* taken from Gundel (1999) means non-contrastive focus, and *aboutness-topic* means non-contrastive topic. Finally, *bg* (background) means the constituent is neither *focus* nor *topic*, which typically does not involve additional marking but may be forced by particular positions in a sentence.

Comparing to the previous version presented in Song and Bender (2011) and other approaches in previous literature, the type hierarchy illustrated in Figure 9.1 allows greater flexibility. First, Fig-

[6]Such a marking system has not been observed, but it is included into the hierarchy as a counterpart of *contrast-or-topic*.

ure 9.1 shows us that *contrast*, which is in a sister relation to *non-topic* and *non-focus*, behaves independently of *topic* and *focus*. Second, *focus-or-topic* and *contrast-or-topic* can help in the modelling of the discrepancies between forms and meanings in information structure (e.g. focus/topic-fronting, *wa* or (*n*)*un*-marked focus in Japanese and Korean, etc.), and represent ambiguous meanings involving a classification across *focus*, *topic*, and *contrast*. Third, *non-topic* and *non-focus* also facilitate more flexible representation for informatively undetermined items in some languages. For example, case-marked NPs can convey either focus or background meaning of in Japanese (Heycock, 1994). That is, since a Japanese case marker (i.e. *ga* for nominatives) can convey two information structure meanings (*focus* and *bg*), the marker itself has to be less specifically represented as *non-topic*. Note that *non-topic* is the supertype of both *focus* and *bg*. Finally, *bg* is made use of as one of the components of information structure.

Using ICONS involves several fundamental points in operation as follows. First, ICONS represents information structure as a binary relation between two elements. In other words, this dissertation regards *clause* as the locus where information structure is determined.[7] Second, ICONS behaves analogously to HCONS and RELS in that values of *info-str* are gathered up from daughters to mother up the tree. The value type of ICONS, HCONS, and RELS is *diff-list*, which incrementally collects linguistic information during the formation of parse trees. Additionally, ICONS and HCONS share almost the same format of feature structure. Both are, so to speak, accumulator lists. The value type in the *diff-list* of ICONS is *info-str*, and that of HCONS is *qeq*, both of these include two attributes to represent a binary relation (i.e. TARGET to CLAUSE, and HARG to LARG). Third, despite the similarity in structure, RELS and HCONS make a difference from ICONS in terms of how to function in the semantics.[8] RELS and HCONS directly engage in the building up of the logical form, and also interact in an intimate manner with each other. ICONS does not function in exactly the same way though information structure has to do with truth-conditions (Partee, 1991). Fourth, HCONS and ICONS behave differently also in generation. ICONS-based sentence generation is carried out via subsumption check, using the type hierarchy whose value type is *icons*

[7] [CLAUSE *individual*] and [CLAUSE-KEY *event*] at first blush might look like an inconsistency. However, *event* is a subtype of *individual* in the current type hierarchy of the LinGO Grammar Matrix system. Roughly speaking, *individual* (an immediate subtype of *index*) is the lowest meaningful supertype of *ref-ind* for nominals and *event* for verbals.

[8] Ann Copestake indicated this at the 9th DELPH-IN Summit held in Saarland.

or its subtypes (e.g. *info-str*). That is, the generator first creates all potential sentences that logically fit in the input MRS without considering the constraints on ICONS, and then postprocesses the intermediate results to filter out sentences mismatching the values in the ICONS list. Chapter 15 deals with the details of ICONS-based generation.

9.2.1 ICONS

ICONS is newly added to structures of type *mrs* (i.e. under CONT) as shown in (7).

(7)
$$\begin{bmatrix} \textit{mrs} & \\ \text{HOOK} & \begin{bmatrix} \textit{hook} & \\ \text{GTOP} & \textit{handle} \\ \text{LTOP} & \textit{handle} \\ \text{INDEX} & \textit{individual} \\ \text{XARG} & \textit{individual} \\ \text{ICONS-KEY} & \textit{info-str} \\ \text{CLAUSE-KEY} & \textit{event} \end{bmatrix} \\ \text{RELS} & \textit{diff-list} \\ \text{HCONS} & \textit{diff-list} \\ \text{ICONS} & \left\langle ! \ ..., \begin{bmatrix} \textit{info-str} & \\ \text{CLAUSE} & \textit{individual} \\ \text{TARGET} & \textit{individual} \end{bmatrix}, ... \ ! \right\rangle \end{bmatrix}$$

An ICONS element has two features: namely TARGET and CLAUSE. When an element is information-structure marked and also is exhibited as an EP, the element is represented as a value of TARGET, which has a coreference with the ARG0 of the EP (a.k.a. distinguished variable). That is to say, each type name indicates which information structure meaning is associated with the EP, and the connection between them is specified by the co-index between TARGET and ARG0. On the other hand, the clause that the element is dependent upon is represented as a value of CLAUSE, which also has a coreference with the INDEX of the predicate that functions as the semantic head of the clause.

To take a simple example in advance, (8a) can be represented as the following AVM (8b). Note that in (8a) the subject ***Kim*** is B-accented and the object *the* BOOK is A-accented.

(8) a. **Kim** reads the BOOK.

b.
$$\begin{bmatrix} \textit{mrs} \\ \text{LTOP} \quad \boxed{h1} \\ \text{INDEX} \quad \boxed{e2} \\ \text{RELS} \quad \left\langle \begin{bmatrix} \textit{proper\_q\_rel} \\ \text{LBL} \quad \boxed{h3} \\ \text{ARG0} \quad \boxed{x5} \\ \text{RSTR} \quad \boxed{h4} \\ \text{BODY} \quad \boxed{h6} \end{bmatrix}, \begin{bmatrix} \textit{named\_rel} \\ \text{LBL} \quad \boxed{h7} \\ \text{ARG0} \quad \boxed{x5} \\ \text{CARG} \quad \textit{kim} \end{bmatrix}, \begin{bmatrix} \textit{\_read\_v\_rel} \\ \text{LBL} \quad \boxed{h8} \\ \text{ARG0} \quad \boxed{e2} \\ \text{ARG1} \quad \boxed{x4} \\ \text{ARG2} \quad \boxed{x9} \end{bmatrix}, \begin{bmatrix} \textit{exist\_q\_rel} \\ \text{LBL} \quad \boxed{h10} \\ \text{ARG0} \quad \boxed{x9} \\ \text{RSTR} \quad \boxed{h11} \\ \text{BODY} \quad \boxed{h12} \end{bmatrix}, \begin{bmatrix} \textit{\_book\_n\_rel} \\ \text{LBL} \quad \boxed{h13} \\ \text{ARG0} \quad \boxed{x9} \end{bmatrix} \right\rangle \\ \text{HCONS} \quad \left\langle \begin{bmatrix} \textit{qeq} \\ \text{HARG} \quad \boxed{h4} \\ \text{LARG} \quad \boxed{h7} \end{bmatrix}, \begin{bmatrix} \textit{qeq} \\ \text{HARG} \quad \boxed{h11} \\ \text{LARG} \quad \boxed{h13} \end{bmatrix} \right\rangle \\ \text{ICONS} \quad \left\langle \begin{bmatrix} \textit{contrast-or-topic} \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x5} \end{bmatrix}, \begin{bmatrix} \textit{semantic-focus} \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x9} \end{bmatrix} \right\rangle \end{bmatrix}$$

In (8b), the first element in ICONS is specified as *contrast-or-topic*, which stands for the information structure meaning that ***Kim*** (potentially) delivers. Likewise, the second element in ICONS indicates that *the* BOOK is evaluated as containing *semantic-focus*. The connection between the elements in ICONS and the EPs in RELS is determined by the coreference between TARGET of each ICONS element and ARG0 of EP(s). The first element in ICONS has $\boxed{x5}$ for TARGET, and the first and the second EPs in RELS have the same value. Likewise, the TARGET of the second element in ICONS is co-indexed with the fourth and the fifth EPs' ARG0. The values of CLAUSE indicate which EP is the head in the clause. In this case, the verb *reads* plays the role as indicated by $\boxed{e2}$. The clues to determine information structure meanings are built up incrementally by lexical and phrasal rules with an interaction of the type hierarchies. In this case, the rules for identifying each information structure value are (hypothetical) lexical rules that constraining the A and B accents. When a specific *info-str* value is created by such a rule, this value is gathered up to the tree via *diff-list* (p. 20).

What is of importance in this way of representation is that the intermediate types in this hierarchy allow for underspecified representations. As discussed several times thus far, the grammar of many

human languages does not fully pin down the information structure role an element plays. but does provide partial information about it. For example, *contrast-or-topic* on the first ICONS value is not a terminal node in Figure 9.1, which means ***Kim*** in (8a) can be interpreted as one of its subtypes; such as either *contrast-focus*, *contrastive-topic*, or *aboutness-topic*.[9] This flexible representation is crucial in a computational model of handling natural language sentences.

9.2.2 ICONS-KEY and CLAUSE-KEY

In (7), there are two pointers under HOOK. They are required in the compositional construction of the ICONS list in an incremental way.

On one hand, ICONS-KEY makes both the phrase/ lexical structure rules and the lexical entries contribute partial information to the same ICONS element. When an *info-str* element can be inserted into the ICONS list, we may not specifically know which information structure meaning the element carries because information structure markings often provide only partial information. The meaning can be more constrained by multiple different sources when the parse tree is further constructed. For example, *wa* in Japanese in itself is assigned *contrast-or-topic*, but this meaning can be more specified later (e.g. *topic*, *contrast-topic*, and *contrast-focus*) by other syntactic operations such as scrambling. Thus, it is necessary to use a pointer in order to impose a more specific constraint on an *info-str* element already augmented in the ICONS list. ICONS-KEY is used for this purpose.

On the other hand, the value of CLAUSE cannot be identified until which clause the constituent belongs to is identified. Thus, when an *info-str* element is inserted into the ICONS list, the value of CLAUSE is not specified yet in most cases. This value should be filled in later by using another pointer called CLAUSE-KEY. Each ICONS-KEY|CLAUSE is not lexically bound. The CLAUSEs are naturally identified at the clausal level. In other words, the CLAUSE values have to remain unbound until each clause an individual is overtly expressed in is chosen.[10] There are two assumptions to be noted. The first is that individuals play an information structure role only with respect to overt

[9]The specific choice amongst them can be determined by the contextual information.

[10]This strategy is different from the approach presented in Song and Bender (2012), in which *verbal-lex* and *headed-icons-phrase* take the responsibility of linking CLAUSE-KEY to the INDEX of heads. The main reason of the change in strategy is that using *headed-icons-phrase* ends up with introducing too many subtypes of *head-comp-phrase*. This runs against the spirit of the HPSG formalism (i.e. reducing redundancy and using a minimal number of grammatical types).

clauses. That is, if an utterance contains no items that can play a role of the semantic head, the utterance is assumed to have no CLAUSE binding.[11] The second is that clauses in this context do not include non-finite (i.e. tenseless) clause. That is, whether or not a verbal type has a clausal dependent (subject or complement) is dependent upon whether or not the dependent involves a verbs in which a tense is identified. For example, underlined VPs in the following are not clausal arguments. In other words, the number of clauses in an utterance is the same as the number of tensed VPs in the utterance.

(9) a. Kim seems to sleep.

b. Kim tried to sleep.

c. Kim saw Fido sleeping.

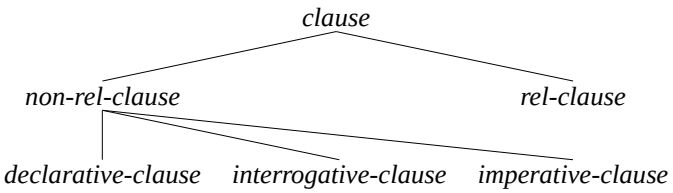
d. Kim made Fido sleep.

e. Kim promised Lee to leave.

f. Kim believed Lee to have left.

The framework of the LinGO Grammar Matrix employs a type hierarchy representing clausal types, as sketched out in (10). The *clause* hierarchy is already implemented in the core of the LinGO Grammar Matrix system (i.e. `matrix.tdl`).

(10)



Among the nodes in (10), *non-rel-clause* and *rel-clause* are responsible for constraining the CLAUSE values. The values of CLAUSEs that elements in ICONS have are co-indexed with the INDEX of the semantic head of the CLAUSE (i.e. the INDEX being the ARG0 of some EP whose LBL is LTOP). This constraint on *non-rel-clause* is represented in (11), in which CLAUSE-KEY has a coreference with its INDEX.

[11] There are some utterances in which no verbal item is used in human language. First, if an utterance is vocative (e.g. *Madam!*), the information structure value of the entire utterance can be evaluated as *focus*, following the findings obtained from the corpus study (§7.1.4). Second, in languages that do not make use of copula (e.g. Russian) copula constructions include non-verbal predicates. In this case, since the complement plays the semantic head role, the value of CLAUSE is bound to the complement.

(11)
$$\begin{bmatrix} \textit{non-rel-clause} \\ \text{NON-LOCAL} \mid \text{REL} \;\; \textit{0-dlist} \\ \text{HD} \begin{bmatrix} \text{HOOK} & \begin{bmatrix} \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} \mid \text{CLAUSE} & \boxed{1} \\ \text{CLAUSE-KEY} & \boxed{1} \end{bmatrix} \\ \text{NON-LOCAL} & \begin{bmatrix} \text{QUE} & \textit{0-dlist} \\ \text{REL} & \textit{0-dlist} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Because every element in a single clause shares the same CLAUSE-KEY, this coreference is also applied all information structure values' ICONS|CLAUSE in ICONS.[12] For instance, lexical types that involve an intransitive ARG-ST (e.g. an intransitive verb *bark* in English) inherit from the AVM (12). The CLAUSE-KEY of the subject has a coreference with the verb's CLAUSE, but the specific value is not yet identified.

(12)
$$\begin{bmatrix} \textit{intransitive-lex-item} \\ \text{LKEYS} \mid \text{KEYREL} \mid \text{ARG1} \; \boxed{1} \\ \text{HOOK} \mid \text{CLAUSE-KEY} \quad \boxed{2} \\ \text{ARG-ST} \left\langle \begin{bmatrix} \text{HOOK} \mid \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY.CLAUSE} & \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix}$$

The CLAUSE values (not yet specified) of the elements in the ICONS list are specified when a clause is constructed by (11). The same goes for adjuncts in a single clause. Adjuncts (e.g. attributive adjectives, adverbs, etc.) and the heads they are modifying share the same value of CLAUSE. That is, the ICONS-KEY|CLAUSE and CLAUSE-KEY of NON-HEAD-DTR has a coreference with the ICONS-KEY|CLAUSE of HEAD-DTR. For more information about this is given in §10.2 (p. 210).

In `matrix.tdl`, the subtypes of *head-subj-phrase* also inherit from the types at the bottom in (10) (e.g. *declarative-clause*, etc.). Hence, the instance types (e.g. *decl-head-subj-phrase* and *decl-head-opt-subj-phrase*) naturally bear the constraint (11). In other words, instances of *head-subj-phrase* are responsible for the binding of CLAUSE-KEY.

---

[12] Petya Osenova (p.c.) says that the constraint on *non-rel-clause* shown in (11) is incompatible with some interrogative sentences in Bulgarian. What she points out is that *0-list* in NON-LOCAL|QUE can cause a problem in that Bulgarian employs multiple *wh*-fronting (Grewendorf, 2001). Nonetheless, the constraint on *non-rel-clause* is presented as is, because the current proposal focuses on information structure in the LinGO Grammar Matrix. The *wh*-fronting is beyond the scope of the present work, but should be addressed in future work.

9.2.3 Summary

As discussed thus far, in order to see the entire picture of how information is packaged in an utterance, it is necessary to look at (i) which element has (ii) which information structure relation to (iii) which clause. In particular, if an utterance is made up of two or more clauses, an entity can have an information structure relation (e.g topic, focus, and so on) with each clause, and the relations are not necessarily the same. Using a binary relation meets this need; TARGET for (i), CLAUSE for (iii), and a value of *info-str* (i.e. a node in the type hierarchy) for (ii). The items on the ICONS list are feature structures of type *info-str*, which indicate which index (the value of TARGET) has a property of information structure and with respect to which clause (the value of CLAUSE). As represented by the event variable associated with the head of the clause, an information structure meaning that each individual conveys will be represented in MRS as a value of ICONS, which our infrastructure of machine translation can refer to at the stages of transfer and generation.

***9.3 Markings* (*mkg*)**

The information structure marking itself is recorded via a morphosyntactic feature MKG (MarKinG) inside of SYNSEM|CAT, which places lexical and syntactic constraints on forms expressing information structure meanings. MKG features are exclusively concerned with markings of information structure.

MKG plays two roles in handling information structure; one is theoretically driven, and the other is practical. First, MKG contributes to resolving discrepancies between form and meaning in information structure. As mentioned earlier, the MKG value reflects the morphosyntactic marking, but does not necessarily coincide with the semantic value. For instance, *wa* in Japanese and (*n*)*un* in Korean (as discussed in §5.1) can sometimes convey a contrastive focus reading as exemplified in (13): *ecey-nun* 'yesterday-NUN' in the answer should be evaluated as conveying a meaning of contrastive focus. In this case, the value of MKG that *ecey-nun* has (under CAT) is *tp*, but the information structure value in semantic representation is *contrast-focus*.

(13) Q: Kim-i onul o-ass-ni?
Kim-NOM today come-PST-INT
'Did Kim come today?' [kor]

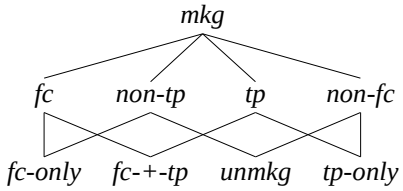


Figure 9.2: Type hierarchy of *Mkg*

A: ani. (Kim-un) ecey-nun o-ass-e.
No. Kim-NUN yesterday-NUN come-PST-DECL
'No. Kim came yesterday.' [kor]

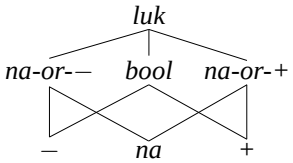
Second, MKG also functions as a flag feature for blocking overgeneration. The typical instantiation that might be overgenerated but for MKG is *topic-comment* constructions, which the next subsection elaborates on.

The type *mkg* is used as the value of the feature MKG, and introduces two further features, FC (FoCus-marked) and TP (ToPic-marked).

(14) $\begin{bmatrix} \text{MKG} & \begin{bmatrix} \text{FC} & \textit{luk} \\ \text{TP} & \textit{luk} \end{bmatrix} \end{bmatrix}$

The value type of TP and FC is *luk*, which is a supertype of *bool* (boolean) and *na* (not-applicable).[13] *Luk*, consisting of six subtypes including +, –, and *na*, can capture the marking type of constituents more flexibly than *bool* that consists of only two subtypes: namely + and –.

(15)



The value of MKG is always a subtype of *mkg*, as sketched out in Figure 9.2, in which *tp* is constrained as [TP +], *non-tp* as [TP *na-or-*−], *fc* as [FC +], and *non-fc* as [FC *na-or-*−]. Types at the bottom line multiply inherit from the intermediate supertypes, and thereby both FC and TP are fully specified. Instantiations of which *mkg* values are assigned to which information structure markings are as follows.

[13]The idea of using *luk* comes from the ERG (English Resource Grammar, Flickinger 2000).

Focus and topic markers in some languages have a fairly straightforward value of MKG. For instance, the contrastive topic marker in Vietnamese *thì* presented in (1) has [MKG *tp-only*]. The focus clitics *é* and *á* in Rendile exemplified in (p. 68) have [MKG *fc-only*]. The clitic *=m* in Ingush conveys a contrastive focus meaning (p. 87), which also involves [MKG *fc-only*]. The two types of Cantonese particles (p. 69), such as *aa4* for topic and *aa3* for focus, have [MKG *tp-only*] and [MKG *fc-only*], respectively.

The A and B accents in English, in line with the analysis in Hedberg (2006), are also straightforwardly assigned. The A-accent (H*) is responsible for conveying a non-contrastive focus meaning, whereas the B-accent (L+H*) can be used to express topic (irrespective of contrastive or non-contrastive) or contrastive focus. The A-accent exclusively used for marking focus has [MKG *fc-only*], while the B-accent has a less specified value such as [MKG *tp*]. That is, the value of MKG|FC remains underspecified.

Lexical markers in Japanese and Korean only partially constrain the meaning. As is well-known, Japanese and Korean employ three types of NP markings; (i) case-marking (e.g. *ga* and *i/ka* for nominatives), (ii) *wa* and (*n*)*un*-marking, and (iii) null-marking (expressed as Ø in the examples presented thus far). The distinction among their MKG values is crucially used in handling the interaction between lexical markings and scrambling in these languages (discussed in detail in §12.3 (p. 253)). First, the case markers have [MKG *unmkg*], given that they are not essentially markers of expressing information structure, although they indirectly have to do with some information structure meanings (i.e. *non-topic* (Heycock, 1994)). Yet, *unmkg* does not necessarily imply a case-marked constituent cannot be used for focus or topic. Note that, in this dissertation, information structure markings are neither a necessary condition nor a sufficient condition for information structure meanings. Second, the MKG value of *wa* and (*n*)*un* may be either *tp* or a fully specified type such as *tp-only*. This dissertation supports the former, because contrastively used markers and non-contrastive used ones show different prosodic behaviour from each other in Japanese and Korean (Chang, 2002; Nakanishi, 2007). For example, as already provided in Chapter 8 (p. 152), Chang argues that non-contrastive ('thematic' in his terminology) topic has [STR $< 1 >$] while contrastive topic has [STR $< 3 >$] in Korean. If we can deploy a resolution system distinguishing the difference between them in the future, the value of MKG|FC would not remain underspecified, and thereby the information structure will be more concretely constrained. In other words, non-contrastively topic-

marked constituents will have [MKG *tp-only*], whereas contrastively topic-marked ones will have [MKG *fc-+-tp*]. *Fc-+-tp* as shown in Figure 9.2 means both values of MKG|FC and MKG|TP are +. Note that these values do not violate the theorem that focus and topic are mutually exclusive. Since MKG is exclusively concerned with markings, *fc-+-tp* does not imply the constituent is regarded as containing both focus and topic. [MKG|TP +] will come from the lexical information of (*n*)*un*, and [MKG|FC +] will be obtained from the prosodic information of the constituent (i.e. [STR $< 3 >$]). However, a completely reliable system for detecting prosody in for Japanese and Korean, to my knowledge, is non-existent for now. The value of MKG|FC of the topic markers, thus, needs to remain underspecified in the current work. Finally, null-marked phrases in Japanese and Korean should be evaluated as remaining underdetermined in information structure markings (i.e. *unmkg*).

The MKG feature also plays a role in calculating the extent of focus projection. As surveyed in the previous chapter, most previous HPSG-based analyses of information structure assume that prosody expressing focus is responsible for spreading the meaning of focus to the larger constituents (Bildhauer, 2007, among others). However, focus is projected into the larger phrases not only by means of prosody but also by lexical markers in some cases (Choe, 2002). The component that responsibly gives rise to focus projection is [MKG|FC +] within the proposal of this dissertation. The A-accent in English and the prosodic pattern of marking focus in Spanish inherently have [MKG|FC +], which originally comes from [UT|DTE +], if we include Bildhauer's phonological structure and related rules in the grammars.

9.4 Sentential Forms (*sform*)

The value of ICONS can be constrained by phrasal types as well as lexical types. In order to capture a generalization about syntactic combination between two phrases with respect to information structure, a type hierarchy representing sentential forms is required. Recall that many previous studies argue that information structure contributes to a sentential grammar (Lambrecht, 1996; Engdahl and Vallduví, 1996; Paggio, 2009; Song and Bender, 2011). Building on the previous literature, I propose Figure 9.3 as the classification of phrasal types. The main purpose of *sform* is to arrange information structure components in a sentence. This type hierarchy does not have to do with linear orders of the components, unlike Figure 8.1 given in the previous chapter (p. 148).

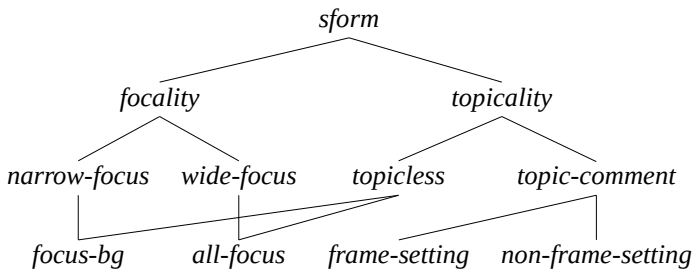


Figure 9.3: Type hierarchy of *Sform*

Lambrecht (1996) proposes information structure is deeply associated with how a sentence is formed. Engdahl and Vallduví (1996), likewise, regard information structure (information packaging in their terminology) as a part of sentential grammar. Paggio (2009) provides a hierarchy representing sentential forms for Danish as shown in Figure 8.1, which is quite similar to Figure 9.3. Paggio's type hierarchy ends up with various phrasal rules that simultaneously inherit from other fundamental phrasal rules, and this method is also made use of in Song and Bender (2011). Song and Bender (2011) analyze scrambling in Japanese and Korean (i.e. in the OSV order) in such a way that a combination between scrambled objects and VPs forms an instance whose type inherits from both *head-comp-phrase* and *topic-comment*. This dissertation follows the same strategy for placing constraints on phrase structure types with respect to information structure.

However, there is a methodological difference between what was proposed in previous literature and that in this dissertation. In previous studies, the types standing for sentential forms characterize the linear order of components. For instance, the instruction-types provided in Engdahl and Vallduví (1996), such as *link-focus*, *link-focus-tail*, *all-focus*, and *focus-tail*, and the node names in the hierarchies of Paggio (2009) and Song and Bender (2011), such as *topic-focus* and *topic-bg-focus*, reflect constraints on which elements are followed by the other(s). In Figure 9.3, by contrast, only *topic-comment* is constructed based on the linear order. All the others merely stand for components that the construction is composed of, without respect to the linear order. *Focus-bg* in Figure 9.3, which is normally used for clefts constructions, does not mean *focus* is followed by *bg*. Focused constituents are postposed in the cleft constructions in Korean (Kim and Yang, 2009), but the cleft constructions are instances of *focus-bg*. In the current work, the linear order of the components

is manipulated by phrase structure rules in each language grammar, particularly in tandem with *head-initial* and *head-final*.

The types of *sform* operate with MKG features to stratify meaning of information structure at the phrase level. The *sform* types are inherited by phrase structure rules. Not all phrase structure rules inherit from *sform* types, but if a specific syntactic operation is used for expressing information structure (e.g. scrambling ins Japanese and Korean), the rule for the constructions inherits from something in Figure 9.3.

Since sentential forms are basically a matter of how two phrases are combined with each other, *sform* inherits from *binary-headed-phrase* (made up of HEAD-DTR (head-daughter) and NON-HEAD-DTR (non-head-daughter)). We may ask why it is necessary to refer to MKG features of daughters in building up parse trees and why *sform* is required to be additionally introduced as a single phrase structure type. Several types of constructions use *sform*. Those include (i) the preverbal/postverbal position of focused constituents, (ii) cleft constructions, (iii) comment markers (e.g. *shì* in Mandarin Chinese (von Prince, 2012) and *ba* in Abma (Schneider, 2009)) that always entail focus projection, and (iv) scrambling in Japanese and Korean (Choi, 1999; Ishihara, 2001; Song and Bender, 2011). These are respectively relevant to (i) *narrow-focus*, (ii) *focus-bg*, (iii) *wide-focus*, and (iv) *topicless vs*. *topic-comment*.

Sform is bipartitely divided into *focality* and *topicality*, which indicates marking (i.e. values of MKG) and/or meaning (i.e. values of ICONS) of components of information structure in the arguments. *Sform* and its subtypes, as presented below, place constraints on MKG, which implies sentences are realized depending on information structure markings of elements. Since *sform* also places constraints on ICONS, it serves to relate the marking to the meaning.

Focality takes *fc-only* as the value of MKG, which indicates the phrase includes a focus-marked constituent. *Focality* is divided into *narrow-focus* and *wide-focus*. The distinction between them, however, is not necessarily equivalent to argument focus *vs*. predicate focus (Lambrecht, 1996; Erteschik-Shir, 2007), because verbs can bear *narrow-focus*. As shown in (16), only the MKG value on the mother is restricted in *focality*. The value is used for further composition: Some phrase structure rules prevent focus-marked constituents (i.e. specified as [MKG|FC +]) from being used as the daughter. Some phrase structure rules, on the contrary, require an explicitly focus-marked constituent as the daughter.

(16) $\begin{bmatrix} \textit{focality} & \\ \text{MKG} & \textit{fc-only} \end{bmatrix}$

Topicality is mainly concerned with how the topic is realized in a sentence. *Topicality* does not have any specific constraint for now, because *topicless* and *topic-comment* are unlikely to share a feature cross-linguistically. Nonetheless, it is introduced into the hierarchy considering symmetry with *focality*. Subtypes of *topicality* have the following AVMs in which HD and NHD are short for HEAD-DTR and NON-HEAD-DTR respectively.

(17) a. $\begin{bmatrix} \textit{topicless} & \\ \text{HD}\,|\,\text{MKG} & \textit{non-tp} \\ \text{NHD}\,|\,\text{MKG} & \textit{non-tp} \end{bmatrix}$ b. $\begin{bmatrix} \textit{topic-comment} & \\ \text{MKG} & \textit{tp} \\ \text{NHD}\,|\,\text{MKG} & \textit{tp} \end{bmatrix}$

Note that *topic-comment* has a constraint on the MKG value of the mother, just as *focality* above has a constraint of [MKG *fc-only*]. In *topic-comment* constructions (e.g. *as for* ... constructions), topics are followed by other constituents. Once a construction is identified as *topic-comment*, there are two options in further composition. If there exists another topic in the left side, and the topic is frame-setter, then further composition is allowed. Otherwise, the *topic-comment* instance cannot be used as a head-daughter in further composition. The subtypes of *topic-comment* (i.e. *frame-setting* and *non-frame-setting*) details this distinction.

As noted, not all sentences have topics. Presumably, all cleft constructions are *topicless*. Accordingly, a constituent with [MKG|TP +] cannot be the non-head-daughter in cleft constructions. For example, cleft clauses in Korean show a strong tendency to be exclusive to (*n*)*un*-marked constituents, as exemplified below.

(18) a. ku chayk-ul/*un ilk-nun salam-i/un Kim-i-ta.
the book-ACC/NUN read-REL person-NOM/NUN Kim-COP-DECL
'It is Kim that reads the book.'

b. Kim-i/*un ilk-nun kes-i/un ku chayk-i-ta.
Kim-NOM/NUN read-REL thing-NOM/NUN the book-COP-DECL
'It is the book that Kim reads.' [kor]

The distinction between *topicless* and *topic-comment* is especially significant in topic-prominent languages, such as Chinese, Japanese, and Korean, in which forms of marking topics play an important role in syntactic configuration (Li and Thompson, 1976; Huang, 1984). Lambrecht (1996)

regards (19), in which *inu* 'dog' is combined with the nominative marker *ga* instead of the so-called topic marker *wa*, as a topicless sentence.[14] That means not all subjects are topics.

(19) inu ga hasitte iru.
dog NOM running
'The dog is running.' [jpn] (Kuroda, 1972, p. 161)

In line with Lambrecht's claim, this dissertation provides for this with the type *topicless*. The difference between *topicless* and *topic-comment* performs a role in constructing Japanese and Korean grammars, which is partially proposed in Song and Bender (2011). For instance, *head-subj-rule* and *head-comp-rule* in these languages need to be divided into several subrules, depending on whether the non-head-daughter of the rules are *wa* or (*n*)*un*-marked or not. The rules dependent upon the value of MKG in Japanese and Korean are provided in §12.3.

There is a need to refine the meaning of *topicless*. On one hand, it indicates that topic is not realized in surface form, not that there is no topic at all in the utterance. For example, *topicless* in Japanese means the non-head-daughter of the phrase is not *wa*-marked. For example, since *inu ga* 'dog NOM' in (19) is not a *wa*-marked constituent, and it constitutes the sentence with the predicate *hasitte iru* 'running' as a non-head-daughter, the sentence ends up with *topicless*. On the other hand, MKG has to do with only overtly expressed items. An utterance sometimes has an implicit topic which is not overtly expressed. Topic-drop, for instance, often occurs in topic-prominent languages, which involves an implicit topic. It is true that dropped topics in the current work surely have a representation in the ICONS, but they are not concerned with MKG.

Narrow-focus and *focus-bg* come under *focality*, but constraints on them are language-specific. This is because they are not reflected in the linearization of components. For example, assume two hypothetical languages Language A[15] and B, which have a symmetrical property as follows.

(20) a. [Language A] employs SVO as its basic word order.

b. Focused constituents in [Language A] are realized in the immediate preverbal position.

c. Additionally, there is an optionally used accent, which expresses focus.

---

[14] Kuroda (1972) regards (19) as a subjectless sentence.

[15] Language A is hypothetically modeled quite analogously to Hungarian (É. Kiss, 1998; Szendrői, 1999). Hungarian is known as adopting SVO word order preferentially (Gell-Mann and Ruhlen, 2011), though it is sometimes reported that the word order in Hungarian is pragmatically conditioned (i.e. no dominant order (Kiefer, 1967)).

(21) a. [Language B] employs SOV as its basic word order.

b. Focused constituents in [Language B] are realized in the immediate postverbal position.

c. The same as (20c)

Based on (20-21), the object in SOV word order in Language A and the objects in SVO word order in Language B are narrowly focused. They participate in *narrow-focus* as a non-head-daughter. Both [OV] in Language A and [VO] in Language B are instantiated as *head-comp-phrase*, but the former is constrained by *head-final* in which the head (i.e. the verb) follows its complement, while the latter is constrained by *head-initial* in which the head precedes. Thus, from a cross-linguistic perspective, linear order does not have to be used as a key to constrain *narrow-focus*. On the other hand, a distinction between HEAD-DTR and NON-HEAD-DTR cannot be used for constraining *narrow-focus*, either. For instance, focused constituents in clefts behave as the head of cleft clauses realized as relatives. In other words, while both focused items in [OV/VO] in Language A and B respectively are non-head-daughters, the focused ones in cleft are head- daughters. In a nutshell, it is true that *narrow-focus* and *focus-bg* require some constraints on information structure marking and meaning, but the constraints must be placed language by language or construction by construction.

In other words, the type *focus-bg* does not have the same constraints in all languages. At least, there are two subtypes of *focus-bg* across languages: one where the HD involves [MKG *fc*] and one where the NHD does. For example, the cleft constructions (as an instance of *focus-bg*) in English basically inherit the following AVM. More specific constraints can be imposed language-specifically.

(22) $\begin{bmatrix} \textit{focus-bg} & \\ \text{HD} \mid \text{MKG} & \textit{fc} \\ \text{NHD} \mid \text{MKG} & \textit{unmkg} \end{bmatrix}$

This AVM serves to prevent constituents with information structure markers from being used in cleft constructions.[16]

To present another instance, *narrow-focus* in Language A can be constrained as follows. Note that the values on HD and NHD are in reverse to those in (22).

[16]For instance, cleft constructions in Korean are not compatible with (*n*)*un* in cleft clauses.

(23) $\begin{bmatrix} \textit{narrow-focus} & \\ \text{HD} \mid \text{MKG} & \textit{unmkg} \\ \text{NHD} \mid \text{MKG} & \textit{fc-only} \end{bmatrix}$

(23) also explains ungrammaticality of pseudo sentence (24c) in Language A; HD of *narrow-focus* requires a minus value as a value of MKG|FC, which conflicts with a focus accent that falls on the verb. Because presumably there is no other way to construct (24c), a pseudo sentence (24c) remains ungrammatical.

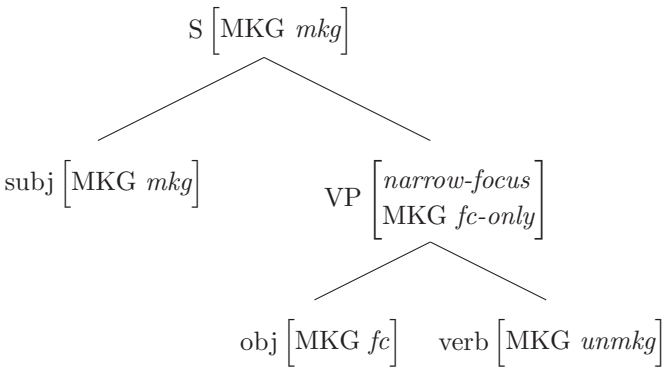
(24) a. subj verb obj. (in the neutral word order)

b. subj obj verb. (focus on obj)

c. *subj obj VERB. (a focus-marking accent on verb)

A sample derivation for (24b) can be sketched out as (25), showing only information structure markings and sentential forms. Note that MKG is seen only locally, because it is not a head feature. Thus, the value would not be transmitted to the higher nodes if it were not for an extra constraint.

(25)



Wide-focus, next, is particularly related to realization of comment markers as mentioned earlier. For instance, Mandarin Chinese employs *shì*[17] as exemplified below, which indicates the remaining part after it is in the focus domain (von Prince, 2012). In a similar vein, Li (2009) regards *shì* as a marker responsible for contrastive meanings: The constituents after *shì* are contrastively focused.

(26) Zhāngsān [shì [xuéxí yīxué]].
Zhangsan SHI study medicine
'Zhangsan studies medicine.' [cmn] (von Prince, 2012, p. 336)

[17] von Prince (2012) says this *shì* is different from the copula *shì* (i.e. homonym).

Thus, the type of construction licensed by *shì* (and comment markers in other languages such as Abma (Schneider, 2009)) has to inherit (27). Note that this constraint is language-universal, unlike *narrow-focus*. In the context of grammar engineering for the LinGO Grammar Matrix system, (27) is encoded into `matrix.tdl`, while the AVM for *narrow-focus* could be either empty or encoded in `mylang.tdl`. In accordance with (27), any constituents after the comment marker should not be topic-marked.

(27) $\begin{bmatrix} \textit{wide-focus} & \\ \text{HD} \mid \text{MKG} & \textit{fc} \\ \text{NHD} \mid \text{MKG} & \textit{fc} \end{bmatrix}$

All-focus inherits from both *wide-focus* and *topicless*.[18]

(28) $\begin{bmatrix} \textit{all-focus} & \\ \text{NHD} \mid \text{MKG} & \textit{fc} \end{bmatrix}$

Finally, it is necessary to discriminate between *frame-setting* and *non-frame-setting*. As mentioned before, this use of the MKG feature aims to pass up appropriate values; in particular, when a topic-marked constituent occurs in the leftmost position. [NHD|L-PERIPH +] in (29a) imposes this constraint. L-PERIPH (Left-PERIPHeral) will be discussed in the next chapter (§10.3.1).

(29) a. $\begin{bmatrix} \textit{frame-setting} & \\ \text{NHD} \mid \text{L-PERIPH} & + \end{bmatrix}$

b. $\begin{bmatrix} \textit{non-frame-setting} & \\ \text{HD} \mid \text{MKG} & \textit{fc-only} \end{bmatrix}$

Chapter 3 (§3.3.3) has already confirmed that topics that function to restrict a frame of what the speaker is speaking of (i.e. so-called frame-setting topics) can appear multiply, whereas an ordinary topics cannot. For example, left-dislocated NPs cannot occur more than once, without affecting grammaticality, as shown in (30a). However, frame-setters such as *yesterday* in (30b) can occur multiple times in the sentence-initial position as presented in (30d). In other words, *topic-comment* constructions can be used as another comment, and do not constrain the value of MKG of the head-daughter. However, they cannot be used again for *non-frame-setting*.

[18]Regarding the status of *all-focus*, Lambrecht (1996, p. 232) argues that there is a clear difference from other types of focused constructions in that the pragmatic core of *all-focus* is "an absence of the relevant presuppositions". The argument sounds convincing, but this dissertation does not represent such pragmatic information on the AVMs of *sform*.

(30) a. *Kim, the book, he read it.

b. Yesterday, Kim read the book.

c. The book, Kim read it.

d. Yesterday, the book, Kim read it.

To sum up, *sform* is concerned with syntactic combination between two phrases with respect to information structure. This places constraints on both MKG and ICONS, relating the marking to the meaning. In other words, *sform* makes information structure marking and meaning interact with each other. Using the type hierarchy is adapted from the proposal of Paggio (2009). What we have in common is that if a phrase structure rule is related to expressing information structure, it can multiply inherit from both a specific type of *sform* and an ordinary phrase structure type, such as *head-subj-phrase*, *head-comp-phrase*, etc. The main difference between Paggio's approach and mine is that my *sform* hierarchy does not reflect on the linear order of components.

9.5 Graphical Representation

Song and Bender (2012) suggest representing constraints on information structure in the style of the dependency graphs of DMRS (Dependency MRS; Copestake 2009) for ease of exposition. In the same manner, the remainder of this dissertation makes use of dependency graphs to present information structure relations between individuals and clauses.

In these graphs, the ICONS values are represented as links between informatively contentful elements (introducing the referential index as the value of TARGET) and verbs (introducing the event variable as the value of CLAUSE) and as unary properties of verbs themselves. The direction of arrow stands for the binary relation between a TARGET (an entity) and a CLAUSE that the TARGET belongs to. The start point indicates the constituent that occupies the CLAUSE-KEY within the clause. The end point refers to the constituent whose INDEX is shared with the TARGET, and whose ICONS-KEY|CLAUSE is co-indexed with the CLAUSE-KEY. The node name on each arrow indicates the information structure value that the binary relation has, such as *focus*, *topic*, and so forth.

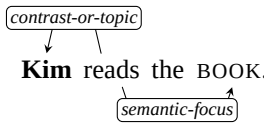
For example, a dependency graph (31b), which stands for the binary relations in the list of ICONS, is a shorthand version of the corresponding MRS representation (31a), which stands for

Kim *reads the* BOOK. in which the B-accented ***Kim*** conveys meaning of contrast and/or topic, and the A-accented BOOK bears non-contrastive focus.[19]

(31) a.

$$\begin{bmatrix} mrs \\ \text{LTOP} \quad \boxed{h1} \\ \text{INDEX} \quad \boxed{e2} \\ \text{RELS} \quad \left\langle \begin{bmatrix} proper\_q\_rel \\ \text{LBL} \quad \boxed{h3} \\ \text{ARG0} \quad \boxed{x5} \\ \text{RSTR} \quad \boxed{h4} \\ \text{BODY} \quad \boxed{h6} \end{bmatrix}, \begin{bmatrix} named\_rel \\ \text{LBL} \quad \boxed{h7} \\ \text{ARG0} \quad \boxed{x5} \\ \text{CARG} \quad kim \end{bmatrix}, \begin{bmatrix} \_read\_v\_rel \\ \text{LBL} \quad \boxed{h8} \\ \text{ARG0} \quad \boxed{e2} \\ \text{ARG1} \quad \boxed{x4} \\ \text{ARG2} \quad \boxed{x9} \end{bmatrix}, \begin{bmatrix} exist\_q\_rel \\ \text{LBL} \quad \boxed{h10} \\ \text{ARG0} \quad \boxed{x9} \\ \text{RSTR} \quad \boxed{h11} \\ \text{BODY} \quad \boxed{h12} \end{bmatrix}, \begin{bmatrix} \_book\_n\_rel \\ \text{LBL} \quad \boxed{h13} \\ \text{ARG0} \quad \boxed{x9} \end{bmatrix} \right\rangle \\ \text{HCONS} \quad \left\langle \begin{bmatrix} qeq \\ \text{HARG} \quad \boxed{h4} \\ \text{LARG} \quad \boxed{h7} \end{bmatrix}, \begin{bmatrix} qeq \\ \text{HARG} \quad \boxed{h11} \\ \text{LARG} \quad \boxed{h13} \end{bmatrix} \right\rangle \\ \text{ICONS} \quad \left\langle \begin{bmatrix} contrast\text{-}or\text{-}topic \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x5} \end{bmatrix}, \begin{bmatrix} semantic\text{-}focus \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x9} \end{bmatrix} \right\rangle \end{bmatrix}$$

b.



The arc from *reads* to **Kim** means the index of **Kim** has a *contrast-or-topic* relation to the clause represented by READS. The arc from *reads* to BOOK, likewise, means the index of BOOK has a *semantic-focus* relation to the index of READS. The root arrow on READS indicates the verb is linguistically underspecified with respect to the clause that it heads.

9.6 Summary

This chapter raises three necessities for representing information structure via ICONS: resolving discrepancies between forms and meanings in information structure, facilitating underspecifiability

[19]Note that this dissertation, according to the argument of Hedberg (2006), regards the A-accent (i.e. marked as SMALL CAPS) as prosodic means expressing non-contrastive focus (i.e. *semantic-focus*), and the B-accent (i.e. **boldfaced**) as conveying one of the meanings of non-contrastive topic, contrastive topic, or sometimes contrastive focus.

for allowing flexible and partial constraints, and capturing the fact that information structure relations are between expressions and particular clauses. Additionally, one working hypothesis is used in the ICONS-based representation: Semantically empty and syncategorematic items are informatively empty. Building upon them, I provide three type hierarchies: (i) *info-str* whose value types stratify information structure meaning, (ii) *mkg* indicating morphosyntactic markings of information structure, and (iii) *sform* that works with MKG relating to the *info-str* value. ICONS is added into *mrs*, and the value is a *diff-list* of *info-str*. ICONS identifies which element has which information structure relation to which clause. For this purpose, the typed feature structure of *info-str* includes TARGET and CLAUSE. TARGET has a coreference with the EP's INDEX (i.e. *individual*), and CLAUSE is determined by the subtype(s) of *clause*. In addition, ICONS-KEY and CLAUSE-KEY are used as a pointer during the construction of parse trees. MKG has two features; one is FC (FoCus-marked), and the other is TP (ToPic-marked). These features are independent of meanings represented as an *info-str* type. The next chapter will discuss how these elements are used to impose constraints on information structure and represent it in MRS.

Chapter 10

INDIVIDUAL CONSTRAINTS: SPECIFICS OF THE IMPLEMENTATION

This chapter dives into the details of implementing ICONS (Individual CONStraints) into MRS (Minimal Recursion Semantics (Copestake et al. 2005), or Meaning Representation System in the current context) via constraining lexical types and phrasal types. §10.1 shows how information structure is dealt with in various lexical types. §10.2 gives an explanation of information structure constraints on phrasal types. Next, §10.3 presents three additional constraints for configuring information structure in a specific way. Building upon the hierarchies and constraints presented thus far, §10.4 illustrates how information structure is represented via ICONS in four languages (English, Japanese, Korean, and Russian).

10.1 Lexical Types

This section largely addresses which lexical item inherits from which *icons-lex-item* type out of *no-icons-lex-item*, *basic-icons-lex-item*, *one-icons-lex-item*, and *two-icons-lex-item*. The first two AVMs do not inherently have an *info-str* element in the list of ICONS, but information-structure related rules can insert a value into the list for the second type. That is, if there is a clue to identify its information structure value, a value of *info-str* is introduced into the the list of ICONS and its TARGET is co-indexed with the HOOK|INDEX of the word. The last two inherently have non-empty ICONS lists. That means that they lexically have an *info-str* value in ICONS.

(1) a.
$$\begin{bmatrix} \textit{no-icons-lex-item} \\ \text{MKG} \begin{bmatrix} \text{FC} & \textit{na} \\ \text{TP} & \textit{na} \end{bmatrix} \\ \text{ICONS} \left\langle ! \;\; ! \right\rangle \end{bmatrix}$$

b.
$$\begin{bmatrix} \textit{basic-icons-lex-item} \\ \text{ICONS} \left\langle ! \;\; ! \right\rangle \end{bmatrix}$$

c.
$$\begin{bmatrix} \textit{one-icons-lex-item} \\ \text{ICONS} \left\langle ! \begin{bmatrix} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

d.
$$\begin{bmatrix} \textit{two-icons-lex-item} \\ \text{ICONS} \left\langle ! \begin{bmatrix} \end{bmatrix}, \begin{bmatrix} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

Lexical entries that cannot be marked with respect to information structure inherit from *no-icons-lex-item* (1a). In other words, information structure markings are not-applicable to them, which is specified as [MKG [FC *na*, TP *na*]]. For example, relative pronouns and expletives in English are instances of *no-icons-lex-item*. Other contentful items introducing an EP inherit from one out of (1b-d). The choice among them depends on how many clauses are subordinate to the lexical type. The prefixes represent how many clauses are created by the type: *Basic-* means the lexical type does not include any clausal subject or clausal complement in ARG-ST. *One-* means either a clausal subject or a clausal complement is subordinate to lexical type. *Two-* means there exist a clausal subject and also a clausal complement. If a verbal type forms a monoclausal construction, the type of *icons-lex-item* of the verbal item is the basic one (i.e. *basic-icons-lex-item*). If a verbal type has ARG-ST information that includes one or more clausal argument(s) (i.e. multiclausal), its *icons-lex-item* type is either *one-icons-lex-item* (a sentential complement or a sentential subject) or *two-icons-lex-item* (both of them). The extra *info-str* values in (1c-d) are required for this purpose. If a verbal item assigns an information structure value to its subordinate clause(s), the verbal item should be either of *one-icons-lex-item* or *two-icons-lex-item*.

10.1.1 Nominal Items

Nominal items, including common nouns, proper nouns, and pronouns, inherit from *basic-icons-lex-item*. One exceptional lexical type includes expletives (e.g. *it* in English), because they cannot be information-structure marked (Lambrecht, 1996). Expletives inherit from *no-icons-lex-item*. One question regarding *info-str* of nominal items is whether different types of nominals can participate in information structure in the same way and with the same status. The current work does not recognize any difference in the information structure of nominal items. In other words, other than *basic-icons-lex-item*, there is no further information structure constraint on nominal items.

Pronouns have been regarded as a component associated with information structure in a different way (Lambrecht, 1996). Pronouns, roughly speaking, can be bisected into (i) unaccented ones and (ii) accented ones. Lambrecht argues that the distinction between them can be sufficiently explained in terms of information structure. Across languages, (i) unaccented pronouns preferentially involve topics. This finding is bolstered by the evidence that unaccented pronouns are the most frequently

used form of expressing topic in Spoken French (Lambrecht, 1986). Besides, unaccented pronouns cannot be used for expressing focus, because they are incompatible with (2).

(2) Focus Prominence: Focus needs to be maximally prominent. (Büring, 2010, p. 277)

On the other hand, (ii) accented pronouns can be divided again into (ii-a) ones with a topic-marking accent, and (ii-b) ones with a focus-marking accent. Lambrecht illustrates the linguistic distinction between (ii-a) and (ii-b) in Italian as follows.

(3) a. IO PAGO.
I pay.
'I'll pay.' [ita]

b. Pago IO.
pay I
'I'll pay.' [ita] (Lambrecht, 1996, p. 115)

The preverbal pronoun IO in (3a) expresses a topic, with a rising intonation contour. On the other hand, the pronoun conveying a focus meaning in (3b) occurs sentence-finally, and has a falling intonation contour, indicating the end of the assertion.

From a theoretical point of view, it seems clear that pronouns show different behaviours in packaging information.[1] However, the current work, based on text processing, cannot deploy such a division. Phenomena related to (ii) can be modelled with hypothetical suffixes such as '-a' and '-b', however the responsibility is borne by the hypothetical suffixes or alternatively lexical rules introducing the prosodic information, not the pronouns themselves.

[1]The nominal items also differ from each other in discourse status. Kaiser (2009) argues that the use of different kinds of referring expressions is relevant to the salience of the antecedents; the more salient antecedent it refers to, the more reduced a form (e.g. dropped subjects) appears. That is, selecting a type of referential form largely hinges on how salient the antecedent is. The discourse status of nominal categories that take *ref-ind* (a subtype of *individual*) as the value type of HOOK|INDEX is represented as COG-ST (COGnitive-STatus) in the current LinGO Grammar Matrix system (Bender and Goss-Grubbs, 2008). Discourse status is related to information status (e.g. given *vs.* . new); COG-ST covers information status from a higher level. Information status, as discussed in Chapter 3, has often been studied in tandem with information structure, but it is neither a necessary nor a sufficient condition for information structure. In sum, since discourse status is not directly responsible for representing information structure, the current work leaves discourse-related information to future work.

10.1.2 Verbal Items

The analysis proposed here uses the event variable associated with the head of the clause to stand in for the clause, and as a result, the lexical types for verbs (typical clausal heads) need to be constrained appropriately. Most contentful verbal items inherit from either *basic-icons-lex-item*, *one-icons-lex-item*, or *two-icons-lex-item*. Verbs inherently have lexical information about how many elements exist in the ICONS list and how they are bound with the semantic head in the clause that the elements belong to. That means the number of elements in the list of ICONS depends on how many clausal dependents a verbal type has. This information is specified inside the ARG-ST of verbal types. If a verb takes no clausal phrase(s) as its dependent(s), the verb locally constitutes a monoclausal phrase. In this case, no element is required to be included in the ICONS list (i.e. *basic-icons-lex-item*). In some cases, a verb lexically places an information structure constraint on its subordinated clause(s). If the ARG-ST of a verbal type includes either one clausal subject or one clausal complement, the verbal type constitutes a locally embedded constructions in which one clause is subordinate to the main clause (i.e. *one-icons-lex-item*). Sometimes, both the subject and one of the complements can be clausal. In this case, two elements of *info-str* are needed in the ICONS list (i.e. *two-icons-lex-item*).

There is an exception: some semantically and informatively empty verbal items are not able to be marked with respect to information structure. These verbs inherit from *no-icons-lex-item*. For example, semantically empty copulae (e.g. specificational copulae English) are incapable of contributing an ICONS element.

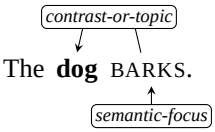
Main Verbs

Because main verbs in principle can be marked with respect to information structure, they do not inherit from *no-icons-lex-item*. Excluding *no-icons-lex-item* for the reason, main verbs can be one of these three types of *icons-lex-item*: *basic-icons-lex-item*, *one-icons-lex-item*, or *two-icons-lex-item*.

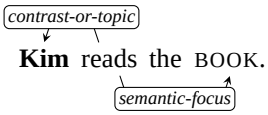
Common verbs that constitute a monocalusal construction (e.g. intransitives (e.g. *bark* in (4a), transitives (e.g. *read* in (4b), and ditransitives (e.g. *give* in (4c)), inherit from *basic-icons-lex-item*. Causative verbs (e.g. *make* in (4d)) and perception verbs (e.g. *see* in (4e)) also inherit from *basic-icons-lex-item*, because their verbal complements, such as *bark* in (4d) and *barking* in (4e), are

tenseless (i.e. infinite). Thus, all dependents, including the subject and the complements, are bound to the verb that functions as the semantic head in the sentence (i.e. an element that takes the INDEX in the finite clause). That means the CLAUSEs of the dependents are co-indexed with the HOOK|INDEX of the main verb by *non-rel-clause* (p. 181) which *decl-head-subj-phrase* inherits from. Note that some relations of information structure are not captured in the following graphs. This is because if there is no specific clue to identify information structure meaning that a constituent conveys, no value is gathered into the list of ICONS. For ease of exposition, in the following examples, the leftmost elements (i.e. subjects) are B-accented (conveying *contrast-or-topic*) and the rightmost elements are A-accented (conveying *semantic-focus*).

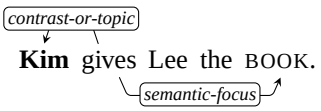
(4) a.



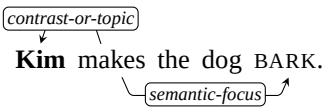
b.



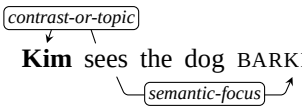
c.



d.

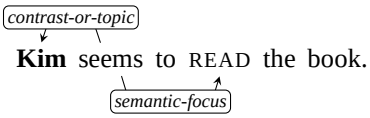


e.

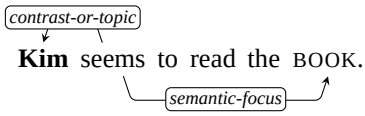


Raising and control verbs have the same mapping type in *info-str*; every dependent marked with respect to information structure within a single clause has a co-index between its CLAUSE and the INDEX of the semantic head in the clause (i.e. matrix clause verb).

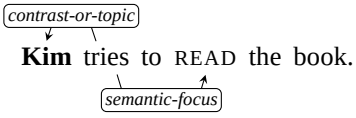
(5) a.



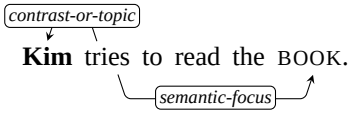
b.



c.



d.

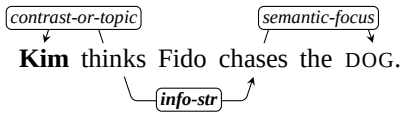


For example, *the* BOOK in (5b) is syntactically a complement of *read*, but is informatively bound to *seems* whose INDEX is connected to the INDEX in the clause. Additionally, it is interesting that *Kim* has an *info-str* role in the matrix finite clause, even though it is not the semantic argument of

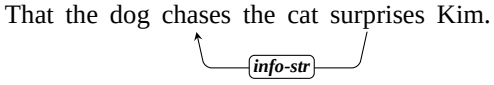
seems. The same goes for a control verb *try* in (5c-d).[2]

Several verbal types take clausal complements as shown in (6a) or clausal subjects as shown in (6b), which inherit from *one-icons-lex-item*.

(6) a.



b.



In (6a), the arrow from *chases* and *dog* is locally established within the embedded clause. The binary relation in the embedded clause does not have to do with the main verb *thinks*. The main verb *thinks* also has an arrow to the subject *Kim* in the local domain. The key point of this example is the arrow from the main verb *thinks* to the verb in the embedded clause *chases*. This arrow is introduced by the element in the ICONS list of *think* (*one-icons-lex-item*), and shows which information structure relation the embedded clause has to the matrix clause. Likewise, the arrow from *surprises* and *chases* in (6b) represents the inherent *info-str* element in the ICONS list of *surprise*.

Both subjects and complements can be clausal at the same time. In these cases, it is necessary to inherit from *two-icons-lex-item*. A typical example can be found in pseudo-clefts, including *wh*-clefts, and inverted *wh*-clefts as exemplified in (7).[3] In (7), the matrix verb *is* inherently has two elements of *info-str* in the ICONS list: The CLAUSE value of the first element is its INDEX, and the TARGET is co-indexed with the INDEX of the verb in the clausal subject (i.e. *happened*). The CLAUSE of the second is still linked to its INDEX, and the TARGET is co-indexed with the INDEX of the verb in the complement (i.e. *caught*).

(7) [What happened] is [they caught her without a license]. <S1A-078 #30:2:A>

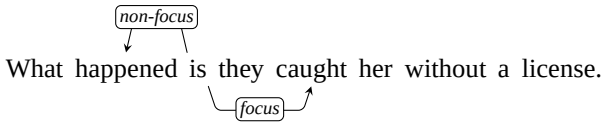
The dependency graph corresponding to (7) is presented in (8). Note that the second arrow in (7) is specified as *focus*. That is, the clausal complements in *wh*-clefts (e.g. *they caught her without a license* in (7)) is focused (Kim, 2007). Since the other constituents cannot be assigned focus, the

[2]As is well-known, raising verbs (e.g. *seem*, *appear*, *happen*, *believe*, *expect*, etc.) and control verbs (e.g. *try*, *hope*, *persuade*, *promise*, etc.) display several different properties, such as the semantic role of the subject, expletive subjects, subcategorization, selectional restriction, and meaning preservation (Kim and Sells, 2008). Nonetheless, they have *basic-icons-lex-item* in common as their supertypes, because they do not take tensed clauses as complements.

[3](7) is originally taken from the ICE-GB (Nelson et al. 2002), and the expressions in angled brackets represent the indices of each sentence in the corpus.

clausal subject in (7) is specified as *non-focus*. The verbal entry *is* includes these values as lexical information.

(8)



Adjectives

Predicative adjective items are the same as the verbal items presented thus far. The copula *is* in (10) is assumed to be semantically and informatively empty, and the adjectives function as the semantic head of these sentences. This constraint is specified in the following AVM, which pass up the CLAUSE-KEY of the second argument (i.e. the adjective).

(9) $$\begin{bmatrix} \textit{copula-verb-lex} \\ \text{ARG-ST} \left\langle \left[\text{ICONS-KEY} \mid \text{CLAUSE } \boxed{1}\right], \left[\text{CLAUSE-KEY } \boxed{1}\right] \right\rangle \end{bmatrix}$$

Happy in (10a) and *fond of* in (10b), which do not constitute multiclausal constructions, inherit from *basic-icons-lex-item*. Next, *obvious* in (10c-d) takes clausal subjects, and *sure* and *curious* in (10e-f) take clausal complements. These inherit from *one-icons-lex-item*.

(10) a. Kim is happy.

b. Kim is fond of apples.

c. That the dog barks is obvious.

d. It is obvious that the dog barks.

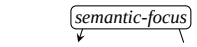
e. Kim is sure that the dog barks.

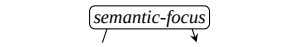
f. Kim is curious whether the dog barks.

There are also raising adjectives (e.g. *likely*) and control adjectives (e.g. *eager*), which also inherit from *basic-icons-lex-item*, just as raising and control verbs.

Attributive adjectives are different in that they introduce no *info-str* value that take their own event variable as the value of CLAUSE. Attributive adjectives and the nouns they are modifying share the value of CLAUSE, which is co-indexed with the INDEX of the verb heading the clause

that they are part of. For example, the arrows on *big* in (11) come from the main verb of each sentence. This linking strategy is constructionally constrained by *head-mod-phrase*. §10.2 gives an explanation of how this linking is achieved via *head-mod-phrase*.

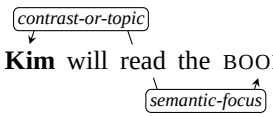
(11) a. 

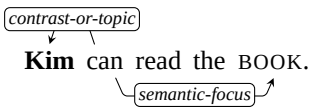
b. 

There is a distinction between attributive and predicative adjectives with respect to building up the list of ICONS, but there is no necessity to use an extra lexical rule to discriminate them. What is of importance is incrementally gathering *info-str* values into the list of ICONS, and this strategy is achieved by phrase structure rules, such as *head-comp-rule* for predicative ones (as specified in the ARG-ST of (9)) and *head-mod-rule* for attribute ones (see (25) presented in the next section).

Auxiliaries

As far as ICONS is concerned, auxiliaries in English are divided into two subtypes. One contributes no predicate and no ICONS element, and thereby inherits from *no-icons-lex-item*. The other introduces an EP to RELS, and thereby inherits from *basic-icons-lex-item*. Since complements of auxiliaries are always non-finite, we do not have to see *one-icons-lex-item* or *two-icons-lex-item* for auxiliaries. For example, *will* in (12a) is semantically empty, and does not occupy the INDEX of the clause. Such an auxiliary, therefore, does not have any *info-str* element in ICONS, either. Instead, the main verb, such as *read* in (12a), has arrows to each of its dependents. By contrast, *can*[4] in (12b) has arrows to all individuals that introduce *info-str* into the clause.

(12) a. 

b. 

These two types of auxiliaries inherit from different lexical types. First, *will* is an instance of *no-icons-lex-item*, which prevent this auxiliary from participating in articulation of information structure. Second, *can* inherit from the following AVM. The CLAUSE value of the subject (i.e. the first element in ARG-ST) is co-linked to its own CLAUSE-KEY, and the second argument (i.e. the VP)

[4]Its LKEYS|KEYREL|PRED is specified as '_can_v_modal_rel' in the ERG.

also shares the CLAUSE value with its own CLAUSE-KEY.[5]

(13)
$$\begin{bmatrix} \textit{trans-first-arg-raising-lex-item-1} \\ \text{CLAUSE-KEY} \mid \text{CLAUSE } \boxed{1} \\ \text{ARG-ST} \left\langle \begin{bmatrix} \text{ICONS-KEY.CLAUSE} & \boxed{1} \end{bmatrix}, \begin{bmatrix} \text{ICONS-KEY.CLAUSE} & \boxed{1} \\ \text{CLAUSE-KEY} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

The main verb which serves as the complement of modal auxiliaries can sometimes take clausal complements. In this case, the CLAUSE-KEY is still occupied by the auxiliary as sketched out in (14): The arrow to the verb in embedded clause headed by *chases* is lexically introduced by *think*, which inherits from *one-icons-lex-item*. The CLAUSE-KEY of the second *info-str* that *think* introduces is yet unbound in the VP *think Fido chases the dog*. Building up *head-subj-phrase*, the CLAUSE-KEY that *chases* has in relation to the matrix clause is co-indexed with the INDEX of *can*, finally. Recall that the value of CLAUSE is bound when one clause is identified (§9.2.1). *Head-subj-phrase*, which is a subtype of *clause* and *non-rel-clause*, serves to identify which EP occupies the INDEX of the clause and fills in the value of CLAUSE.

(14) *contrast-or-topic* *semantic-focs*
Kim can think Fido chases the DOG.
info-str

Copulae

Copulae, generally speaking, have at least three usages as exemplified in (15). Among them (15b-c) have the same properties in English, and thereby the different names might be useless. That is, a single lexical entry is used for both of them. Yet, because there are some languages in which the second and the third types are lexically different, I would use the three different names herein for convenience sake.

(15) a. Kim is the student. (identificational)

b. Kim is happy. (specificational)

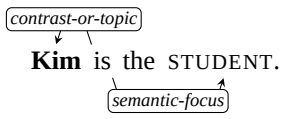
c. Kim is in Seattle. (locative)

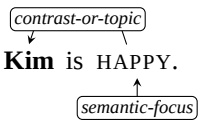
[5]Note that (13) actually contains more constraints, such as HCONS.

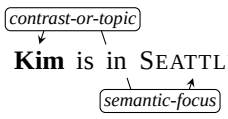
(15a) can be paraphrased into *Kim is identical to the student.*, while (15b-c) cannot. Traditionally, identificational copulae in many languages are treated as ordinary transitive verbs, whose ARG-ST includes one NP for the subject and the other NP for the complement (i.e. a two-place predicate). Identificational copulae, thus, are assumed to be contentful and thereby introduce an EP, whose LKEYS|KEYREL|PRED value would be something like '_be_v_id_rel'.

In contrast, the others are semantically empty items that do not introduce any EP into the list of RELS in MRS. Thus, the semantic heads of (15b-c) are respectively computed as *happy* and *in*, not the copula *is*. Specificational copulae also have a difference from locative copula. Both are lexically the same in English, but the copula in (15c) can be replaced by a locative verb, such as *exists*. In some languages, specificational copula and locative copula are lexically different; for example, *i* as a specificational copula *vs*. *iss* as a locative verb in Korean, *shì* as a copula *vs*. *zài* as a locative verb in Mandarin Chinese. Since the locative verb is not semantically void in such a language, the lexical entry for the locative verb has a PRED value like '_be+located_v_rel'.

Identificational copulae inherit from *basic-icons-lex-item*, while the others inherit from *no-icons-lex-item*. That means the semantic head that occupies the INDEX of the clause in (15b-c) are *happy* and *in*. In other words, the CLAUSE-KEY in (15b-c) is linked to the INDEX of *happy* and *in*, respectively. (16a-c) stand for the information structure of (15a-c), respectively.

(16) a. 

b. 

c. 

10.1.3 Adpositions

Adpositions normally inherit from either *basic-icons-lex-items* or *one-icons-lex-item*. Every information-structure marking adposition inherits from *one-icons-lex-item*. If an adposition does mark information structure by itself, it inherits from *basic-icons-lex-item*. Adpositions that inherit from *basic-icons-lex-items* can have an ICONS element later when other means of marking information structure (e.g. an accent on *under* in (17b)) is additionally used.

(17) Q: Did Kim put the book on the desk?
A: No. Kim put the book under the desk.

Prepositions in English do not inherit from *one-icons-lex-items*, because there is no information-structure marking preposition in English. Japanese has both types. As discussed thus far, information-structure marking postposition, such as *ga* (nominative) and *o* (accusative) and *wa* (contrast or topic), are instances of *one-icons-lex-item*. That means they introduce one element into the ICONS list. The TARGET of the ICONS element is co-index with the INDEX of their complement (i.e. XP that they are attached to), and the ICONS-KEY of each postposition is lexically specified: *non-topic* for *ga* and *o*, *contras-or-topic* for *wa*. Other than these, focus particles syntactically classified as postpositions in Japanese are also instances of *one-icons-lex-item*. These include *dake* 'only', *shika* 'except', *mo* 'also', and so on (Hasegawa, 2011; Hasegawa and Koenig, 2011). They behave in the same manner as *ga*, *o*, and *wa*, but the value is *focus*. Other postpositions that do not mark information structure in Japanese inherit from *basic-icons-lex-item*. These include *made* 'till', *kara* 'from', etc.

10.1.4 Determiners

Determiners inherit from either *one-icons-lex-item* or *basic-icons-lex-item*, depending on whether or not they mark information structure by themselves. English does not have determiners that inherit from *one-icons-lex-item*, because there is no information-structure marking determiner. It is reported that some languages employ information-structure marking determiners. For example, Lakota[6] uses a definite determiner *k'uŋ* to signal contrastive topic. These determiners inherently include an ICONS element (i.e. *one-icons-lex-item*).

Determiners in English may bear the A-accent as shown in (18).

(18) a. Kim reads THE book.

b. Kim reads SOME/ALL books.

However, the individuals associated with what *focus* is assigned to are the nouns, not the determiners themselves. That is, the focused items in (18) are *book*(*s*), not the determiners. Thus, when a determiner has an ICONS element, its TARGET should be co-indexed with the INDEX of the NP. For example, the A-accented ALL in (18b) is constrained as follows.[7]

[6]This is a Siouan language spoken in Dakota. §14.5.1 provides more information on *k'uŋ* in Lakota.

[7]This dissertation employs two hypothetical suffixes ('-a' for the A-accent and '-b' for the B-accent) and the suffixes

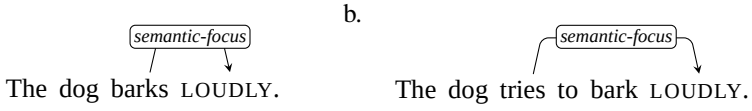
(19)
$$\begin{bmatrix} \text{STEM} & \langle \textit{all} \rangle \\ \text{MKG} & \textit{fc-only} \\ \text{SPEC} & \left\langle \begin{bmatrix} \text{INDEX } \boxed{1} \\ \text{ICONS-KEY } \boxed{2} \end{bmatrix} \right\rangle \\ \text{ICONS} & \left\langle ! \boxed{2} \begin{bmatrix} \textit{semantic-focus} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

Notably, the *info-str* value that determiners assign to the specified NPs should be consistent with the ICONS-KEY of the NPs; for example, 'THE **book**' is ill-formed, because the *semantic-focus* that the determiner involves is inconsistent with the *contrast-or-topic* the noun carries.[8]

10.1.5 Adverbs

Adverbs just inherit from *basic-icons-lex-item*. (20) is illustrative of the information structure relation that adverbs have within a clause. Just as with attributive adjectives, the relation is bound to the HOOK|INDEX of the semantic head within the clause.

(20) a. b.



10.1.6 Conjunctions

First of all, all conjunctions that take adverbial clauses as their complement inherit from *one-icons-lex-item*. They have their CLAUSE-KEY linked to the INDEX of the main clause's semantic head. Note that the semantic head of the matrix clause is co-indexed with the element in HEAD|MOD. They also have their TARGET linked to the INDEX of their complement (i.e. the semantic head of the adverbial clause).

---

are attached by lexical rules. Two lexical rules presented in (37) later take nominal and verbal items as their daughter. In addition to them, there could be one more lexical rule that takes determiners as their daughter. These rules are not presented in this dissertation.

[8] §12.1.1 provides more discussion on information structure values of quantifiers.

(21)
$$\begin{bmatrix} \textit{subconj-word} & \\ \text{HEAD} \mid \text{MOD} & \left\langle \left[\text{INDEX} \quad \boxed{1} \right] \right\rangle \\ \text{VAL} \mid \text{COMPS} & \left\langle \left[\text{INDEX} \quad \boxed{2} \right] \right\rangle \\ \text{ICONS} & \left\langle ! \begin{bmatrix} \text{TARGET} & \boxed{2} \\ \text{CLAUSE} & \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

Second, conjunctions that involve temporal adverbial clauses, such as *when*, *before*, and *after*, are related to *topic*, if they appear before the main clause (Haiman, 1978). That means that the information structure value between the two semantic heads should be *topic*. This value is assigned by the temporal conjunctions. Third, conditional conjunctions (e.g. *if* and *unless* in English) also assign *topic* to the element in ICONS (Ramsay, 1987). §11.3 provides more information about temporal and conditional conjunctions. Fourth, causal conjunctions, such as *because* in English, and *weil* in German, differ in different languages with respect to information structure relation to the matrix clause (Heycock, 2007). Therefore, their information structure value is language-specifically constrained. That is, causal conjunctions in some languages (e.g. English) have an ICONS element whose value is *info-str*, and those in other language have an ICONS element whose value is more specific.

Coordinating conjunctions, such as *and* and *or*, are another story. First, each coordinand can have its own information structure relation to the semantic head in the clause if it is marked with respect to information structure. Second, coordinands in a single coordination may have different information structure values from each other. For example, *Kim* and ***Sandy*** in (22B) have the same status in the syntax of coordination, but ***Sandy*** is contrastively focused as vetted by the correction test. In this case, while *Kim* introduces no ICONS element, ***Sandy*** introduces an ICONS element and the element is assigned *contrast-or-topic* by the B-accent.

(22) A: Kim and Lee came.
B: No. Kim and **Sandy** came.

Third, the coordinate phrase itself also can have an information structure relation to the semantic head. For instance, the fronted constituent in (23) (specified as *focus-or-topic*) is a coordinate phrase.

(23) The book and the magazine, Kim read.

In this case, an ICONS element that indicates the information structure relation between the coordinate phrase *The book and the magazine* and the main verb *read* is added into C-CONT|ICONS.[9]

10.2 Phrasal Types

Information structure can also be restricted by phrase structure rules. Phrasal types can be roughly divided into *unary-phrase* and *binary-phrase*. First, ICONS is an accumulator list: ICONS is implemented as a *diff-list*, and the element are gathered up to the tree using *diff-list* append. Second, the information-structure related features (e.g. MKG and L/R-PERIPH) are shared between mother and daughter in a *unary-phrase*, with no further constraint. For instance, *unary-phrase* is defined as in (24). L-PERIPH and R-PERIPH in (24) have not yet been mentioned. They impose an ordering constraint on constituents with respect to expressing information structure. The next section (§10.3.1) discusses how they contribute to constraining information structure at the phrasal level.

(24)
$$\begin{bmatrix} \textit{unary-phrase} & \\ \text{MKG} & \boxed{1} \\ \text{LIGHT} & - \\ \text{L-PERIPH} & \boxed{2} \\ \text{R-PERIPH} & \boxed{3} \\ \text{ICONS} & \begin{bmatrix} \text{LIST} & \boxed{4} \\ \text{LAST} & \boxed{6} \end{bmatrix} \\ \text{C-CONT}\,|\,\text{ICONS} & \begin{bmatrix} \text{LIST} & \boxed{5} \\ \text{LAST} & \boxed{6} \end{bmatrix} \\ \text{HD} & \begin{bmatrix} \text{MKG} & \boxed{1} \\ \text{L-PERIPH} & \boxed{2} \\ \text{R-PERIPH} & \boxed{3} \\ \text{ICONS} & \begin{bmatrix} \text{LIST} & \boxed{4} \\ \text{LAST} & \boxed{5} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Third, there are five basic subtypes of *binary-headed-phrase*: (i) *basic-head-subj-phrase*, (ii) *basic-head-comp-phrase*, (iii) *basic-head-spec-phrase*, (iv) *basic-head-mod-phrase-simple*, and (v) *basic-head-filler-phrase*. The first three are the same as the previous versions, but with [C-CONT|ICONS <! !>] added. The empty *diff-list* in C-CONT|ICONS means that these rules never contribute

[9]Information structure in coordinated phrases would be an interesting research topic. In particular, since LinGO Grammar Matrix system includes a library of coordination, this idea needs to be implemented and tested though it is left to future work.

ICONS elements. *Basic-head-mod-phrase-simple*, in addition to the existing constraints, is further constrained as follows: The ICONS-KEY|CLAUSE and CLAUSE-KEY of NON-HEAD-DTR has a coreference with the ICONS-KEY|CLAUSE of HEAD-DTR. That means the modifier and the modificand share the same CLAUSE-KEY. These ICONS-KEY and CLAUSE-KEY indicate which clause the adjunct is focused or topicalized within. Additionally, an empty ICONS list is added.

(25)
$$\begin{bmatrix} \textit{basic-head-mod-phrase-simple} \\ \text{HD} \mid \text{HOOK} \mid \text{ICONS-KEY} \mid \text{CLAUSE} \quad \boxed{1} \\ \text{NHD} \mid \text{HOOK} \begin{bmatrix} \text{ICONS-KEY} \mid \text{CLAUSE} & \boxed{1} \\ \text{CLAUSE-KEY} & \boxed{1} \end{bmatrix} \\ \text{C-CONT} \mid \text{ICONS} \left\langle ! \;\; ! \right\rangle \end{bmatrix}$$

Finally, *basic-head-filler-phrase* does not include [C-CONT|ICONS <! !>], because this phrase may or may not contribute ICONS elements. §14.2.4 provides an explanation of its role in configuring information structure. Related constructions include clause-initial/final focus constructions, focus/topic fronting, and so forth.

10.3 Additional Constraints on Configuring Information Structure

Other than the hierarchical constraints presented in the previous chapter, there must be some additional constraints in order to implement information-structure related phenomena within the LinGO Grammar Matrix system. L/R-PERIPH in (§10.3.1) and LIGHT in (§10.3.2), as flag features, impose a constraint on the position of components of information structure. The former is newly introduced, while the latter has already been implemented in the system. PHON in (§10.3.3), newly introduced, is adapted from Bildhauer (2007).

10.3.1 Periphery

As surveyed in Chapter 4, syntactic positioning is one of the means to express information structure. The positions associated with focus include (i) clause-initial (e.g. in Akan, Ingush, Yiddish, etc.), (ii) clause-final (e.g. in Russian, Bosnian Croatian Serbian, etc.), (iii) preverbal (e.g. in Hungarian, Basque, Turkish, etc.), and (iv) postverbal (e.g. in Portuguese, Chicheŵa, etc.). The common position for topics is sentence-initial though some languages (e.g. Danish) do not use the initial positions to signal topic.

This dissertation, in order to implement the constraint on periphery, proposes to use L/R-PERIPH. These two flag features aim to constrain the first two positions (i.e. (i) clause(sentence)-initial or (ii) clause-final). The remaining two positions (i.e. (iii) preverbal and (iv) postverbal) are constrained by the feature called LIGHT, discussed in the next subsection. Even though they have something to do with syntax and semantics, flag features take part in syntactic configuration and semantic computing only in an indirect way. They traditionally tend to be immediately under SYNSEM. The same goes for L/R-PERIPH. Their value type is *luk*, but they are usually constrained as + or – (i.e. *bool*).

[L-PERIPH +] indicates the constituent with the feature cannot be combined with another constituent leftward. [R-PERIPH +] likewise indicates there must be no other constituent on the right side of the constituent. In other words, a constituent involving out of these features has to be peripheral in word order unless there is an exceptional rule. (a) The constituent that has [L-PERIPH +] should be in the most left position within the single clause (i.e. clause-initial). (b) The constituent that has [R-PERIPH +] should be in the most right position, and thereby should be clause-final.

One of the representative cases in which both features are required can be found in Russian, which places contrastively focused constituents in the clause-initial position and non-contrastively focused ones in the clause-final position (Neeleman and Titov, 2009). Thus, the clause-initial constituent (*contrast-focus*) has [L-PERIPH +], and the clause-final constituent (*semantic-focus*) has [R-PERIPH +]. Russian has several more examples that clearly have to do with periphery (Gracheva, 2013): Russian employs a clitic *li*,[10] which should appear in the second position of an utterance. This clitic modifies the immediately preceding constituent (i.e. the most left-peripheral item), and sometimes assigns *contrast-focus* to it, possibly depending on the part of speech of the constituent it attaches to and context.[11] Notably, *li* imposes the [L-PERIPH +] constraint on the left-located constituents. For example, the emphasized constituents in (26)[12] can be evaluated as containing *contrast-focus* in some context.

---

[10]According to Gracheva (2013), there is one more constraint on *li*; the sentence should be interrogative. That is, the sentential force of the utterance is conditioned as [SF *ques*] by *li*.

[11]Gracheva (p.c.) would not say *li* is definitely a contrastive-focus marker. She thinks more research is necessary, which would be a topic for an entire dissertation. Suffice it to say that the clitic *li* is possibly and partially related to contrastive-focus meanings.

[12]Those examples are provided by Varya Gracheva (p.c.).

(26) a. **Na rynke** li Ivan kupil popugaya?
On market-PREP li Ivan-NOM buy-PST.SG.M parrot-ACC
'Was it **in the market** that Ivan bought a parrot?'

b. **Govoriashego** li popugaja kupil Ivan?
Talking-SG.MASC.ACC li parrot-ACC buy-PST.SG.M Ivan-NOM
'Did Ivan buy a **talking** parrot?' [rus]

Gracheva provides other clitics possibly signaling specific information structure meanings in Russian. *-to*, *že*, and *ved'* are also related to periphery of the modificands in a similar way.[13] That implies L-PERIPH and R-PERIPH play an important role in configuring information structure in Russian-like languages.

L-PERIPH can also be used for imposing a restriction on the position of topics in topic-first languages (e.g. Japanese and Korean). The leftmost (i.e. sentence-initial) and probably topic-marked constituent in topic-first languages should involve [L-PERIPH +], disallowing the appearance of any other constituents on its left side. One exceptional case to this restriction is *frame-setting*, because a series of constituents functioning as frame-setters can show up in the sentence-initial position. This is a phenomenon which seems language-universal (Li and Thompson, 1976; Chafe, 1976; Lambrecht, 1996). Now that *topic-comment* and its subtypes have an additional constraint on L-PERIPH, the AVMs offered in §9.4 (p. 192) are extended as follows.[14]

(27) a. $\begin{bmatrix} \textit{topic-comment} & \\ \text{L-PEPIPH} & + \\ \text{MKG} & \textit{tp} \\ \text{NHD} & \begin{bmatrix} \text{MKG} & \textit{tp} \\ \text{L-PERIPH} & + \end{bmatrix} \end{bmatrix}$

b. $\begin{bmatrix} \textit{non-frame-setting} & \\ \text{HD} & \begin{bmatrix} \text{MKG} & \textit{fc-only} \\ \text{L-PERIPH} & - \end{bmatrix} \end{bmatrix}$

[13]Russian has been known to employ pragmatically conditioned word order (Rodionova, 2001). In that case, the pragmatic condition largely refers to information structure. For the reason, it seems that a variety of means are used for expressing information structure in Russian. It looks relatively complex how sentences are configured in Russian, heavily depending on information structure.

[14]Note that they are a short version of AVMs. L-PERIPH is a feature under SYNSEM in the actual description in TDL.

The more specific rules that inherit from these will be presented with reference to specific constructions of articulating information structure (e.g. scrambling in Japanese and Korean) in Chapter 12.

L-PERIPH also plays a role in focus/topic-fronting constructions. If a language places focused constituents in the clause-initial position and also has the topic-first restriction, the fronted constituents are associated with *focus-or-topic*, as suggested before. Yiddish typically exhibits such a behaviour as repeatedly presented in (28). Yiddish is a V2 language with a neutral word order of SVO, in which the verb (i.e. the syntactic head in a sentence) should occur in the second position of the linear order (Jacobs, 2005). Therefore, if the object is focused in Yiddish, the linear order, as exemplified (28b), should be OVS, not OSV. The same goes for sentences in which adverbials are fronted as shown in (28c-d).[15]

(28) a. Der lerər šrajbt di zacn mit krajd afn tovl.
The teacher writes the sentences with chalk on the blackboard (neutral) [ydd]

b. Di zacn šrajbt der lerər mit krajd afn tovl.
the sentences writes the teacher with chalk on the blackboard
'It's the sentence (not mathematical equations) that the teacher is writing with chalk on the blackboard.' [ydd]

c. mit krajd šrajbt der lerər di zacn afn tovl.
with chalk writes the teacher the sentences on the blackboard
'It's with chalk (not with a crayon) that that the teacher is writing the sentence on the blackboard.' [ydd]

d. afn tovl šrajbt der lerər di zacn mit krajd.
on the blackboard writes the teacher the sentences with chalk
'It's on the blackboard (not the notepad) that that the teacher is writing the sentence with chalk.' [ydd] (Jacobs, 2005, p. 224)

As discussed before (§4.4), (28a) in which the subject occurs sentence-initially may sound ambiguous. This is like ordinary focus/topic-fronting constructions in other languages. However, in any cases the focused or topicalized constituent should be first, and is constrained as [L-PERIPH +].

---

[15] In fact, the translations in (28) provided by Jacobs (2005) follow the notion that the focused XPs in cleft constructions exhibit exhaustive inferences (i.e. contrastive meaning). Chapter 12 addresses this interpretation in detail (§12.4).

10.3.2 Lightness

Preverbal and postverbal focus positions are constrained by LIGHT, which already exists in the Matrix core (i.e. `matrix.tdl`) in order to distinguish words from phrases. Using LIGHT for discriminating words and phrases is inspired by the "Lite" feature Abeillé and Godard (2001) suggest.[16] LIGHT is a feature under SYNSEM because it is also a flag feature.

[LIGHT +] is attached to words, while [LIGHT –] is attached to phrases. The value of LIGHT, whose type is *luk*, is sometimes co-indexed with that of HC-LIGHT originally taken from the ERG. HC stands for Head-Complement. The purpose of using HC-LIGHT is to indicate whether a *head-comp-phrase* projected from a head is regarded as light or heavy. If an element in a parse tree has [LIGHT +], it indicates the element has not yet been pumped into an instance of a phrasal type. As for verbal nodes in parse trees, the distinction between V and VP is naturally made by the value of LIGHT.

Preverbal and postverbal foci are always realized as *narrow-focus* presented in the previous chapter (p. 190). In addition to this constraint, I argue that preverbal and postverbal focus can be combined with only Vs with [LIGHT +]. Basque, for instance, is known for the preverbal focus position. In the Basque sentence below, *Jonek* 'Jon' is signaled as *focus*, which is immediately followed by *irakurri du* 'read has'.

(29) Eskutitza, Jonek irakurri du
letter Jon read has
'JON has read the letter.' [eus] (Ortiz de Urbina, 1999, p. 312)

My analysis of the sentence is as follows:[17] The auxiliary verbal item *du* 'has' takes *irakurri* 'read' as its complement, but they are still regarded as behaving as a word. That is, *irakurri du* has [HC-LIGHT +], which shares a coreference with LIGHT. After *irakurri* and *du* form a *head-comp-phrase* together, *Jonek* is combined with *irakurri du* including [LIGHT +], which constitutes a *head-subj-phrase*. Because the *head-subj-phrase* (i.e. *Jonek irakurri du*) now has [LIGHT –], no

[16] Crowgey and Bender (2011, p. 54) also make use of this feature to impose a constraint on negation in Basque: "The feature LIGHT is defined on synsems with a value *luk*. Lexical items are [LIGHT +], while phrases are [LIGHT –]. This stipulation ensures that the verbal complex rule applies before the auxiliary picks up any arguments in any successful parse."

[17] This is not necessarily analogous to proposal of Ortiz de Urbina (1999), whose grammatical framework is movement-based.

more preverbal foci can take place. Crowgey and Bender (2011) provide a similar analysis to mine. They argue that Basque has a constraint (30) with respect to the variation of word order.

(30) If the lexical verb is to the left of the auxiliary, then the lexical verb must be left-adjacent to the auxiliary. (Crowgey and Bender, 2011, p. 49)

This constraint explains ungrammaticality of (31), in which the main verb and the auxiliary are not adjacent to each other. That implies that a main verb plus an auxiliary (e.g. *irakurri du*) behave as a single cluster of verb, featured as [LIGHT +].

(31) *Liburu irakurri Mirenek du
book.ABS.SG read.PERF Mary.ERG.SG 3SGO.PRES.3SGA
'Mary has read a book.' [eus] (Crowgey and Bender, 2011, p. 48)

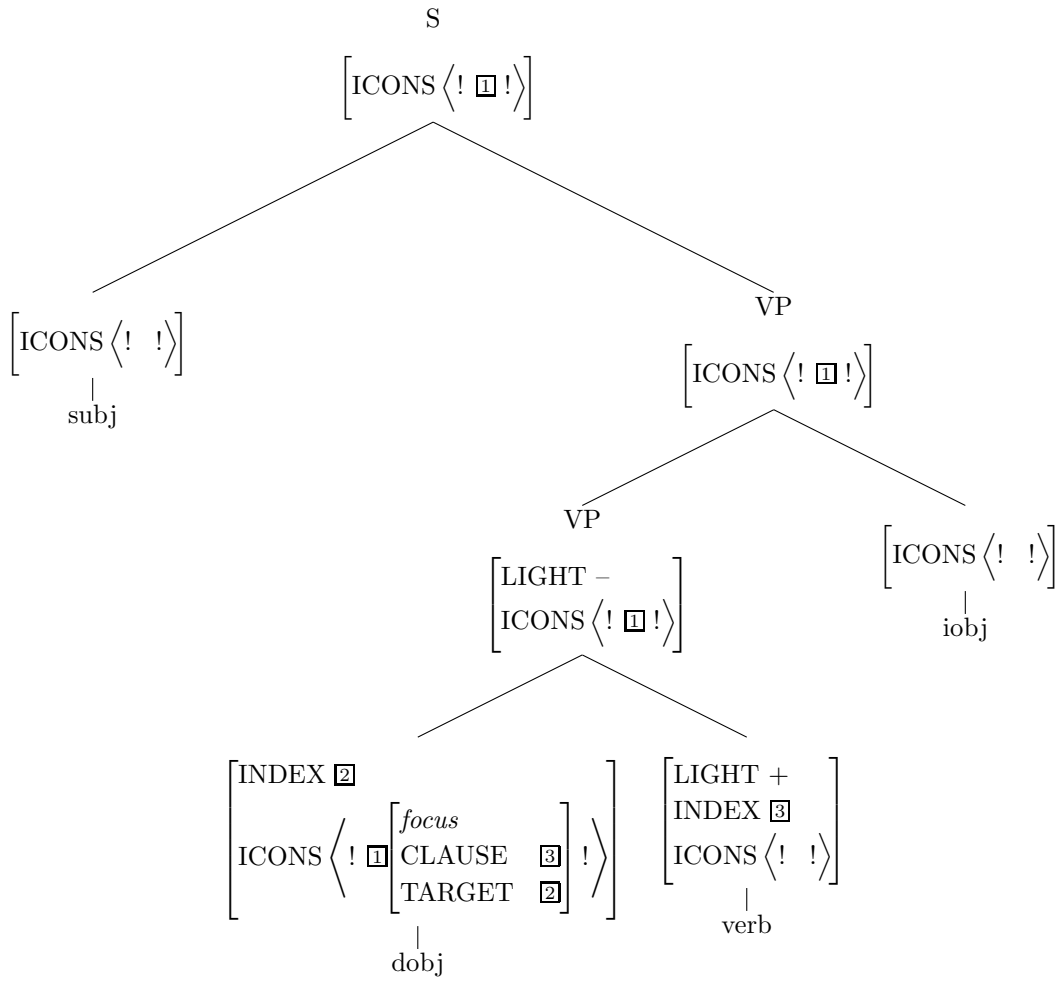
For more explanation about constraints on preverbal/postverbal foci, two pseudo sentences in Language A (presented in the previous chapter (p. 189)) can be instantiated as shown in (32). If Language A, whose word-order properties are repeated in (33), has ditransitive verbs, and the ordinary order between objects is [indirect object (*iobj*) + direct object (*dobj*)], then (32a) is in the basic word order.

(32) a. subj verb iobj dobj. (neutral)
b. subj dobj verb iobj. (focus on dobj)

(33) a. [Language A] employs SVO as its basic word order.
b. Focused constituents in [Language A] are realized in the immediate preverbal position.
c. Additionally, there is an optionally used accent, which expresses focus.

A sample derivation for (32b) in which the direct object is focused and preverbal is illustrated below.

(34)



The focused item *dobj*, which is not *in situ*, is combined with *verb* involving [LIGHT +] before anything else. They constitutes a *head-comp-phrase*, which has now [LIGHT –]. Next, the VP takes *iobj*, which is *in situ*, as the second complement, and forms another VP as *head-comp-phrase*. Finally, the subject is combined with the second VP into a *head-subj-phrase*. In this case, the first and the second *head-comp-phrase* are realized as two different rules. The first one puts constraints on both the NON-HEAD-DTR (e.g. *dobj* in (32b)) and the HEAD-DTR (e.g. *verb*); an information structure value *focus* is assigned to the NON-HEAD-DTR and the HEAD-DTR is in need of [LIGHT +]. The second one does not signal any specific values of information structure, but requires [LIGHT –] of the HEAD-DTR. Notably, this analysis is not applied to sentences in the neutral word order. For example, *subj* in (32a) is in the immediately preverbal position, but it is *in situ* in the neutral word order. Thus, it is not analyzed as containing *focus*.

10.3.3 Phonological Structure

The phonological structure proposed in Bildhauer (2007) is not completely applied to the customization system as is, because the phonological behaviours in many languages remain hitherto unknown. The structure itself is implemented into `matrix.tdl` in TDL, but any other further rules are not implemented in this dissertation. The set of the phonology-related features are introduced in `matrix.tdl` so that developers of Matrix-derived grammars could use them in future.

Prosodic patterns in Japanese and Korean, with respect to information structure, have been substantially revealed by phonetic experiments (Jun et al., 2007; Ueyama and Jun, 1998; Jun and Lee, 1998). Prosodic behaviours of information structure in Spanish are well-summarized in Bildhauer (2007) as well. Yet, this dissertation has little interest in them, because the main purpose of the current work is to create a grammar library for information structure in the LinGO Grammar Matrix system. The system is built for text-based processing, and has not yet reflected phonological information in a significant manner. It is left to the future research to implement prosodic rules in Japanese, Korean, Spanish, and other languages.

Nevertheless, Bildhauer (2007) proposes four levels of phonological structure, consisting of (i) prosodic word, (ii) phonological phrase, (iii) intonational phrase, and (iv) phonological utterance, and two intonational typed feature structures, including (v) pitch accents, and (vi) boundary tones. Among them, this dissertation is not concerned with the first three structures, because it is difficult to obtain an acoustic system to resolve the prosodic levels reliably. In other words, the rules presented in Chapter 8 (p. 160) are tentatively disregarded in the current work. The last three are largely related to focus projection, but the rules for them are also altered to be suitable for implementation in DELPH-IN grammars. The altered rules, such as the focus-prominence rule and focus-projection rule, are presented in Chapter 13 which is especially concerned with how to calculate the spreading of focus.

10.3.4 Sign

A revised version of *sign*, which is now defined in the LinGO Grammar Matrix system, is presented in Appdendix A.

10.4 Sample Derivations

This section provides sample derivations, which briefly show how information structure works with ICONS in several different types of languages. The languages that this section presents are English, Japanese, Korean, and Russian. The type of the ICONS-KEY value of a constituent, which points to an element of the ICONS list, can be constrained by (i) accents responsible for information structure meanings, (ii) lexical rules attaching information structure marking morphemes, (iii) particles like Japanese *wa* combining as heads or modifiers with NPs, and/or (iv) phrase structure rules corresponding to distinguished positions.

10.4.1 English

Pitch accents primarily serve to express information structure meanings in English as shown in (35).

(35) a. The DOG barks.
b. The **dog** barks.

In other words, English imposes a constraint on the A and B accents. In the current work, they are hypothetically realized as suffixes (e.g. '-a', '-b'), whose lexical rules are as follows respectively. That is, (35a-b) are actually encoded into *The dog-a barks.* and *The dog-b barks.* respectively as an input string for parsing and an output string from generation.

First of all, UT|DTE, adapted from Bildhauer (2007), aims to calculate focus projection in Chapter 13 (§13.2). The current work gives the value of UT|DTE a coreference with that of MKG|FC, because focus projection is not always licensed by prosodic means across languages (Choe, 2002). That means MKG|FC is responsible for spreading the focus domain to the larger phrases, and the value should be the same value as the value of UT|DTE in English.

(36) *lex-rule* →
$$\begin{bmatrix} \text{UT} \mid \text{DTE} & \boxed{1} \\ \text{MKG} \mid \text{FC} & \boxed{1} \end{bmatrix}$$

Next, (37a-b) are the lexical rules of the A and B accents. Each of their PA values of them, taken from Bildhauer's hierarchy (14), stands for H* and L+H* in the ToBI format, respectively. MKG for the A-accent is valued as *fc-only* and accordingly UT|DTE is also valued as +. MKG for the B-accent has *tp*, whose FC remains underspecified and has a coreference with UT|DTE. Because the

A and B accents exhibit which information structure meaning are conveyed in a fairly direct way, they include *semantic-focus* and *contrast-or-topic* into the list of ICONS. What is to be noted is that the value of MKG|FC and its co-indexed value of DTE are only related to marking of information structure. Even though they have a plus value, that does not necessarily mean that the constituent conveys meaning of focus. In other words, [MKG|FC +] indicates only F(ocus)-marking.

(37) a.

fc-lex-rule →

$$\begin{bmatrix} \text{UT}\,|\,\text{DTE} & + \\ \text{PA} & \textit{high-star} \\ \text{MKG} & \textit{fc-only} \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{C-CONT}\,|\,\text{ICONS} & \left\langle ! \; \boxed{2} \begin{bmatrix} \textit{semantic-focus} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \\ \text{DTR} & \begin{bmatrix} \text{HEAD } +nv \end{bmatrix} \end{bmatrix}$$

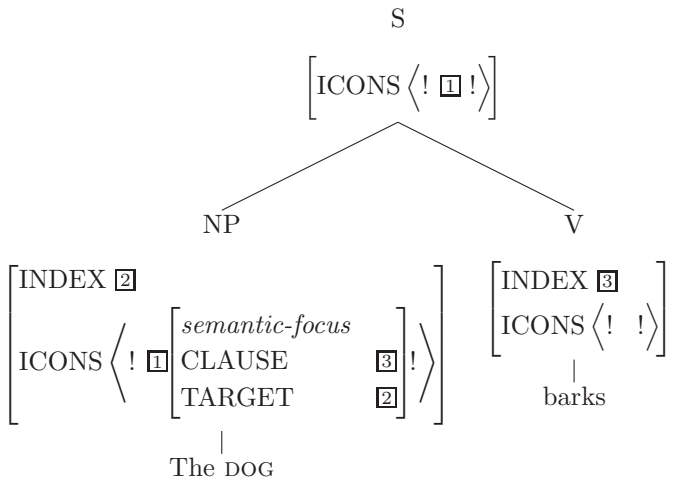
b.

tp-lex-rule →

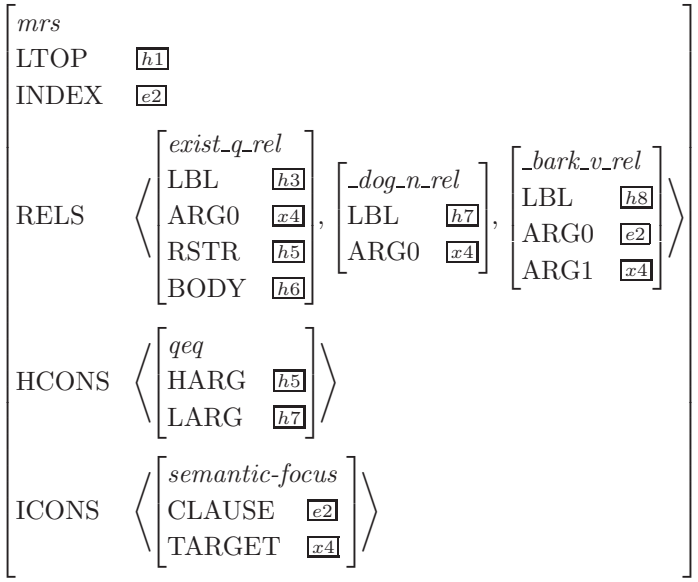
$$\begin{bmatrix} \text{UT}\,|\,\text{DTE} & \textit{luk} \\ \text{PA} & \textit{low-high-star} \\ \text{MKG} & \textit{tp} \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{C-CONT}\,|\,\text{ICONS} & \left\langle ! \; \boxed{2} \begin{bmatrix} \textit{contrast-or-topic} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \\ \text{DTR} & \begin{bmatrix} \text{HEAD } \textit{noun} \end{bmatrix} \end{bmatrix}$$

Building upon these rules, (35a) in which DOG bears the A-accent for expressing *semantic-focus* is constructed as (38). The corresponding MRS and the dependency graph standing for the information structure of (35a) are presented in (39). The utterance forms a clause, and the clausal type (i.e. *declarative-clause* imposing [SF *prop-or-ques*]) is inherited by the phrase structure type (i.e. *head-subj-phrase*). Applying this constraint on *clause* as presented before, the CLAUSE-KEY of the NP *the dog* points to the INDEX of the HEAD-DTR (i.e. the verb *barks*). Herein, the TARGET of the NP is co-indexed with its INDEX, and the CLAUSE is co-indexed with the INDEX of the verb. The TARGET and CLAUSE of *barks* is recursively linked. Each value in the *diff-list* of ICONS is collected into higher phrases.

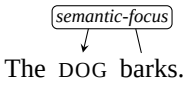
(38)



(39) a.



b.



10.4.2 Japanese and Korean

In Japanese and Korean, the distinction between lexical markers (i.e. *ga vs. wa* in Japanese and *i/ka vs.* (*n*)*un* in Korean) are responsible for delivering different meanings in information structure. Note that case-marking NPs in Japanese and Korean do not always correspond to A-accented NPs in English: NPs with *ga* or *i/ka* in Japanese and Korean basically involve *non-topic*. On the other hand, A-accented NPs in English are straightforwardly interpreted as containing meaning of semantic (i.e., non-contrastive) focus.

(40) a. inu ga hoeru.
dog NOM bark [jpn]
b. inu wa hoeru.
dog WA bark
'The dog barks.' [jpn]

(41) a. kay-ka cic-ta.
dog-NOM bark-DECL [kor]
b. kay-nun cic-ta.
dog-NUN bark-DECL
'The dog barks.' [kor]

In Japanese, *ga* and *wa* are adpositions, following the convention in Jacy (Siegel and Bender 2002). The information structure value that the null marker assigns to constituents is in line with Yatabe (1999). As for the null marker Ø, the grammar developed here uses a lexical rule,[18] which is different from Yatabe's proposal (i.e. the null marking system as particle ellipsis). Instead, a lexical rule places a constraint on the null marker (i.e. in `lrules.tdl`). Nonetheless, this dissertation agree with Yatabe's argument on the information structure meaning of null-marked constituents in Japanese (and Korean). Yatabe claims that *ga* cannot be dropped when the *ga*-marked expression is focused, which implies null-marked phrases (mostly NPs) should be evaluated as containing *non-focus*. The AVMs for them are provided in (42). Note that the values of MKG are not in accordance with the values of their complement's ICONS-KEY, and the element of ICONS specifies the relation of the complement, not the adposition itself.

(42) a.
$$\textit{nom-marker} \rightarrow \begin{bmatrix} \text{STEM} & \langle \textit{ga} \rangle \\ \text{CASE} & \textit{nom} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{MKG} & \textit{unmkg} \\ \text{COMPS} & \left\langle \left[\text{INDEX } \boxed{1} \right] \right\rangle \\ \text{ICONS} & \left\langle ! \ \boxed{2} \begin{bmatrix} \textit{non-topic} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

[18]This instance can be a daughter of a unary rule (i.e. *bare-np-phrase*) which promotes the word to a phrase.

b.

wa-marker →

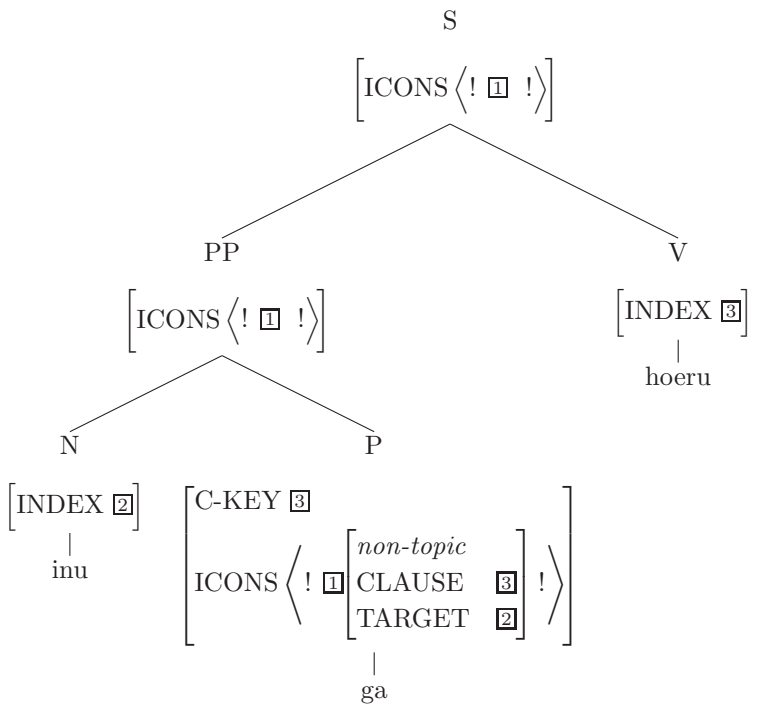
$$\begin{bmatrix} \text{STEM} & \langle \textit{wa} \rangle \\ \text{CASE} & \textit{case} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{MKG} & \textit{tp} \\ \text{COMPS} & \left\langle \left[\text{INDEX } \boxed{1} \right] \right\rangle \\ \text{ICONS} & \left\langle ! \ \boxed{2} \begin{bmatrix} \textit{contrast-or-topic} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

c.

null-lex-rule →

$$\begin{bmatrix} \text{CASE} & \textit{case} \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{MKG} & \textit{unmkg} \\ \text{C-CONT} \mid \text{ICONS} & \left\langle ! \ \boxed{2} \begin{bmatrix} \textit{non-focus} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

(43)

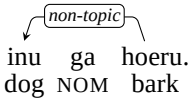


The sample derivation for (40a), in which *inu* 'dog' is combined with *ga* responsible for *non-topic*, is illustrated in (43).[19] The corresponding MRS representation is given in (44a), and the

[19]In HPSG-based analyses for syntactic configuration in Japanese, PPs are important. That means the case markers

graphical version is in (44b).

(44) a.
$$\begin{bmatrix} mrs \\ \text{LTOP} \quad \boxed{h1} \\ \text{INDEX} \quad \boxed{e2} \\ \text{RELS} \quad \left\langle \begin{bmatrix} \_inu\_n\_rel \\ \text{LBL} \quad \boxed{h3} \\ \text{ARG0} \quad \boxed{x4} \end{bmatrix}, \begin{bmatrix} exist\_q\_rel \\ \text{LBL} \quad \boxed{h5} \\ \text{ARG0} \quad \boxed{x4} \\ \text{RSTR} \quad \boxed{h6} \\ \text{BODY} \quad \boxed{h7} \end{bmatrix}, \begin{bmatrix} \_hoeru\_v\_rel \\ \text{LBL} \quad \boxed{h8} \\ \text{ARG0} \quad \boxed{e2} \\ \text{ARG1} \quad \boxed{x4} \end{bmatrix} \right\rangle \\ \text{HCONS} \quad \left\langle \begin{bmatrix} qeq \\ \text{HARG} \quad \boxed{h6} \\ \text{LARG} \quad \boxed{h3} \end{bmatrix} \right\rangle \\ \text{ICONS} \quad \left\langle \begin{bmatrix} non\text{-}topic \\ \text{CLAUSE} \quad \boxed{e2} \\ \text{TARGET} \quad \boxed{x4} \end{bmatrix} \right\rangle \end{bmatrix}$$

b.


The CLAUSE-KEY of the nominative marker *ga* is identified with its own ICONS-KEY|CLAUSE. Second, when the *head-comp-phrase* combines *inu* and *ga*, the ICONS-KEY|CLAUSE of *inu* is identified with the CLAUSE-KEY of *ga*. The ICONS-KEY of *ga* is passed up to the mother (Semantic Inheritance Principle). When the *head-subj-phrase* combines *inu ga* and *hoeru*, the ICONS-KEY|CLAUSE of the subject *inu ga* is identified with the INDEX of *hoeru*.

With respect to Korean, this dissertation basically assumes that the marking systems (e.g. *wa* or (*n*)*un*-marking, case-marking, and null-marking) in Japanese and Korean share the same properties in terms of information structure, given that counterexamples to this assumption are very rare. The counterexample in which *wa* and (*n*)*un* in two languages show different behaviour is reported in some Japanese dialects.[20] For example, Tokyo dialect does not show any difference from Korean with respect to using topic markers, but Kansai dialect sometimes makes a subtle difference in *wa*-marking from (*n*)*un*-marking in Korean. Despite the similarity in syntax, phrase structures in Korean

(e.g. *ga* for nominatives), the *wa* marker, and a null-marker are adpositions that take the NPs that they are attached to as the complement, and constitute PPs. The reason why the combination between NPs and the markers should be PPs in Japanese has been already explained in several previous HPSG-based studies (Gunji, 1987; Siegel, 1999; Yatabe, 1999).

[20] Yo Sato, p.c.

have been analyzed differently from those in Japanese. In a nutshell, *ga* and *wa* in Japanese are dealt with as words, whereas *i/ka* and (*n*)*un* in Korean are treated as suffixes. Because postpositions are crucially employed in the building blocks of a clause in most analyses of Japanese syntax (Sato and Tam, 2012), the combination between nouns and these markers (e.g. *ga*, *wa*, etc.) forms a PP, rather than an NP. Kim and Yang (2004), in contrast, regard the lexical markers in Korean (e.g. *i/ka*, (*n*)*un*, etc.) as affixes, rather than adpositions.[21] That means that the combination between nouns and their markers still remains as an NP. This dissertation respects the two different analyses of these languages, as they exist. Technically speaking in the context of grammar engineering, the adpositions *ga* and *wa* in Japanese are treated as independent lexical entries, while the morphemes *i/ka* and (*n*)*un* in Korean are dealt with by lexical rules. Accordingly, the derivation of an NP plus *i/ka* or (*n*)*un* is created at the lexical level. The inflectional rules are as follows.

(45) a.

$$\textit{nom-lex-rule} \rightarrow \begin{bmatrix} \text{INFOSTR-FLAG} & + \\ \text{CASE} & \textit{nom} \\ \text{MKG} & \textit{unmkg} \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{C-CONT} \mid \text{ICONS} & \left\langle ! \; \boxed{2} \begin{bmatrix} \textit{non-topic} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \\ \text{DTR} & \left[\text{INFOSTR-FLAG } - \right] \end{bmatrix}$$

b.

$$\textit{nun-lex-rule} \rightarrow \begin{bmatrix} \text{INFOSTR-FLAG} & + \\ \text{CASE} & \textit{case} \\ \text{MKG} & \textit{tp} \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{C-CONT} \mid \text{ICONS} & \left\langle ! \; \boxed{2} \begin{bmatrix} \textit{contrast-or-topic} \\ \text{TARGET } \boxed{1} \end{bmatrix} ! \right\rangle \\ \text{DTR} & \left[\text{INFOSTR-FLAG } - \right] \end{bmatrix}$$

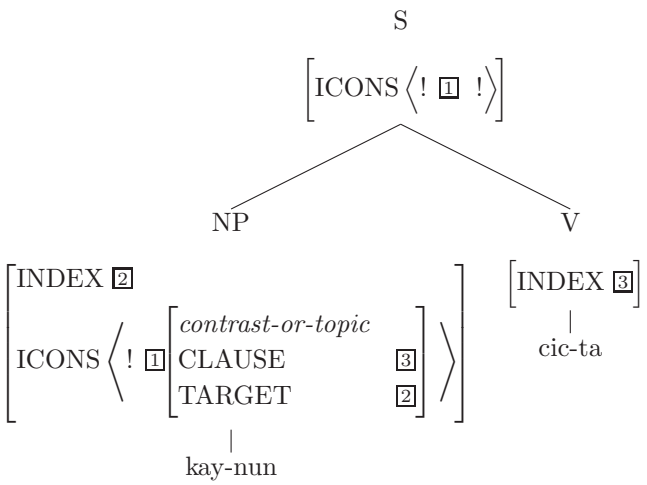
[21] Another approach to Korean postpositions is given in Ko (2008), who insists that Korean postpositions should be analyzed as clitics attaching to either the preceding lexical item or weak syntactic heads sharing syntactic feature values of the complement phrase.

c.

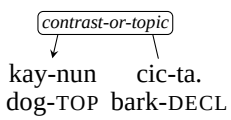
$$\textit{null-lex-rule} \rightarrow \begin{bmatrix} \text{CASE} & \textit{case} \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{MKG} & \textit{unmkg} \\ \text{C-CONT}\,|\,\text{ICONS} & \left\langle\, !\ \boxed{2}\begin{bmatrix}\textit{non-focus}\\ \text{TARGET}\ \boxed{1}\end{bmatrix}!\,\right\rangle \end{bmatrix}$$

Note that they are in complementary distribution. They share the same slot in the morphological paradigm. Additionally, *null-lex-rule* is required to constrain information structure of null-marked constituents. The AVM is the same as that in Japanese, because Yatabe (1999)'s analysis of how the null marker contributes to information structure is straightforwardly applied to Korean (in my intuition), although the null marker is not regarded as an ellipsis in this dissertation. Though they are realized as inflectional rules in the morphological paradigm, the values of MKGs are identical to those in Japanese, and the elements in the ICONS lists are the same as the values of COMPS|ICONS-KEY in Japanese. Altogether, analyzing Korean sentences is syntactically similar to those in English, and informatively similar to those in Japanese. The sample derivation for (41b), in which the (*n*)*un*-marked *kay* 'dog' is associated with *contrast-or-topic*, is sketched out in (46a). The graphical representation is also shown in (46b).

(46) a.



b.



10.4.3 Russian

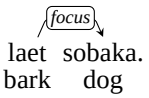
Russian employs its relatively free word order to mark focus with clause-final constituents bearing non-contrastive focus (i.e. *semantic-focus*) (Neeleman and Titov, 2009). Notably, constituents *in situ* can also convey focus meaning, if they involve a specific prosody for expressing focus. Thus, in (47a) in the basic word order, the focus can fall on either the subject *sobaka* or the verb *laet*, or both (i.e. *all-focus*).[22]

(47) a. Sobaka laet.
dog bark
'The dog barks.' [rus]

b. Laet sobaka.
bark dog
'The DOG barks.' [rus]

Headed rules can have subtypes which handle information structure differently, resolving the type of an ICONS element or leaving it underspecified. For example, the Russian allosentences of (47) are instances of *head-subj-phrase*, but the first one (*sobaka laet*), in which the subject is *in situ*, is licensed by a subtype that does not resolve the ICONS value, while the second one (*laet sobaka*), in which the subject is marked through being postposed, is licensed by the one which does. Hence, the subject *in-situ* is specified as *info-str* (i.e. underspecified), whereas the overtly postposed subject is specified as *focus*. Consequently, (47a-b) are graphically represented as follows.

(48) a.
sobaka laet.
dog bark

b.


In order to construct a derivation tree for (47b) whose word order in not neutral, it is necessary to implement several additional devices: an additional flag feature [INFOSTR-FLAG *luk*] , a unary phrase structure rule *narrow-focused-phrase*, and *head-subj-phrase*. First, a flag feature INFOSTR-FLAG is immediately under SYNSEM like L/R-PERIPH and LIGHT.[23] This flag fea-

[22]Russian also employs prosody to signal focus, too. They could be modelled in the same way of using '-a' and '-b' in English. Nonetheless, they are not considered here for ease of explanation, and instead this subsection concentrates on syntactic positioning.

[23]Although they are housed in the same position, not all languages use INFOSTR-FLAG, while L/R-PERIPH and LIGHT are commonly used in human language. Thus, this flag feature is not included in the basic *synsem*.

ture serves to indicate whether the current phrase is information structure-marked and includes an ICONS element. *Narrow-focused-phrase* takes only a constituent with [INFOSTR-FLAG –] as its daughter, and assign [INFOSTR-FLAG +] to itself. *Narrow-focused-phrase* introduces an *focus* element into the C-CONT|ICONS. Finally, *head-subj-phrase* in Russian inherits from both *basic-head-subj-phrase* and *head-initial*. It takes a constituent that has both [INFOSTR-FLAG +] and [R-PERIPH +] as its non-head-daughter. These indicate that the subject is marked with respect to information structure and it is also clause-final. *Narrow-focused-phrase* and *head-subj-phrase* are represented in the following AVMs. The derivation tree for (47b) is sketched out in (50).

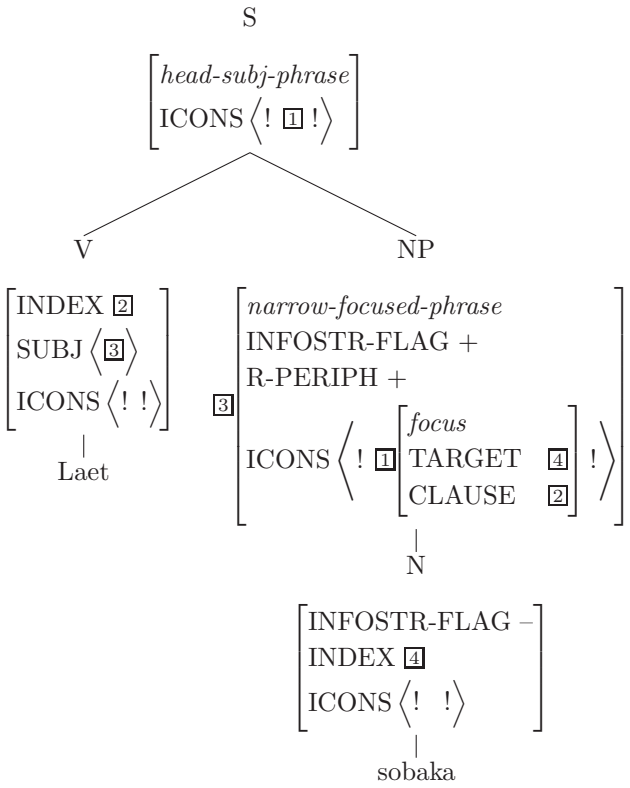
(49) a.

$$\begin{bmatrix} \textit{narrow-focused-phrase} & \\ \text{INFOSTR-FLAG} & + \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{HD}\,|\,\text{INFOSTR-FLAG} & - \\ \text{C-CONT}\,|\,\text{ICONS} & \left\langle ! \ \boxed{2}\begin{bmatrix}\textit{focus} \\ \text{TARGET}\ \boxed{1}\end{bmatrix} ! \right\rangle \end{bmatrix}$$

b.

$$\begin{bmatrix} \textit{head-subj-phrase} & \\ \text{NHD} & \begin{bmatrix}\text{INFOSTR-FLAG} + \\ \text{R-PERIPH} +\end{bmatrix} \end{bmatrix}$$

(50)



10.5 Summary

This chapter addressed specifics of implementing ICONS into lexical and phrasal types for constraining information structure. Lexical types inherit from either four types of *icons-lex-item*: *no-icons-lex-item*, *basic-icons-lex-item*, *one-icons-lex-item*, and *two-icons-lex-item*. Both *no-icons-lex-item* and *basic-icons-lex-item* have an empty ICONS list, and *no-icons-lex-item* additionally has [MKG [FC *na*, TP *na*]]. This contraint indicates that lexical entries which inherit from *no-icons-lex-item* cannot be marked with respect to information structure; for instance, relative pronouns, expletives, etc. Nominal items normally inherit from *basic-icons-lex-item*, while verbal items can inherit from one of them depending on how many clauses are subordinated to the verbal item. Adpositions and determiners inherit from either *basic-icons-lex-item* or *one-icons-lex-item*, adverbs inherit from *basic-icons-lex-item*, and syncategorematic items inherit from *no-icons-lex-item*. Conjunctions may or may not introduce a *topic* value into ICONS depending on which type of adverbial clauses they involve. There are five phrasal types: *basic-head-subj-phrase*, *basic-head-comp-phrase*, *basic-head-spec-phrase*, *basic-head-mod-phrase-simple*, and *basic-head-filler-phrase*. *Basic-head-subj-phrase* is crucial in that CLAUSE is identified in the construction. *Basic-head-mod-phrase-simple* and *head-filler-phrase* have some extra constraints to specify which element is linked to which clause. There are three additional constraints for elaborating on properties of information structure: L/R-PERIPH, LIGHT, and PHON. L/R-PERIPH constrain the clause-initial/final constituents with respect to information structure, and likewise LIGHT is used for constraining preverbal and postverbal constituents. PHON is added into `matrix.tdl` for future work with acoustic resolution systems.

Chapter 11

MULTICLAUSAL CONSTRUCTIONS

As discussed previously, one of the main motivations of using ICONS (Individual CONstraints) is to capture binary relations across clauses. This chapter addresses two points of clausal constraints in representation of information structure. First, this chapter looks into how the relation between matrix clauses and non-matrix clauses are represented via ICONS. Since ICONS is a way of representing a relation between an individual and a clause that the individual belongs to, it is also needed to identify the relation between two clauses in a single sentence. Second, this chapter also looks at which restrictions non-matrix clauses have with respect to information structure. Many previous studies argue that information structure in non-matrix clauses is differently formed from that in matrix clauses. According to Kuno (1973), *wa* in Japanese is seldom used in relative clauses. Similarly, (1a) indicates that English normally disallow left dislocation in relative clauses.

(1) a. *A man who your book$_i$ could buy it$_i$.

b. Un uomo che, il tuo libro$_i$, lo$_i$ potrebbe comprare.
A man who, your book, could buy it. [ita] (Rizzi, 1997, p. 306)

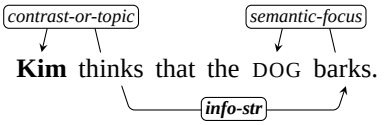
However, this restriction is language-specific. Embedded clauses in some languages exhibit properties of root clauses as shown in (1b). Furthermore, Haegeman (2004) argues that topic fronting can occur in non-root clauses even in English under certain circumstances, such as adversative clauses, *because* clauses, and sometimes conditional clauses. This restriction has to do with so-called embedded root phenomena (Heycock, 2007). Root clause might be simply defined as not embedded clause, but there exist some counterexamples to such a definition. It is known that root phenomena has an effect on the appearance of topics in embedded clauses: Portner and Yabushita (1998) insist that topic should only be interpretable with the wide scope on the root clause. For instance, OSV word order constructions in English and *wa*-marking in Japanese typically exhibit a root effect. They tend not to appear in non-root clauses.

Non-matrix clauses can be roughly classified into at least three types. These are complement clauses (§11.1), relative clauses (§11.2), and adverbial clauses (§11.3). Each section in this chapter looks into linguistic factors that have an influence on information structure in each clausal type, and provides an HPSG/MRS-based analysis of the clausal type.

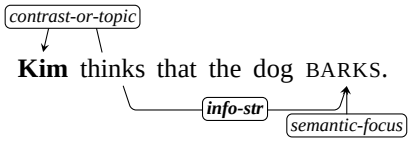
11.1 Complement Clauses

The issues that this section addresses include (i) how components of information structure are constituted in complement clauses, and (ii) how assignment of information structure values is conditioned in different complement clauses. Dependency graphs in (2) show the basic mechanism of indexing between TARGETs and CLAUSEs in multiclausal constructions. In accordance with the AVMs presented hitherto, (2a-c) are the representations of a sentence *Kim thinks that the dog barks*. In (2a), the subject in the matrix clause is B-accented, and the subject in the embedded clause is A-accented. Hence, they are assigned *contrast-or-topic* and *semantic-focus*, respectively (p. 219). The arc from the main verb *thinks* to the verb in the embedded clause *barks* comes from the lexical information of the main verb, which inherits from *one-icons-lex-item*. That is, *thinks* has one inherent element in the list of ICONS, which links its its own INDEX to INDEX of *barks*).[1]

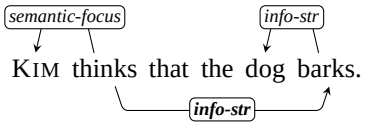
(2) a.



b.



c.



11.1.1 Background

Topic can sometimes take place in complement clauses, largely depending on the characteristics of the predicate in the main clause. The properties that license topics to appear in complement clauses include speech acts, semi-factives, quasi-evidentials (Roberts, 2011). Maki et al. (1999)

[1]These underspecified *info-str* elements are not fully desirable. An analysis that allows specific ICONS elements relating the two clauses but does not entail inserting these underspecified ones is left for future work.

argue that topic fronting in embedded clauses in English and appearance of *wa* in embedded clauses in Japanese commonly shows four characteristics as given in (3).[2]

(3) a. Embedded topicalization is possible in complement clauses of bridge verbs (e.g. *believe*, *sinziteiru* 'believe').

b. Embedded topicalization is possible in interrogative clauses.

c. Embedded topicalization is impossible in complement clauses of factive verbs (e.g. *regret*, *kookaisiteiru* 'regret') and noun-complement clauses.

d. Embedded topicalization is impossible in an adjunct clause and in a sentential subject. (Maki et al., 1999, p. 8–10)

Heycock (2007), in a similar vein, elaborates on cases in which embedded clauses have a root function. According to Heycock's analysis, the main criterion to distinguish whether sentential subjects/complements exhibit root phenomena or not is assertion. In other words, whether a topic can occur in subordinate clauses is influenced by whether the subordinate clause is asserted. A five-way division of predicates is offered as follows.

(4) a. Class A predicates (e.g. "say", "report", "be true", "be obvious"). The verbs in this group are all verbs of saying. Both the verbs and the adjectives in this group can function parenthetically, in which case the subordinate clause constitutes the main assertion of the sentence. It is claimed however that if the subordinate clause occurs in subject position (as in, e.g. "*That German beer is better than American beer is true*") it is not asserted.

b. Class B predicates (e.g. "suppose", "expect", "it seems", "it appears"). In this group also the predicates can function parenthetically, and in this case the subordinate clause is asserted. The distinction between this group and Group A is not made entirely clear, although it is noted that Class B predicates allow "Neg raising" and tag questions based on the subordinate clause.

c. Class C predicates (e.g. "be (un)likely", "be (im)possible", "doubt", "deny") have complements which are not asserted.

d. Class D predicates (e.g. "resent", "regret", "be odd", "be strange"); these factive predicates have complements which are argued to be presupposed, and hence not asserted.

---

[2]In the previous studies, topic-marking (a.k.a. topicalization) and meaning of topic seem to be used without distinction. Nonetheless, we can say that topic can be marked even in embedded clauses and the topic-marked constituents can be potentially interpreted as conveying topic meaning.

e. Class E predicates (e.g. “realize”, “know”); these semifactives (factives that lose their factivity in questions and conditionals) have a reading on which the subordinate clause is asserted. (Heycock, 2007, p. 189)

Based on the division presented in (4), complement clauses may or may not contain a topicalized phrase, depending upon whether the predicate of the matrix clause belongs to Class A, B, or E.

Moreover, contrastive topic can relatively freely appear in embedded clauses. Bianchi and Frascarelli (2010) examine embedded topicalization in English. In short, their conclusion is that contrastive topics (C-Topics in their terminology) can be interpreted within complement clauses. In other words, although the complement clauses are not endowed with assertive force, an interpretation of contrastive topic is acceptable by native speakers. This claim shows a similarity to the analysis of (*n*)*un*-marked phrases in Korean relative clauses (§3.3.3). If (*n*)*un* appears in relative clauses as presented below, the (*n*)*un*-marked constituent is evaluated as containing a contrastive meaning (Lim, 2012).

(5) hyangki-nun coh-un kkoch-i phi-n-ta.
scent-NUN good-REL flower-NOM bloom-PRES-DECL
‘A flower with a good scent blooms.’ [kor] Lim (2012, p. 229)

11.1.2 Analysis

Two restrictions are factored into constraining information structure in complement clauses.

First, topic fronting can happen even in embedded clauses. Some languages, such as Italian, do not impose any restriction on topic fronting in embedded clauses as exemplified in (1b). Even in languages which have such a restriction (e.g. English, Japanese, and Korean), constituents can be topicalized in embedded clauses if the constituents carry a contrastive meaning as shown in (5). The topicalized constituents in complement clauses would to be more precisely represented as *contrast-topic*. At least in English, Japanese, and Korean, such a difference does not matter in generating sentences, because the meaning difference between them is not marked in surface forms.[3] One potential problem can be found in languages which employ different marking systems for contrastive topics and non-contrastive topics. Recall that Vietnamese uses *thì* for expressing

[3]It also needs to be studied how to specialize the focus/topic fronting in these embedded contexts in future work.

contrastive topics, which does not go for marking non-contrastive topic (Nguyen, 2006). Topic-marking systems in embedded clauses in Vietnamese-like languages need to be further examined in future work.

Second, if the main verb is one of the members of verbs of saying (*say*), semi-factive verbs (*realize*), and quasi-evidential verbs (*it appears*), the complement clause can be asserted, and thereby the structural relation between main and complement clauses is normally (but not always) specified as *focus*. Otherwise, the complement clause has an underspecified relation (i.e. *info-str*) to its matrix clause(s).

For instance, since the main verb *appears* in (6) is quasi-evidential, it has an arrow to *read* in complement clauses, whose value is specified as *focus*. Note that the syntactic subject *it* is an expletive (i.e. semantically and informatively vacuous), and does not have any information structure relation to the clause.[4]

(6) ~~It~~ appears ~~that~~ Kim read the books.
focus

Appears in (6), accordingly, has the following structure.

(7)
$$\begin{bmatrix} \text{STEM} & \langle \textit{appears} \rangle \\ \text{CLAUSE-KEY} & \boxed{1} \\ \text{COMPS} & \left\langle \left[\text{INDEX} \quad \boxed{2} \right] \right\rangle \\ \text{ICONS} & \left\langle \begin{bmatrix} \textit{focus} \\ \text{TARGET} \quad \boxed{2} \\ \text{CLAUSE} \quad \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

11.2 Relative Clauses

Which information structure value is assigned to the head noun modified by relative clauses? The behaviours of information structure shown by relative clauses have been analyzed from three points of view in previous literature: (i) Relative clauses assign topic to their modificands or the relative pronouns (Kuno, 1976; Bresnan and Mchombo, 1987; Jiang, 1991; Bjerre, 2011). (ii) Relative clauses do not always give a topic meaning to the head NPs (Ning, 1993; Huang et al., 2009).

[4]There could be some counterexamples to this analysis. Further work will examine the full range of information structure relations that complement clauses have to their main clauses.

(iii) Relative clauses signal focus on the head nouns (Schachter, 1973; Schafer et al., 1996). This subsection examines the three standpoints one by one, and offers a new approach to properties of information structure that relative clauses have in relation to their relativized constituents.

11.2.1 Background

First, Bresnan and Mchombo (1987) and Bjerre (2011) claim that relative pronouns indicate the topic function, as stated before. For convenience sake, the analysis that Bresnan and Mchombo provide is shown again below.

(8) The car [which you don't want ___] is a Renault.
TOPIC OBJ
(Bresnan and Mchombo, 1987, p. 757)

Bjerre (2011) provides a similar analysis to (8). This analysis is presented in an example (9) in Danish. Bjerre, making use of clefting as a tool to diagnose focus, claims that (9b), which has a clefted relative pronoun *den*, is of dubious acceptability. Whereas (9a), in which an interrogative pronoun *hvem* is clefted, sounds normal. Note that focus and topic are mutually exclusive.

(9) a. Som komponist er det naturligvis vigtigt, at lytterne ved,
as composer is it of course important that listeners.DEF know
hvem det er der har skrevet den musik, de lytter til.
who it is there has written that music they listen to
'As a composer it is of course important that the listeners know
who it is that has written the music they are listening to.'

b. ???Som komponist er det naturligvis vigtigt, at lytterne kender
as composer is it of course important that listeners.DEF know
den musik hvilken det er der lyttes til.
that music which it is there listen.PRS.PAS to
'As a composer it is of course important that the listeners know
that music which it is that is listened to.' [dan] (Bjerre, 2011, p. 279)

However, it is almost impossible for this dissertation to apply their approach to the current work for three reasons. The first two reasons are based on the distributional properties of relative pronouns, and the other one is due to system-internal factors. The first and the most important reason is that relative pronouns do not necessarily exist in all human languages. Japanese and Korean, for

example, do not have relative pronouns, and relative clauses in these languages are constructed in a different way (Baldwin, 1998; Kim and Park, 2000). If relative pronouns were universally to be evaluated as bearing the topic function, all relative clauses in Korean and Japanese would be topicless constructions. Second, relative pronouns can be missing in some circumstances even in English (e.g. those corresponding to object nouns in restrictive readings). Since English is not a topic dropping language (Chinese, Japanese, Korean, etc.), the dropped relative pronouns are rather difficult to explain with respect to information structure. Finally, as hypothesized in §9.1.4, because relative pronouns are syncategorematic, their lexical type inherits from *no-icons-lex-items* which has an empty list of ICONS. Hence, relative pronouns within the current work cannot participate in building up the list of ICONS, though they can perform a role to signal information structure values to their heads and/or dependents.

The first argument above is partially grounded upon the following constraint (Bresnan and Mchombo, 1987, 19f). That is to say, when there is no relative pronoun, relativized constituents would play the same role, which may give an answer to the first two problems posed above.

(10) The thematic constraint on relative clauses: A relative clause must be a statement about its head noun. (Kuno, 1976, p. 420)

Kuno provides several examples in Japanese and English to verify (10). First of all, Kuno argues (11a) is derived from not (11b) but (11c), in which *sono hon* 'the book' occurs sentence-initially with the topic marker *wa* to signal the theme (i.e. topic in this dissertation). Recall that the constituent associated with aboutness topic usually shows up in the initial position in Japanese (Maki et al., 1999; Vermeulen, 2009).

(11) a. [Hanko-ga yonda] hon
Hanako-NOM read book
'The book that Hanako read'

b. [Hanko-ga sono hon-o yonda] hon
Hanako-NOM the book-ACC read book

c. [[sono hon-wa]$_{\text{theme}}$ Hanko-ga yonda] hon
the book-WA Hanako-NOM read book [jpn] (Kuno, 1976, p. 419)

At first glance, the explanation about the linguistic phenomena presented above sounds reasonable. It seems clear that relative clauses present certain constraints on information structure. Yet, it is still necessary to verify whether the head nouns modified by relative clauses always and cross-linguistically carry the meaning of topic. There are several previous studies providing a counterargument to (10).

Huang et al. (2009), from a movement-based standpoint, basically accepts that topics and relative clauses share some characteristics with *wh*-constructions as A′-movement structures. The common properties notwithstanding, they argue that relative clause structures are not derived from topic structures for two reasons, contra Kuno (1976). First, a topic relation does not license a relative construction in Chinese. For instance, if a topic structure were sufficient for relativization in Chinese, (12b) and its relativized counterpart (12c) would be equally acceptable.

(12) a. yiwai fasheng-le
accident happen-LE
'An accident happened.'

b. tamen, yiwai fasheng-le
they accident happen-LE
'(As for) them, an accident happened.'

c. *[[yiwai fasheng-le de] neixie ren]
accident happen-LE
textscde those person
'the people such that an accident happened' [cmn] (Huang et al., 2009, p. 212–213)

In other words, Kuno's claim (10) is not cross-linguistically true. Second, Ning (1993) reveals that a relativized construction may be well-formed even though its corresponding topic structure is ill-formed. Thus, the well-formedness of a topic structure is neither necessary nor sufficient for the acceptability of a corresponding relative structure at least in Mandarin Chinese.

Schachter (1973) probes into the relationship between focus constructions (e.g. clefts) and restrictive relative constructions, and concludes that they bear a striking likeness to each other. On the basis of the findings from four languages including English, Akan, Hausa, and Ilonggo,[5] Schachter

[5]an Austronesian language spoken in the Philippines

sets up a hypothesis: both constructions syntactically necessitate the promotion of a linguistic item from an embedded clause into its main clause, and semantically have to do with foregrounding (i.e. making a specific part of a sentence conspicuous at the expense of the rest). This argument goes sharply against (10). The following examples in Akan [aka] and Ilonggo [hil] show that constructions involving relative clauses and focus constructions are structurally quite similar to each other.

(13) a. àbòfr̀á áà míhúù nó
child that I.saw him
'a child that I saw'

b. àbòfr̀á nà míhúù nó
child that I.saw him
'It's a child that I saw.' [aka] (Croft, 2002, p. 108)

c. babayi nga nag- dala sang bata
woman that AG.TOP- bring NONTOP child
'the woman that brought a child'

d. ang babayi and nag- dala sang bata
TOP woman TOP AG.TOP- bring NONTOP child
'It was the woman who brought a child' [hil] (Ibid. p. 108)

One difference between (13a) and (13b) is which marker (i.e. a relative marker *vs*. a focus marker) is used. As exemplified earlier in §4.3, *nà* in (13b) behaves as a focus marker in Akan, and is in complementary distribution with a relative marker *áà* in (13a). The same goes for (13c) and (13d) in Ilonggo. The relative marker *nga* and the second topic marker *ang* share the same position to draw a boundary between the promoted NP and the relative clause or the cleft clause.

The structural similarity notwithstanding, we cannot jump to a conclusion that the head nouns of relative clauses always bear the focus function. First of all, we cannot say that a structural likeness equals likeness of information structure meaning. Although cleft constructions involve something formally similar to relative clauses, that does not mean that they are information structurally similar.

11.2.2 Analysis

In sum, there are opposing arguments about the information structure properties that the head nouns of relative clauses have. Thus, it is my understanding that it is still an open question whether relative

clauses assign their head nouns a focus meaning or a topic meaning. Moreover, previous studies show that the relation of information structure which the head nouns have to their relative clauses could be language-specific. This dissertation does not rush to create a generalization, instead allows a flexible representation: The information structure values of the constituents modified by relative clauses should be *focus-or-topic*, which is the supertype of both *focus* and *topic* within the hierarchy of *info-str* (Figure 9.1). That implies the relativized constituents can be evaluated as delivering either focus in some cases or topic in other cases. In other words, information structure in constructions involving relative clauses is analyzed analogously to focus/topic-fronting constructions (§12.6). The preposed constituents in focus/topic fronting constructions carry ambiguous meanings but for the help of contextual information, and because of this they have to be flexibly specified as *focus-or-topic*. The same motivation goes for relativized constituents.

Building upon the argument above, this dissertation also marks the difference between restrictive reading and non-restrictive readings in information structure values. Sometimes it is necessary to distinguish restrictive relatives and non-restrictive relatives in assigning the value of *info-str*. The distinction between restrictive and non-restrictive readings is also revealed in the corpus study of Part III in the same way: All non-restrictively relativized NPs are analyzed as aboutness topic (tagged with 'ab') with respect to the relative clause in English, Spanish, and Russian. What follows provides two more reasons.

First, restrictive relative clauses and non-restrictive relative clauses have been regarded as having different linguistic behaviours in most previous work. To begin with, there is an orthographic convention in English of setting off non-restrictive relatives with commas, and not using commas for ordinary restrictive relatives.[6] Syntactically, it has been stated that the distinction between restrictive readings *vs*. non-restrictive ones yields different bracketing as presented in (14). The restrictive relative clause in (14a) modifies the head noun *dog* itself, and then the entire NP *dog which Kim chases* is combined with the determiner as *head-spec-phrase*. In contrast, the non-restrictive relative clause in (14b) modifies the NP in which the noun *dog* takes the determiner beforehand. They also show contrastive syntactic behaviour in binding of anaphora (Emonds, 1979), co-occurrence with NPIs (e.g. *any*), and focus sensitive items (e.g. *only*) (Fabb, 1990).

[6]However, this convention is not necessarily followed. In this dissertation, commas are inserted just for ease of comparison.

(14) a. [[The [dog that Kim chases]] barks.]

b. [[[The dog,] which Kim chases,] barks.]

Semantically, they may not share the same truth-conditions.

(15) a. Kim has two children that study linguistics.

b. Kim has two children, who study linguistics.

(15b) implies that Kim has two and only two children, while (15a) does not. For example, if Kim has three children, the proposition of (15b) would not be felicitously used, whereas that of (15a) may or may not be true depending on how many children among them study linguistics. Given that they have different properties in semantics as well as syntax, it is a natural assumption that they behave differently in information structure as well.

Second, beyond the general properties that restrictive relative clauses and non-restrictive relative clauses have, there is a distributional reason for viewing them differently with regard to their information structure.

(16) a. Kim chases the dog that likes Lee.

b. Kim chases the dog, which likes Lee.

c. Kim chases the dog, and it likes Lee.

d. Kim chases the dog, and as for the dog, it likes Lee.

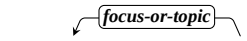
e. Kim chases the dog, and speaking of the dog, it likes Lee.

Unlike restrictive relative constructions such as (16a), non-restrictive constructions such as (16b) can be paraphrased into (16c-e). (16c) reveals that non-restrictive relatives are almost equivalent to coordinated clauses which clearly involve root phenomena (Heycock, 2007, p. 177). In (16c), a pronoun *it* is used as referring to *the dog* in the previous clause, which means *the dog* cannot receive *focus* from the non-restrictive clause in (16b). The focused constituents in the non-restrictive clause should be either the object *Lee* or the VP *likes Lee*. Finally, (16d-e) conclusively pass the test for aboutness topic.

In sum, the semantic head of relative clauses (i.e. the verb in relative clauses) basically has a *focus-or-topic* relation with relativized dependents. Non-restrictive relatives additionally have a

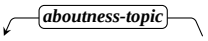
more specific constraint; *aboutness-topic*. The schema of those constraints is exemplified in the following dependency diagrams.

(17) a.



The dog that Kim chases barks.

b.



The dog, which Kim chases, barks.

Consequently, all relative clauses inherit from the following type in that *aboutness-topic* is a subtype of *focus-or-topic*. Note that the information structure relation that the relativized NPs have to the relative clauses should be constructionally added using C-CONT, because the meaning is specified at the phrasal level.

(18)
$$\begin{bmatrix} \textit{rel-clause} \\ \text{HD} \mid \text{INDEX } \boxed{1} \\ \text{NHD} \mid \text{CLAUSE-KEY } \boxed{2} \\ \text{C-CONT} \mid \text{ICONS} \begin{bmatrix} \textit{focus-or-topic} & \\ \text{TARGET} & \boxed{1} \\ \text{CLAUSE} & \boxed{2} \end{bmatrix} \end{bmatrix}$$

The phrase structure type responsible for non-restrictive relative clauses requires us to impose a more specific value (i.e. *aboutness-topic*). This is left to the future work.

11.3 Adverbial Clauses

Adverbial clauses in the analysis of this dissertation may be evaluated as having a relation of either *topic* or just the underspecified value *info-str*, with respect to the main clauses. This depends on the type of subordinating conjunction.[7]

11.3.1 Background

Several previous studies investigate conditional *if*-clauses and temporal *when*-clauses with respect to topichood. Haiman (1978) argues that conditionals are topics, and Ramsay (1987) also argues that *if*/*when* clauses are endowed topichood when they precede the main clauses.[8] That is to say,

[7]Using this strategy, subordinating conjunctions sometimes introduce an underspecified *info-str* element into ICONS like verbal items that take clausal complements (§11.1). These underspecified elements are disadvantageous as mentioned in the first footnote of the current chapter, and a revised analysis in future work will suppress this problem.

[8]Ramsay (1987)'s claim implies that *if*/*when* clauses at the beginning, at the end, and in the middle of an utterance differ in their information structure. The traditional movement-based studies account for the type of variation in conditional and temporal clauses in terms of the so-called the Adjunct Island Constraint (Huang, 1982): Postposed

conditional and temporal clauses have a topic feature at least if they are sentence-initial. Following their argument, the present work assumes that topic is associated with preposed conditional and temporal clauses with respect to the main clauses. Syntactically, because they appear in the sentence-initial position and their function is to restrict the domain of what the speaker is talking of, they serve as *frame-setting* presented in Figure 9.3. As for sentence-final/internal conditional and temporal clauses, their information structure relation to the main clause parsimoniously remains underspecified.

11.3.2 Analysis

Before moving to adverbial clauses, it is necessary to look at the information structure relationship between adverbs and their clauses. Frame-setters, as discussed previously, have several restrictions: (i) they normally appear initially, (ii) they can multiply occur in a single clause, and (iii) they should play a role in restricting the domain of what the speaker is talking about (e.g. spatial, temporal, manner, or conditional). First, the clause-initial constraint can be conditioned by [L-PERIPH +], which renders the constituents left-peripheral. The second constraint can be enforced by *sform*, as presented in Chapter 9: namely *frame-setting vs*. *non-frame-setting*. The third one might be controversial, because information about lexical semantics has not yet been included in the DELPH-IN reference formalism.[9] Future work should refer to lexical semantic information to tell whether an adverb conveys a spatial, temporal, or manner meaning.

The combination between a frame-setting adverb and the rest of sentence should be carried out using a specific subtype of *head-mod-phrase*, which implies *head-mod-phrase* needs to be divided into at least two subtypes; one requires [L-PERIPH +] of NON-HEAD-DTR, and the other requires [L-PERIPH –] of both daughters. Thus, a temporal expression such as *today* in (19) has *topic* relations to the main verb *barks*.

(19) Today the dog barks.
topic

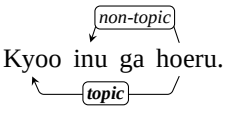
conditional and temporal clauses are adjoined to VPs forming an adjunct island, while preposed ones are moved into IP's specifier position (Iatridou, 1991) or generated *in situ* (Taylor, 2007). In other words, preposed adverbial clauses modify the main sentence, while postposed ones modify the VP.

[9]In principle, such an information should be specified in tandem with WordNet. There is some on-going research which tries to incorporate lexical semantic information within DELPH-IN grammars using WordNets (Bond et al., 2009; Pozen, 2013).

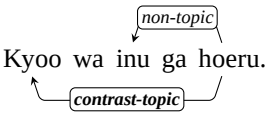
In Japanese and Korean, *wa* and (*n*)*un* can be attached to adjuncts. If they are adjacent to adjuncts, the constituents are normally evaluated as bearing contrastiveness. Following the annotation schema in Chapter 6, if an adjunct is combined with *wa* or (*n*)*un*, the adjunct should be associated with *contrast*, even when it appears in the leftmost position. Consequently, *kyoo* 'today' in the left-peripheral position has a plain *topic* relation, while *kyoo-wa* 'today-wa' has a *contrast-topic* relation to the verb *hoeru* 'bark'.

(20) kyoo (wa) inu ga hoeru.
today (WA) dog NOM bark
'Today, the dog barks.' [jpn]

(21) a.



b.



Turning back to adverbial clauses, my argument is that subordinating conjunctions are responsible for the information structure relation between adverbial clauses and main clauses. First of all, subordinating conjunctions that entail temporal and conditional clauses signal *topic* (Haiman, 1978; Ramsay, 1987), as discussed above. Other subordinating conjunctions assign an underspecified *info-str* value, because there seems to be no clear distinction on the status of information structure. Causal conjunctions, such as *because* in English and *weil* in German, do not show consistency in information structure (Heycock, 2007), which means there is no lexical and phrasal clue to identify the information structure relations.[10] It is less clear how concessive conjunctions, such as (*al*)*though*, configure information structure, though they are known to be partially related to information structure (Chung and Kim, 2009). They are also provisionally treated as underspecified in this analysis. Some conjunctions with multiple meanings, such as *as*, are also assumed to assign an underspecified value, because we do not clearly see the information structure meanings if it were not for contextual information. For example, the lexical entries of *when* and *if* have [ICONS-KEY *topic*], while those of *because*, *though*, *as*, etc. have no such constraint.

(22) *when-subord* →
$$\begin{bmatrix} \text{STEM} & \langle \textit{when} \rangle \\ \text{ICONS-KEY} & \textit{topic} \end{bmatrix}$$

[10]The meaning could be clear by a specific prosodic pattern, like intonation.

All subordinate conjunctions have one *info-str* value in the list of ICONS. However, they do not inherit from any *icons-lex-item* presented in Chapter 9. This is because the TARGET should point to the semantic head (mostly a verb) of the adverbial clause, not the conjunctions themselves, and the CLAUSE is readily and lexically identified to be the INDEX of the main clause. The following AVM presents this co-indexation.

(23)
$$\begin{bmatrix} \textit{subconj-word} & \\ \text{HEAD} \mid \text{MOD} & \left\langle \left[\text{INDEX} \quad \boxed{1} \right] \right\rangle \\ \text{VAL} \mid \text{COMPS} & \left\langle \left[\text{INDEX} \quad \boxed{2} \right] \right\rangle \\ \text{ICONS} & \left\langle ! \begin{bmatrix} \text{TARGET} & \boxed{2} \\ \text{CLAUSE} & \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

As a consequence, adverbial clauses have an information structure relation to the main clauses, as exemplified in (24). The arrows from *reads* to *barks* at the bottom are created by (23). The value *topic* on the arrow of (24a) is specified in (22). In addition, the arrow of (24b) is specified as merely *info-str*, since the subordinate conjunction is devoid of such a constraint.

(24) a. When the dog barks, Kim reads the book.
topic

b. Because the dog barks, Kim reads the book.
info-str

11.4 Summary

This chapter addressed how information structure in multiclausal utterances are represented via ICONS and what kinds of constraints are imposed on the non-matrix clauses. There are three types of non-matrix clauses that this chapter looked into: complement clauses, relative clauses, and adverbial clauses. First, the information structure relation between matrix clauses and their complement clauses largely depend on the verbal type of the main predicate. If the predicate serves to invoke an assertion to the complement clause, the complement clause has the *focus* relation to the main clause. Second, information structure that the head nouns have to their relative clauses depends on the reading of relative clauses. If the relative is restrictive, the head nouns are assigned *focus-or-topic* by the relative clauses. Otherwise, they are assigning a more specific type *aboutness-topic*. Third, information structure in adverbial clauses are impacted by the type of conjunctions and the position of

adverbial clauses. If the adverbial clauses are temporal or conditional and appear sentence-initially, they are assigned *topic*. Other adverbial clauses are preferentially underspecified.

Chapter 12

FORMS OF EXPRESSING INFORMATION STRUCTURE

This chapter looks into specific forms of expressing information structure in human language. Every language presumably has one or more operations to articulate information structure. The operations are strategized sometimes at the lexical level by employing specific lexical items or rules, and sometimes at the phrasal level by using special constructions. §12.1 goes over focus sensitive items, and presents how they are represented in articulation of information structure via ICONS (Individual CONStraints). §12.2 deals with argument optionality from the point that focus is defined in terms of whether or not a constituent is omissible (i.e. optionality). The remaining portion of this chapter addresses specific constructions related to forming information structure. §12.3 probes into scrambling in Japanese and Korean, which are deeply related to arrangement of information structure components. §12.4 delves into cleft constructions, which are the most well-known operation for expressing focus in an overt way. §12.5 refers to passive constructions which have something to do with structuring information in some languages. Lastly, §12.6 and §12.7 investigate two types of syntactic operations which are seemingly similar to each other, but are constructed differently: focus/topic-fronting and dislocation. Although they are not implemented in the present work,[1] I would like to present which information structure meaning the fronted and dislocated constituents conveys for future work.

12.1 Focus Sensitive Items

Lambrecht (1996) provides several intriguing explanations about the lexical properties of some focus sensitive items. First, emphatic reflexives cannot involve a topic interpretation, because they are usually focused in the sentence. Second, for the same reason, NPIs (e.g. *any* in English) and negative words (e.g. *not*, *never*, *no*, *nobody*, *nothing*, and so forth in English) cannot play a topic

[1] In the current version of ERG (1212), there is no distinction between focus/topic fronting and dislocation. Thus, these two constructions are not implemented in the present work, and how to build this representation compositionally is left to future work.

role, either. This means that some lexical categories, such as reflexives, NPIs, and negative words, are inherently incompatible with the topic role. Nonetheless, focus sensitive items do not have the same properties altogether. Rather, there are two subtypes of words with an inherent focus meaning. The nominal categories among them, such as *anybody*, *nobody*, and *nothing*, are focus-sensitive by themselves. In contrast, negative modifiers such as *any*, *not*, *never*, and *no* assign a focus relation not to themselves, but to the constituent they modify. Henceforth, I call the former Type I, and the latter Type II.

(1) a. Focus Sensitive Type I assigns an information structure role (either *non-topic* or *focus*) to itself.

b. Focus Sensitive Type II assigns such a role to its adjacent constituent.

Type I includes *nothing*, *nobody*, etc. These lexical items are contentful, introducing an EP into the list of RELS. Their lexical constraint inherits from *one-icons-lex-item*, and additionally the TARGET of the element in the ICONS list is co-indexed with their INDEX. Lexical items under Type II also inherit from *one-icons-lex-item*, but their TARGET is co-indexed with the INDEX of their modificands. For instance, a lexical entry for *only* (Type II) can be described as (2).

(2) *only* →

$$\begin{bmatrix} \text{STEM} \left\langle \textit{only} \right\rangle \\ \text{HEAD} \mid \text{MOD} \left\langle \begin{bmatrix} \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \end{bmatrix} \right\rangle \\ \text{CONT} \mid \text{ICONS} \left\langle \, ! \, \boxed{2} \begin{bmatrix} \textit{focus} \\ \text{TARGET} \quad \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

Regarding the choice of *info-str* value for this type of lexical item, it could be marked as *contrast-focus* in that *only* has an exhaustive effect (Velleman et al., 2012). However, the current work leaves the value less specified because the corpus study in Part III reveals that there are some sentences in which there is no overtly realized alternative of the constituent modified by *only*.

12.1.1 Quantifiers

Quantifiers also exhibit focus-sensitivity. In particular, Lambrecht (1996) argues that universally quantified NPs can be used as topics, whereas other quantified NPs cannot, as exemplified in (3).

(3) a. As for all his friends, they ...

b. *As for some people, they ... (Lambrecht, 1996, p. 156)

That implies non-universally quantifying determiners, such as *some*, assign *non-topic* to the head as represented in (4).

(4) *some* →
$$\begin{bmatrix} \text{STEM} & \langle \textit{some} \rangle \\ \text{VAL} \mid \text{SPEC} & \left\langle \left[\text{ICONS-KEY} \quad \textit{non-topic} \right] \right\rangle \end{bmatrix}$$

In (3b), what is responsible for putting an *info-str* element into the ICONS list is *as for* when we are not using the hypothetical suffixes '-a' and '-b'. In this case, the *info-str* value of the element is *topic*, the TARGET of the element is co-indexed with the INDEX of *people*, and the element itself is co-indexed with the ICONS-KEY of *people*. However, the ICONS-KEY of *people* is already constrained as *non-topic* by (4). Because this value is inconsistent with the *topic* value introduced by *as for*, *as for some people* is ruled out.

12.1.2 Wh-words

Wh-questions, as stated many times so far, have been employed as a tool to probe the meaning and markings of focus: a technique which looks quite reliable from a cross-linguistic stance. *Wh*-words have often been regarded as inherently containing a focus meaning. That is to say, in almost all human languages, *wh*-words share nearly the same distributional characteristics with focused words or phrases in non-interrogative sentences.

A typological implication is provided in Drubig (2003, p. 5): In a language with *wh*-phrases *ex situ*, the *wh*-phrase usually appears in focus position. This typological argument is convincingly supported by several previous studies in which the linguistic similarity that *wh*-words have to meaning and marking of focus is addressed. According to Comrie (1984) and Büring (2010), Armenian is a language with strict focus position: Focused constituents should appear in the immediately preverbal position (as exemplified earlier (p. 80)). Tamrazian (1991) and Megerdoomian (2011) argue that focused elements and *wh*-words in Armenian show a striking similarity to each other from various points of view. Especially, *wh*-words and focused constituents cannot co-occur, because they occupy the same syntactic position. In other words, *wh*-words should occur in the focus position in Armenian.

(5) a. ov a Ara-in həravir-el?
who AUX/3SG.PR Ara-DAT invite-PERF
'Who has invited Ara' [hye]

b. *ov Ara-in a həravir-el?
who Ara-DAT AUX/3SG.PR invite-PERF [hye]

c. *Ara-in a ov həravir-el?
Ara-DAT AUX/3SG.PR who invite-PERF [hye] (Megerdoomian, 2011)

According to Ortiz de Urbina (1999)'s analysis in Basque, *wh*-words are also in complementary distribution with focused constituents in that both of them (i) occupy the immediately preverbal position,[2] optionally preceded by a constituent with topic meaning, and (ii) seldom occur in embedded clauses. From a transformational perspective, Ortiz de Urbina (1999) argues that *wh*-words and focused items are able to undergo cyclic movement with bridge verbs.[3]

From these linguistic facts and analyses, this dissertation assumes that *wh*-words are inherently focused items which always have the *focus* relation with the clause that they belong to. The linguistic constraint on *wh*-words is represented as the following AVM. Note that *wh*-words are focus sensitive items under Type I (*one-icons-lex-item*). The TARGET is co-indexed with its INDEX.

(6)
$$\begin{bmatrix} \textit{wh-words} & \\ \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{ICONS} & \left\langle ! \boxed{2} \begin{bmatrix} \textit{semantic-focus} \\ \text{TARGET} \ \boxed{1} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

(6) involves two more things. First, as investigated in Gryllia (2009) *wh*-questions are incompatible with contrastive focus. The value of ICONS should be specified as *semantic-focus*. However, *semantic-focus* in (6) does not necessarily provide an answer to *wh*-questions. As discussed in Chapter 3 (§3.4.4), because contrastiveness is heavily speaker-oriented (Chang, 2002), the answerers may alter information structure in a solicited question as they want. The (*n*)*un*-marked *Kim-un* in (40) delivers a contrastive meaning, but the answer does not directly correspond to the question's

[2]Note that the canonical position of focused items in Basque is preverbal as exemplified in Chapter 4.

[3]This dissertation does not present a transformational analysis, but it is clear that both types show a significant similarity to each other.

information structure. Instead, the replier manipulates information structure in order to attract a special attention to *Kim*. In other words, *wh*-words themselves are still assigned *semantic-focus*.

(7) Q: nwuka o-ass-ni?
who come-PST-INT
'Who came?' [kor]

A: Kim-un o-ass-e.
Kim-NUN come-PST-DECL
'Kim came.'
(conveying "I know that at least Kim came, but I'm not sure whether or not others came.") [kor]

On the other hand, functionally speaking, there are two types of interrogatives; informational questions and rhetorical questions. The former explicitly solicits the hearer's reply, while the latter does not. Since rhetorical questions perform a function to express an assertion in a strong and paradoxical manner, their interpretation naturally hinges on the context. For example, (8a) can be ambiguously read as either an informational question or a rhetorical question, and each reading can be paraphrased as (8b-c), respectively. That means the *wh*-elements in rhetorical questions function like a trigger to derive the form of interrogative sentences, but they can also convey quantificational readings as implied by *nobody* in (8c). In other words, it is true that *wh*-words convey focus meaning in *wh*-questions, but not all the sentential forms (i.e., *sform* are necessarily *focus-bg*).

(8) a. Who comes?

b. I'm wondering which person comes.

c. Nobody comes.

Do all *wh*-questions sound ambiguous (at least in English)?[4] We know that in actual speech they do not. This is because the meaning becomes unambiguous depending on where the accent is assigned as illustrated in (9), in which the A-accent falls on different words.

(9) a. WHO comes?
≈ I'm wondering which person comes.

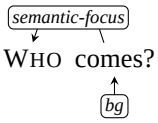
[4]At the level of compositional semantics, rhetorical questions may not have to be considered in linguistic modelling, because they are basically a pragmatic phenomenon. What I want to say here is that other sentential constituents in *wh*-questions can be focused, and this needs to be taken into account in modelling *wh*-questions in a flexible way.

b. Who COMES?
≈ Nobody comes.

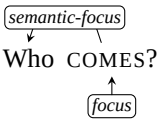
Informational questions, in order to clarify the meaning of an assertion, employ an ordinary intonation pattern (i.e. rise-fall), whereas rhetorical questions involve pitch accent within the intonation contour (Gunlogson, 2001). In other words, the prosodic marking for information structure (i.e. intonation contour driven by pitch accent) has an influence on the interpretation of *wh*-questions. Note that *comes* in (9a) can be eliminated, while both *who* and COMES cannot be missing in (9b). The focus in (9b) comes from the accented verb COMES and spreads to the whole sentence (i.e. all-focus), whereby it forms a different information structure from (9a).

If a *wh*-question is rhetorically used, the entire sentence is in the focus domain. For example, if *Who comes?* is not asked rhetorically, its information structure can be represented as (10a). The verb *comes* in (10a) has *bg*, which implies that *Who* exclusively bears a focus relation in the sentence. In contrast, if the question is rhetorically used, the sentence should be informatively structured as (10b), in which the verb *comes* also has the *focus* relation within the clause (i.e. *all-focus*). Since the choice between them is only contextually conditioned, in an approach to grammar engineering that represents ambiguity via underspecification wherever possible the MRS (Copestake et al. 2005) representing *Who comes?* has to be able to subsume (10a-b). Given that the lowest supertype of *bg* and *focus* is *non-topic*, *wh*-questions should be analyzed as (10c).[5]

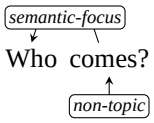
(10) a.



b.



c.



12.1.3 Negative Expressions

Negation is also sensitive to focus (Partee, 1991; Krifka, 2008). For example, negative quantifiers (e.g. *no*), replacing sentential negation (e.g. *not ..., but ...*), and some other constructions including negation such as *neither ...* are associated with focus almost invariably. However, we cannot say negative verbs are assigned *focus* all the time. For example, in the following Q/A pair, the focused element should be the subject KIM. The rest of the reply can be elided in the context.

[5] *Non-topic* on *comes* in (10c) should be introduced by a specific phrase structure rule to constrain *wh*-questions with respect to information structure. A creation of phrase structure rules for interrogative sentences is left to future work.

(11) Q: Who didn't read the book?

A: KIM (didn't read the book).

For the reason, this dissertation argues that the value that negative operators assign to operands is *non-topic*, which can be either *focus* or *bg*.

12.2 Argument Optionality

Argument-optionality (a.k.a. *pro*-drop, including subject-drop and topic-drop) has been assumed to have to do with information structure. The basic explanation about the relationship between dropped elements and articulation of information structure is provided in Alonso-Ovalle et al. (2002), with special reference to subject-dropping in Spanish. Additionally, the distinction between subject-drop and topic-drop has also been studied in Li and Thompson (1976), Huang (1984) and Yang (2002) (as discussed in Chapter 3 (§3.2.3)). Argument-optionality is also crucial in computational linguistics; in multilingual processing, such as (multilingual) anaphora resolution and machine translation (Mitkov et al., 1995; Mitkov, 1999), as well as monolingual processing, such as syntactic parsing and semantic interpretation. Just as with other subfields of language processing, there are two approaches to resolving dropped elements within language applications: First, several rule-based algorithms have been designed to resolve zero anaphora in *pro*-drop languages.[6] Second, there are several (semi-)machine-learning methods to compute zero anaphora in topic-drop languages for the purpose of machine translation.[7]

In this dissertation, I use optionality and omissibility as synonyms. Whether an argument can be elided or not needs to be augmented into an analysis of argument optionality with respect to focality. As discussed thus far, the most noteworthy factor for focused constituents is their inomissibility. If a constituent is omitted then the constituent is not focused. This restriction can be defined as (12). Note that (12a) entails (12b).

(12) a. C is inomissible iff C is focused.

---

[6]These are provided in Han (2006), Byron et al. (2006), and so on.

[7]These can be found in Zhao and Ng (2007), Yeh and Chen (2004), Kong and Ng (2013), and Chen and Ng (2013) for Chinese, Nakaiwa and Shirai (1996) and Matsui (1999) and Hangyo et al. (2013) for Japanese, and Roh and Lee (2003) for Korean.

b. If C is omitted then C is not focused.

For example, Spanish is a subject-drop language. The pronouns are often missing as shown in (13). However, if they have a meaning of *focus*, they have to appear with an accent (Cinque, 1977; Lambrecht, 1996). Therefore, the dropped subject in (13) should be regarded as *non-focus*.

(13) Ø Habla.
speaks
'(He/She/It) speaks.' [spa]

In the Argument Optionality library in the customization system (Saleem, 2010; Saleem and Bender, 2010), both subject dropping and object dropping are described and modelled. The questionnaire requires users to answer to several questions; (i) whether or not subjects/objects can be dropped in user's language, (ii) whether or not the verb needs to have a marker when the subjects/objects are dropped, (iii) whether or not subject-drop only happens in some particular contexts, and (iv) whether or not object-drop is lexically licensed. On top of them, I add one more constraint; The dropped elements are informatively constrained as *non-focus*. This constraint should be written into *basic-head-opt-subj-phrase* and *basic-head-opt-comp-phrase*. These two phrasal types now include some additional constraints on their subjects and complements as follows.

(14) a.

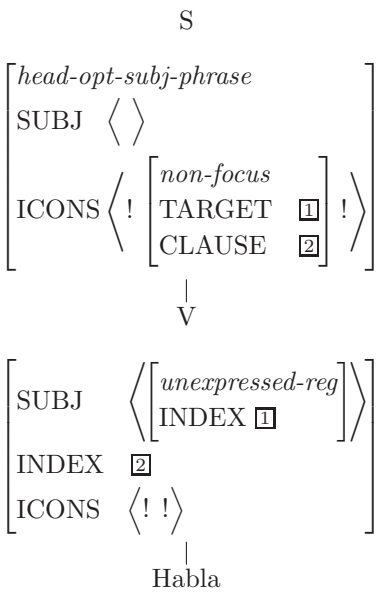
$$\begin{bmatrix} \textit{basic-head-opt-subj-phrase} & \\ \text{HD} \mid \text{VAL} \mid \text{SUBJ} & \left\langle \begin{bmatrix} \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{CLAUSE-KEY} & \boxed{3} \end{bmatrix} \right\rangle \\ \text{C-CONT} \mid \text{ICONS} & \left\langle ! \boxed{2} \begin{bmatrix} \textit{non-focus} & \\ \text{TARGET} & \boxed{1} \\ \text{CLAUSE} & \boxed{3} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

b.

$$\begin{bmatrix} \textit{basic-head-opt-comp-phrase} & \\ \text{HD} \mid \text{VAL} \mid \text{COMPS} & \left\langle ..., \begin{bmatrix} \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{2} \\ \text{CLAUSE-KEY} & \boxed{3} \end{bmatrix} \right\rangle \\ \text{C-CONT} \mid \text{ICONS} & \left\langle ! \boxed{2} \begin{bmatrix} \textit{non-focus} & \\ \text{TARGET} & \boxed{1} \\ \text{CLAUSE} & \boxed{3} \end{bmatrix} ! \right\rangle \end{bmatrix}$$

Building upon (14a), the derivation tree for (13) is sketched out in (15).

(15)



12.3 Scrambling

The typical case in which forms of expressing information structure do not coincide with information structure meanings can be found in the *wa* in Japanese and (*n*)*un* in Korean. NPs in Japanese and Korean, as presented several times, can have three types of marking; case-marking, *wa* or (*n*)*un*-marking, and null-marking (a.k.a. case ellipsis (Yatabe, 1999; Sato and Tam, 2012)). These are in complementary distribution with each other, and the choice among them is largely conditioned by information structure.

(16) Kim-ga/wa/Ø kita.
Kim-NOM/WA/NULL came
'Kim came. [jpn]'

As stated before, *wa* and (*n*)*un* can convey meaning of aboutness topic, contrastive topic, or even contrastive focus[8] (Choi, 1999; Song and Bender, 2011). Case markers are also ambiguously interpreted. They have sometimes been assumed to be associated with focus, but there are quite a

[8]In this vein, *wa* and (*n*)*un* perform the same role as the B-accent in English (Young Chul Jun, p.c.), which can also be used to express non-contrastive topic, contrastive topic, or sometimes contrastive focus (Hedberg, 2006).

few counterexamples which show that all case-marked NPs do not necessarily convey focus meaning in all languages (Heycock, 1994). Null-marking is also conditioned by information structure in some languages: The markers are not omissible if an NP is associated with *focus*. That means the null-marked NPs receive an interpretation of either topic or background (i.e. *non-focus*).

Nevertheless, this does not mean that NPs in Japanese and Korean deliver an informatively knotty meaning all the time. The meanings can be disentangled at the phrasal level, mainly depending on different word orders, such as basic *vs*. scrambling. Scrambling refers to constructions in which one or two objects are followed by the subject. This construction is productively used in Japanese and Korean (i.e. SOV in the basic order *vs*. OSV in the scrambled order). Scrambling has been rather discounted as a dummy operation in syntax and semantics, but Choi (1999) and Ishihara (2001) argue that scrambling has a strong effect on information structure. The contrast between orders with respect to *wa* is exhibited in the following examples.[9]

(17) a. John-wa kono hon-o yonda.
John-WA this book-ACC read
'As for John, he read this book.'

b. Kono hon-wa John-ga yonda.
this book-WA John-NOM read
'As for this book, John read it.'

c. John-ga kono hon-wa yonda.
John-NOM this book-WA read
'John read this book, as opposed to some other book.'
'*As for this book, he read this it.' [jpn] (Maki et al., 1999, p. 7–8)

The first sentence is in the basic word order, in which the subject is topicalized. The second sentence is scrambled, and the fronted object carries a topic meaning (i.e. *contrast-topic*). The third sentence is in the basic word order, but *wa* is attached to the object, not the subject. In that case, the topicalized object should be interpreted as containing contrastiveness (i.e. *contrast-focus*). Regarding the relationship between *wa* or (*n*)*un*-marking and word-order in Japanese and Korean, Song and Bender (2011) provides Table 12.1, adapted from Choi (1999).

[9]There can be one more sentence from this paradigm though Maki et al. (1999) do not include it in their source; *Kono hon-o John-wa yonda*, which is completely grammatical, but the *wa*-marked *John* is interpreted as containing contrastiveness. In order to show the authors' example as is, this sentence is not included in (17).

Table 12.1: Information structure of (*n*)*un*-marked NP

	in-situ	**scrambling**
subject	*topic*	*contrast-focus*
non-subject	*contrast-focus*	*contrast-topic*

According to Table 12.1, the set of allosentences given in (18) have different information structure. In other words, the default meaning of *wa* and (*n*)*un* (i.e. *contrast-or-topic*) can be narrowed down, interacting with word order (e.g. scrambling).

(18) a. Kim wa sono hon o yomu.
Kim WA DET book ACC read (*topic*)

b. sono hon o Kim wa yomu.
DET book ACC Kim WA read (*contrast-focus*)

c. Kim ga sono hon wa yomu.
Kim NOM DET book WA read (*contrast-focus*)

d. sono hon wa Kim ga yomu.
DET book WA Kim NOM read (*contrast-topic*) [jpn]

There is one additional property that *wa* and (*n*)*un* display: They cannot appear in an *all-focus* construction that allows only *semantic-focus* lacking contrastive meanings, as exemplified in (19).

(19) Q: doushita nano
what INT
'What happened?' [jpn]

A: Kim-ga/#wa sono hon-o/#wa yabut-ta.
Kim-NOM/WA DET book-ACC/WA tear-PST [jpn] (Song and Bender, 2011, p. 359)

In syntactic derivation, *topic-comment* presented below plays an important role in creating grammatical rules. The construction itself is [MKG *tp*] so that constituents which have picked up a topic cannot serve as the head daughter of another *topic-comment* phrase.

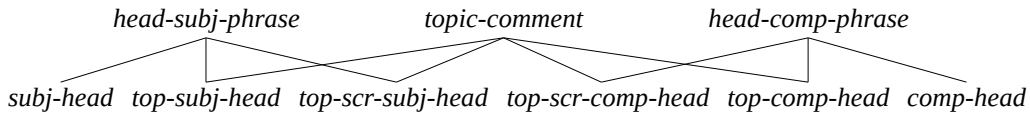


Figure 12.1: Strategy using three phrase structure rules

$$
(20)\ \begin{bmatrix}
\textit{topic-comment} & \\
\text{L-PEPIPH} & + \\
\text{MKG} & \textit{tp} \\
\text{HD}\,|\,\text{MKG}\,|\,\text{TP} & - \\
\text{NHD} & \begin{bmatrix} \text{MKG} & \textit{tp} \\ \text{L-PERIPH} & + \end{bmatrix}
\end{bmatrix}
$$

The phrasal rules, such as *subj-head-rule* and *comp-head-rule*, are classified into subrules, which inherit from two types of head-phrases (i.e. *subj-head-phrase* and *comp-head-phrase*) and optionally *topic-comment*. This type hierarchy is presented in Figure 12.1, in which there are two factors that have an influence on branching nodes; *wa* or (*n*)*un*-marking (i.e. *top-*) and scrambling (i.e. *scr-*).

This tripartite strategy might be controversial in that several types of headed rules are introduced. In the spirit of HPSG, reducing the number of rules should be considered in order to avoid redundancy. From this point of view, the six grammatical rules presented in Figure 12.1 might look rather superfluous. Nevertheless, this dissertation pursues this strategy for several reasons. First, Japanese and Korean are typical topic-prominent languages in which expressing topics plays an important role in configuring sentences (Li and Thompson, 1976; Sohn, 2001). Accordingly, it is my belief that the use of *topic-comment* as one of the major phrase structure types is never ill-conceived in creating Japanese and Korean grammars. Second, if we did not refer to the marking system (i.e. MKG), we would allow too wide an interpretation of scrambled constructions. That is, it would be almost impossible to narrow down the information structure meaning that *wa* and (*n*)*un* inherently carry (i.e. *contrast-or-topic*), if it were not for such discrimination. One alternative analysis would be to treat topicalized and scrambled constituents as a *head-filler-phrase*. However, this is also poor at handling scrambling. Such a *head-filler*-based analysis predicts long-distance dependency (i.e. scrambling across clause boundaries), but it is not likely to happen. Furthermore, the basic *head-comp* and *head-subj* properties are still encoded in single types, and these types are cross-classified

with others to give the more specific rules. That means that there are no missing generalizations. It seems clear that the tripartite strategy is well-motivated and is the most effective way to manipulate information structure in Japanese and Korean.

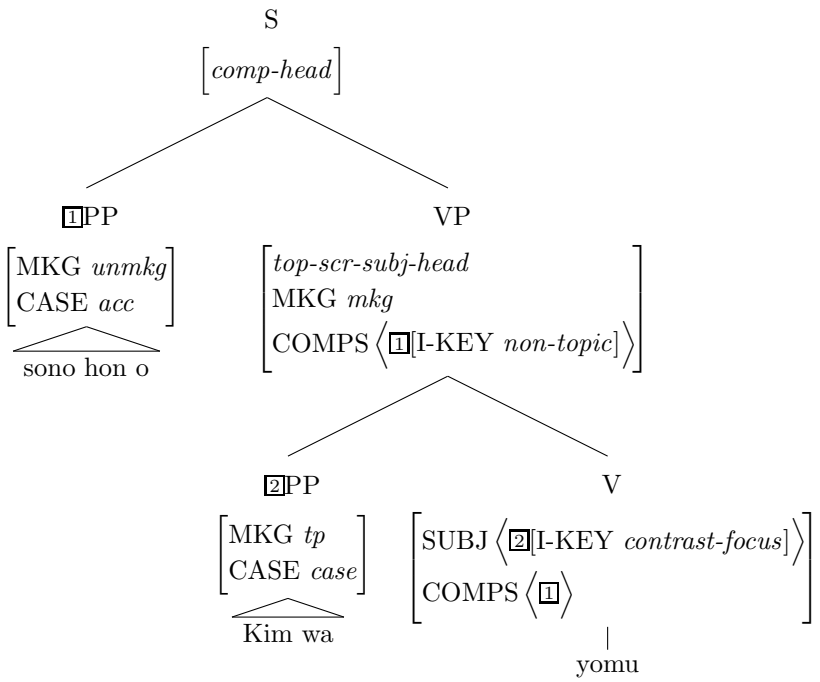
More specific information structure values are assigned by each grammatical rule, which has more specific constraints on both HEAD-DTR and NON-HEAD-DTR. For example, *top-scr-subj-head* and *top-scr-comp-head* impose a value on NON-HEAD-DTR as shown in (21).[10]

(21) a.
$$\begin{bmatrix} \textit{top-scr-subj-head} \\ \text{HD}\,|\,\text{VAL}\,|\,\text{COMPS}\left\langle \begin{bmatrix} \end{bmatrix} \right\rangle \\ \text{NHD}\,|\,\text{ICONS-KEY}\ \textit{contrast-focus} \end{bmatrix}$$

b.
$$\begin{bmatrix} \textit{top-scr-comp-head} \\ \text{HD}\,|\,\text{VAL}\,|\,\text{COMPS}\,\langle\rangle \\ \text{NHD}\,|\,\text{ICONS-KEY}\ \textit{contrast-topic} \end{bmatrix}$$

On the other hand, grammatical rules whose NON-HEAD-DTR is non-topicalized (e.g. *subj-head* and *comp-head*) have to place [MKG|TP *na-or-−*] onto the NON-HEAD-DTR, whose information structure value (i.e. ICONS-KEY) comes from the lexical information of the case markers (i.e. *non-topic*) and the null marker (i.e. *non-focus*). Consequently, the parse trees and dependency graphs for (18b) and (18d) are illustrated in (22) and (23), respectively.

(22) a.



[10]ICONS-KEY is doing some valuable work here, because it lets both the phrase structure rules and the lexical rules/entries contribute partial information to the same ICONS element.

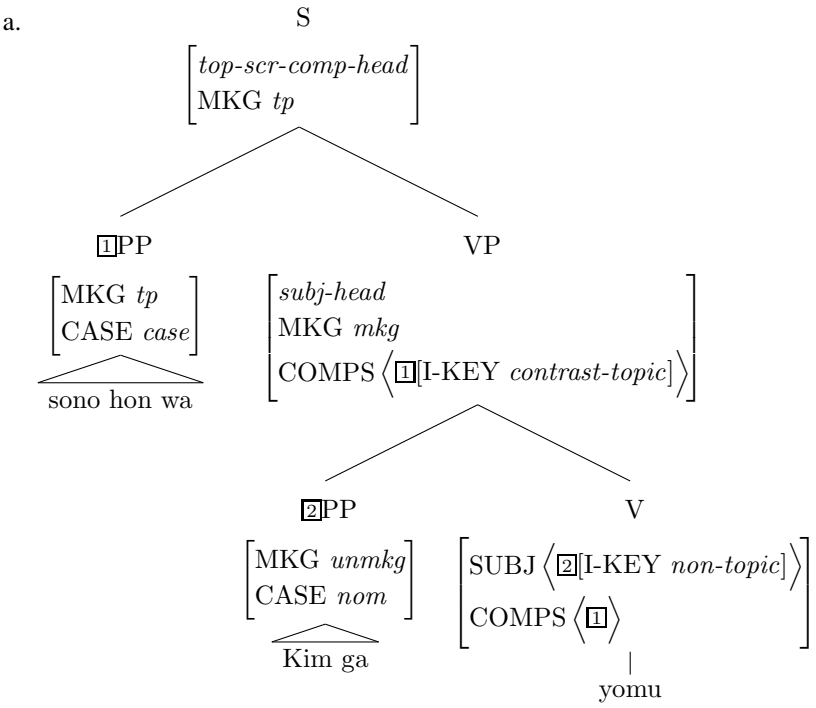
b. *contrast-focus*

sono hon-o Kim-wa yomu.

non-topic

In (22a),[11] the *wa*-marked subject *Kim* is combined with the verb *yomu* 'read'. This combination is an instance of *top-scr-subj-head* which requires [MKG *tp*] of the NON-HEAD-DTR (i.e. *Kim wa*) and assigns *contrast-focus* to the ICONS-KEY of the NON-HEAD-DTR. Since there is no specific constraint on MKG in this phrase structure, the value of MKG remains underspecified (i.e. *mkg*). Next, the VP takes *sono hon o* 'the book' as its complement. Because the fronted object is not *wa*-marked, its information structure meaning is still represented as *non-topic*, which comes from the case-marking adposition *o* 'ACC'.

(23) a.



b. *non-topic*

sono hon-wa Kim-ga yomu.

contrast-topic

On the other hand, since the scrambled object in (23a) is *wa*-marked, the top node is an instance of *top-scr-comp-head*, which assigns *contrast-topic* to the complement. Thus, topic falls on the object, while the case-marked subject conveys *non-topic*.

---

[11] The reason why *sono hon o*/*wa* and *Kim ga*/*wa* are structured as PPs, not NPs is given in Yatabe (1999) and Siegel (1999).

To summarize, scrambling in Japanese and Korean has to do with both lexical markers (e.g. *wa* and (*n*)*un*) and constraints on *topic-comment*. In order to systematize the different values that scrambled arguments and non-scrambled arguments have with respect to information structure, this dissertation proposes a tripartite strategy using cross-classification of three phrase structure types: *head-subj-phrase*, *head-comp-phrase*, and *topic-comment*. [MKG *tp*] is used to control the combination of the different phrase structure rules (and lexical markers) so that the scrambled and non-scrambled versions can be detected and related to their appropriate *info-str* values.

12.4 Cleft Constructions

Clefting is a special syntactic operation expressing focus in a marked way. Quite a few languages, including English, have the focus-related syntactic device called cleft constructions. It is true that all languages have at least one or more ways to express focus, but it is unlikely that all languages have cleft constructions. Cleft constructions normally involve relative clauses, which are not yet implemented in the LinGO Grammar Matrix system. For these reasons, clefts are not included into the information structure grammar library for the customization system. This subsection, instead, deals with how cleft constructions are analyzed within the HPSG/MRS formalism on a theoretical basis.[12]

12.4.1 Properties

Cleft constructions are regarded as showing different behaviours from ordinary focused constructions in syntax as well as semantics. In a nutshell, clefts are associated with exhaustive focus, which renders (24b) infelicitous.

(24) a. JOHN laughed, and so did MARY.

b. #It was JOHN who laughed, and so did MARY. (Velleman et al., 2012, p. 444)

[12]The constraints this subsection provides were tested with a former version of ERG (English Resource Grammar, Flickinger 2000, *ver*. 1111) in order to check whether ICONS can successfully replace the current representation (using the 'discourse relation' _be_v_itcleft_rel) and whether focus in *it*-clefts can be correctly represented via ICONS. The current version of ERG (1212) includes ICONS, but its structure is not exactly the same as what I propose in this dissertation. In future work, I will incorporate information structure into the ICONS in the current version of ERG, and conduct an evaluation on information structure-based machine translation using the resource grammar.

Kim (2012b), in a similar vein, argues clefts cannot coincide with lexical items that conflict with exhaustive focus. For example, *even* cannot be used in the focused XPs of cleft constructions as exemplified below.

(25) *It was even the advanced textbook that the student read. (Kim, 2012b, p. 48)

However, not all identificational foci are always realized as cleft constructions. Identificational foci can be conveyed in some languages using marked word order, as exemplified in the examples in Hungarian (26a) and Standard Arabic (27a). That is, clefting is just a sufficient condition for expressing identificational focus, not a necessary condition.

(26) a. Mari EGY KALAPOT nézett ki magának.
Mary a hat.ACC picked out herself.ACC
'It was A HAT that Mary picked for herself.' [hun] (É. Kiss, 1998, p. 249)

(27) a. RIWAAYAT-AN ʔallat-at Zaynab-u
novel-ACC wrote-she Zaynab-NOM
'It was a NOVEL that Zaynab wrote.' [arb] (Ouhalla, 1999, p. 337)

Other than the difference in semantics, as previously exemplified in Hungarian (26) clefts have a different property in prosodic patterns. Gussenhoven (2007) offers an analysis of cleft constructions in English with respect to information status. The clefted and non-clefted constituents are optionally accented. If the non-clefted constituent is accented, then clefts cause the non-clefted constituent to be interpreted as reactivated information (as presented in the first pair of (41)).[13] On the other hand, if the non-clefted constituent is unaccented, and the clefted one bears the accent as given in the second pair of (41), the clefted and non-clefted part denote new/old information, respectively. It is impossible to have both clefted and non-clefted constituents deliver new information at the same time.

(28) Q: Does Helen know JOHN?
A: It is John/JOHN she DISLIKES.
Q: I wonder who she dislikes.
A: It is JOHN she dislikes. (Gussenhoven, 2007, p. 96)

[13]Gussenhoven (2007) analyzes that the first reply in (41) implies Helen's disfavour to somebody had been discussed recently.

12.4.2 Subtypes

Clefts can be classified into subtypes.[14] These include *it*-clefts, *wh*-clefts (a.k.a. pseudo clefts), and inverted *wh*-clefts (Kim, 2007). Each of them is exemplified in (29), whose skeletons are represented in (30) in turn.[15]

(29) a. *It*-clefts: In fact it's their teaching material that we're using... <S1A-024 #68:1:B>

b. *Wh*-clefts: What we're using is their teaching material.

c. Inverted *wh*-clefts: Their teaching material is what we are using. (Kim, 2007, p. 217)

(30) a. *It*-clefts: It + be + XP_i + cleft clause

b. *Wh*-clefts: Cleft clause + be + XP_i

c. Inverted *wh*-clefts: XP_i + be + cleft clause (Ibid. p. 218)

Kim (2007), building upon this taxonomy, provides a corpus study with reference to ICE-GB (a syntactically annotated corpus of British English, Nelson et al. 2002). Out of 88,357 text units, *it*-clefts occur 422 times (0.47%), *wh*-clefts occur 544 times (0.61%), and inverted *wh*-clefts occur 537 times (0.60%). Other than NPs, various phrasal categories can be focused in cleft constructions. These include APs, AdvPs, PPs, VPs, and even CPs. For example, *it*-clefts can take various types of XPs as the focused constituent.

(31) a. NP: It was [the gauge] that was the killer in the first place. <S1A-010 #126:1:B>

b. AdvP: And it was [then] that he felt a sharp pain. <S2A-067 #68:1:A>

c. Subordinate Clause: It wasn't [till I was perhaps twenty-five or thirty] that I read them and enjoyed them <S1A-013 #238:1:E> (Ibid. p. 220)

One interesting point is that there is a restriction on categorical choice. Kim (2007, p. 220–223) presents the frequency as shown in Table 12.2.

---

[14] From a functional perspective, Kim and Yang (2009) classify cleft constructions into predicational, identificational, and eventual types. Similarly Clech-Darbon et al. (1999) classify cleft constructions in French (basically realized in the form as *C'est ... que/qui ...*) into four types: basic, broad event-related focus, broad presentational focus, exclamatory comment. This taxonomy is not used in the present work.

[15] (29a) is originally taken from the ICE-GB corpus (Nelson et al. 2002), and the bracketed expression after the sentence stands for the indexing number. (29a) is paraphrased into (29b-c) by Kim (2007).

Table 12.2: Frequency of the three types of clefts (Kim, 2007)

Types of XP	NP	AP	AdvP	PP	VP	CP
it-cleft	324	0	18	65	0	16
wh-cleft	136	19	3	14	19	275
inverted *wh*-cleft	518	0	0	0	0	19

Table 12.2 shows the following: *It*-clefts seldom take verbal items as the pivot XP, while *wh*-clefts do not show such a restriction. Inverted *wh*-clefts exclusively put focus on NPs, but there are some exceptional cases in which the focused constituent is clausal as exemplified below.

(32) a. [To feel something you have written has reached someone] is what matters. <S1A-044 #096>

b. [What one wonders] is what went on in his mind. <S1A-044 #096> (Ibid. p. 222)

Though various types of phrases can be focused in clefts, Velleman et al. (2012) argues that only a portion of the pivot is assigned genuine focus. This implies that clefts involve narrowly focused items inside of the pivot XP as Beaver and Clark (2008) argue that the clefts raise an exhaustive reading as a focus-sensitive operator.

This dissertation is exclusively concerned with *it*-clefts, basically following the analysis provided by the ERG. Implementing *it*-clefts in TDL requires a categorical constraint indicated in Table 12.2: non-clausal verbal items are not used as the pivot XPs. Pseudo cleft constructions, such as *wh*-clefts and inverted *wh*-clefts, are left to future work, because free relative clauses need to be separately implemented in relation to ICONS.

12.4.3 Components

Cleft constructions across languages are made up of four components (Gundel, 2002; Kim and Yang, 2009; Kim, 2012b); (i) placeholder, (ii) copula, (iii) pivot XP, and (iv) cleft clause.

(33) [It] [is] [the dog] [that barks].
placeholder copula pivot XP cleft clause

Some languages constitute cleft constructions in the same way as English. For instance, a basic cleft sentence (34) in Norwegian is comprised of all the four components.

(34) Det var Nielsen som vant.
It was Nielsen that won [nor] (Gundel, 2002, p. 113)

However, the first two components are not necessarily used in all languages that employ clefts. The following subsections explore these four components one after another.

Placeholders

For English, placeholders in cleft constructions are usually realized as expletives (i.e. *it* in English) (Pollard and Sag, 1994), but some counterexamples to this generalization exist. Kim (2012b) presents a dialogue in which *it* in a cleft construction is made use of as a referential pronoun, rather than an expletive. Han and Hedberg (2008) exemplify a specific context in which demonstrative pronouns (e.g. *this* and *that*) can be substituted for *it*. Moreover, some languages do not employ any placeholder. For example, clefts in Arabic languages and in Wolof[16] have no counterpart to *it*. In the following examples (Standard Arabic in (35a), Moroccan Arabic in (35b), and Wolof in (35c)), the focused constituents occupy the first position of the sentence, followed by pronominal copulae, such as *hiyya* in (35a) and *huma* in (35b) or an ordinary copula *la* in (35c), and then followed by cleft clauses.

(35) a. ZAYNAB-u hiyya llatii ʔallaf-at l-riwaayat-a.
Zaynab-NOM PRON.she RM wrote-she the-novel-ACC
'It was ZAYNAB who wrote the novel.' [arb]

b. L-WLAD huma Hi sarrd-at (-hum) Nadia.
the-children PRON.they RM sent-she (-them) Nadia
'It was the CHILDREN that Nadia sent.' [ary] (Ouhalla, 1999, p. 341)

c. Fas wi la jaakat bi jënd
horse the COP.3SG merchant the buy
'It is the horse (that) the merchant bought.' [wol] (Kihm, 1999, p. 256)

This dissertation assumes that the placeholder for clefts is conditioned language-specifically. As for the placeholder *it* in English clefts, it is assumed to be a semantically vacuous pronoun (i.e. an expletive) that introduces no EP and involves an empty ICONS list (i.e. *no-icons-lex-item*).

[16] a Niger-Congo language, spoken in Senegal.

Copulae

Copulae participate in cleft constructions. However, because not all languages employ copulae, the use of copulae is language-specific. For example, Russian does not use any copula in clefts, as exemplified below.

(36) Eto [Boris] vypil vodku.
it Boris drank vodka
'It is Boris-FOC (who) drank the vodka.' [rus] (King, 1995, p. 80)

Thus, (ii) copula is not a mandatory cross-linguistic component for constructing clefts.

On the other hand, it is necessary to determine the grammatical status of the copulae in clefts. Kim (2012b) surveys two traditional approaches to cleft constructions: (a) extraposition (Gundel, 1977), (b) expletive (É. Kiss, 1999; Lambrecht, 2001). First, the extraposition analysis assumes *it*-clefts stem from *wh*-clefts; a free relative clause in a *wh*-cleft construction is first extraposed (i.e. right-dislocated) leaving *it* in the basic position, and then *what* in the extraposed clause turns into an ordinary relative pronoun such as *that*. Second, the expletive analysis assumes that the pronoun *it* is a genuine expletive (i.e. generated *in situ*), and the cleft clause is directly associated with the pivot XP. For example, a simple cleft sentence *It is the dog that barks.* can be parsed into (37a-b) respectively. In (37a), the copula *is* takes two complements; one is the pivot XP *the dog*, and the other is the cleft clause *that barks*. In contrast, the copula in (37b) takes only one complement, and the pivot XP and the cleft clause are combined with each other before dominated by the copula.

(37) a. [It [[$_{\text{HEAD-DTR}}$ is the dog] [$_{\text{NON-HEAD-DTR}}$ that barks]]].

b. [It [is [[$_{\text{HEAD-DTR}}$ the dog] [$_{\text{NON-HEAD-DTR}}$ that barks]]]].

Kim provides a hybrid approach between the extraposition analysis and the expletive analysis. For him, the focused XP constitutes a *cleft-cx* with the following cleft clause first, and then the copula takes the *cleft-cx* as a single complement. That is, his analysis takes (37b) as the proper derivation of the cleft sentence.

The ERG parses a cleft sentence in a way similar to the extraposition analysis; along the lines of the parse in (37a). That is, the ERG analyzes *it*-clefts in such a way that the focused XP complements the copula, and the construction introduces a constructional content (i.e. C-CONT) whose

EP is '_be_v_itclefts_rel', and then the VP (i.e. [copula + XP]) is complemented once again by cleft clauses. This follows the traditional approach in which the copula in clefts takes two complements; one for the focused constituent, and the other for the cleft clause.

On one hand, the two HPSG-based analyses have one thing in common: The copula in clefts is a single entry, lexically different from ordinary copulae. On the other hand, they show a difference in the ARG-ST values of the cleft copula. Kim (2012b) argues that the focused XP and the cleft clause is a syntactic unit (as presented in (37b)), which means the cleft clause does not directly complement the copula. That is, the cleft copula has only one element (*cleft-cx*) in its VAL|COMPS list. In contrast, the cleft copula in the ERG is syntactically a ditransitive verb that takes two complements, whose second complement is clausal.[17] The ARG-ST of the *it*-cleft copula is <*it*, XP, CP>.

Pivot XPs

The focused XPs in clefts are assigned *focus*. This constraint implies the pivot XP can be interpreted as containing either *semantic-focus* or *contrast-focus*. Can the pivot XP be contrastively focused? The linguistic surveys say the answer is yes across languages, and this is supported by the fact that clefts pass the correction test (Gryllia, 2009). Gracheva (2013) provides a corpus study with reference to the Russian National Corpus (Grishina 2006), and substantiates that cleft constructions in Russian are compatible with *contrast-focus* using the correction test as shown in (38). Her analysis is also applicable to other languages, such as French in (39) and Mandarin Chinese in (40), as well. Li (2009), especially, regards the *shì ... de* constructions exemplified (40) as the canonical syntactic means of expressing contrastive focus in Mandarin Chinese.

(38) Q: Eto Ivan vypil vodku?
It Ivan drank vodka
'(Was) it Ivan (that) drank vodka?' [rus]

A: (Net.) Eto [Boris] vypil vodku.
(No.) It Boris drank vodka
'(No). It (was) Boris (that) drank vodka.' [rus] (Gracheva, 2013, p. 118)

[17]For example, *tell* in *Kim told Sandy that Pat slept.* is an instance of *clausal-third-arg-ditrans-lex-item* in the current `matrix.tdl` of the LinGO Grammar Matrix system. The cleft copula should be a subtype of the lexical type with some additional constraints on the complements.

(39) Q: Ta fille est tombée dans l'escalier?
Did your daughter fall down the stairs? [fra]
A: Non. c'est le petit qui est tombé dans l'escalier.
No, it's the youngest one [+masc.] that fell down the stairs.
[fra] (Clech-Darbon et al., 1999, p. 84)

(40) Ta shi zai Beijing xue yuyanxue de, bu shi zai Shanghai xue de.
3SG be at Beijing learn linguistics DE NEG be at Shanghai learn DE
'It's in Beijing that he studied linguistics, not in Shanghai. [cmn] (Paul and Whitman, 2008, p. 414)

The corpus study presented in Part III shows that we are incapable of identifying whether the pivot XP is contrastively or non-contrastively focused without reference to the context. Hence, the focused constituents in cleft constructions are assigned the immediate supertype of both *semantic-focus* and *contrast-focus*.

Cleft Clauses

The semantic head of cleft clauses (i.e. the verbs) could be assigned *bg* in line with previous studies which analyze cleft constructions as a *focus-bg* realization (Paggio, 2009). The principle motivation for this comes from the fact that cleft clauses can be freely omitted (Kim, 2012b). However, the first reply in (41), in which the verb in cleft clauses bears the A-accent (i.e. focused), can serve as a counterexample to this generalization. Thus, the verbs in cleft clauses are not specified with respect to information structure meanings.

(41) Q: Does Helen know JOHN?

A: It is John/JOHN she DISLIKES.

Q: I wonder who she dislikes.

A: It is JOHN she dislikes. (Gussenhoven, 2007, p. 96)

There are some additional properties of cleft clauses to be considered. Kim (2012b) claims that cleft clauses show a kind of ambivalent behaviour between restrictive relatives and non-restrictive relatives: the focused XP and the cleft clause are basically combined with each other in the restrictive way, but the combined phrase does not look like a canonical restrictive relative in that

proper nouns and pronouns can be used for the focused XP. Though his argument sounds intriguing, this dissertation does not take this ambivalence into account in revising the implementation of cleft constructions in the ERG, because the basic approach to clefts is different (i.e. *cleft-cx vs*. two complements of the cleft copula).

12.4.4 It-clefts in the ERG

It-clefts in the ERG are constrained by only the specific type of copulae *itcleft-verb*. Building upon the analyses discussed hitherto, I present a revised version of *itcleft-verb* tested in the ERG (*ver*. 1111). The original constraint in the ERG are represented in (42).

(42) *itcleft-verb* →

$$\begin{bmatrix} \text{VAL} \begin{bmatrix} \text{SUBJ} & \left\langle \left[\textit{it-expl} \right] \right\rangle \\ \text{COMPS} & \left\langle \begin{bmatrix} \text{HOOK} \mid \text{LTOP} \ \boxed{1} \\ \text{VAL} \begin{bmatrix} \text{SUBJ } \textit{*olist*} \\ \text{COMPS} \left\langle \ \right\rangle \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{HEAD} & \textit{verb} \\ \text{HOOK} \mid \text{LTOP} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{LKEYS} \mid \text{KEYREL} \begin{bmatrix} \text{PRED} & \textit{\_be\_v\_itcleft\_rel} \\ \text{ARG2} & \boxed{1} \end{bmatrix} \end{bmatrix}$$

(43) is my version which places several constraints on *it*-clefts in accordance with my analysis presented hitherto.

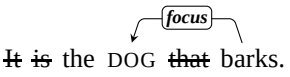
(43) *itcleft-verb* →

$$\begin{bmatrix} \text{VAL} \begin{bmatrix} \text{SUBJ} & \left\langle \left[\textit{it-expl} \right] \right\rangle \\ \text{COMPS} & \left\langle \begin{bmatrix} \text{HOOK} & \begin{bmatrix} \text{LTOP} & \boxed{1} \\ \text{INDEX} & \boxed{2} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SUBJ } \textit{*olist*} \\ \text{COMPS} \left\langle \ \right\rangle \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{HEAD} & \textit{verb} \\ \text{HOOK} & \begin{bmatrix} \text{LTOP} & \boxed{1} \\ \text{INDEX} & \boxed{3} \\ \text{CLAUSE-KEY} & \boxed{3} \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{C-CONT} \mid \text{ICONS} \left\langle \ ! \begin{bmatrix} \textit{focus} \\ \text{TARGET} & \boxed{2} \\ \text{CLAUSE} & \boxed{3} \end{bmatrix} ! \right\rangle \\ \text{LKEYS} \mid \text{KEYREL} \mid \text{ARG2} \ \boxed{1} \end{bmatrix}$$

The most significant difference between them is that ICONS replaces the representation using the 'discourse relation' _be_v_itcleft_p_rel in (42). The focused XPs are assigned *focus* within ICONS,

whose CLAUSE value is linked to the cleft clauses. (42) and (43) have the categorical restriction on the focused XPs in common. This restriction is specified in VAL of the first complement. According to the corpus study Kim (2007) provides, APs and VPs cannot be focused in *it*-clefts as indicated in Table 12.2, while CPs can be used as the focused XP. Other phrasal types, such as NPs, AdvP, and PPs, can freely become the first complement of *itcleft-verb*. This restriction is specified using **olist** in VAL|SUBJ and an empty list of VAL|COMPS.

Building on the AVM in (43), (44) exemplifies how cleft constructions are represented via ICONS. The focused element DOG has a *focus* relation to the cleft clause, the element *barks* in the cleft clause remains underspecified. Note that the expletive ~~*it*~~ and the copula ~~*is*~~ are semantically empty, and thereby they cannot participate in ICONS.

(44) 

12.5 Passive Constructions

This subsection is exclusively concerned with passive constructions in English, in order to revise the related types in the ERG with respect to information structure. Nonetheless, a similar version of revision can be applied to other languages.

Passive constructions have to do with this dissertation in terms of two aspects: namely information structure and semantics-based machine translation.[18]

First, passivization is (partially) relevant to information structure. It has been reported that some languages, such as Spanish (Casielles-Suárez, 2003), exhibit relationship between passivization and the articulation of information structure. Though such a straightforward relationship between these concepts does not hold for all human languages, there seems to be at least some connection. What is the motivation for using passive forms? For this question, it is necessary to look at promoted arguments and demoted arguments differently. One might think that one function of passive is to place a different argument in subject position so that it can be the topic (given the general tendency to align topic with subject). However, the promoted arguments in passives are not always assigned *topic*. For example, the promoted argument in the following sentence conveys a focus meaning,

[18]The argument in this subsection largely comes from what I discussed with Dan Flickinger and Emily M. Bender. The examples in (45) and (46) are also provided by Emily M. Bender.

which is exclusive from a topic meaning in that the promoted argument *the book* corresponds to the *wh*-word in the question.

(45) Q: What was found by Sandy?
A: The book was found by Sandy.

However, the promoted arguments are not always interpreted as *focus*, because some aspect of passivization is clearly motivated by the desire to put something else other than the agent into the canonical topic position (i.e a subject position).

(46) They were looking all over for the book. Finally, it was found by Sandy.

As a result, the best we can say about the promoted arguments is that they are not background (i.e. *focus-or-topic*). At the same time, the demoted arguments, if they appear overtly, have to be marked as *non-topic*. In particular, the demoted arguments can hardly serve as a topic of a sentence (at least in English), because NPs with *topic* are normally preferable in sentence-initial position.

Second, active/passive pairs are related to machine translation as well as monolingual paraphrasing. Presumably they share the same truth-conditions monolingually, and exhibit structural divergence multilingually. For example, in English, passives are used productively and constraints on passivization are relatively weak. In contrast, Japanese and Korean, which tend to downplay the role of passives, have stronger constraints on passivization.[19] In the ERG (*ver.* 1111), the passive constructions constructionally introduce an EP (i.e. using C-CONT), whose predicate is '_parg_d_rel'. The original constraint using the 'discourse relation' is represented as follows.

(47) *passive-verb-lex-rule* →

$$\begin{bmatrix} \text{VAL} & \begin{bmatrix} \text{SUBJ} & \left\langle \left[\text{INDEX} \quad \boxed{2} \right] \right\rangle \\ \text{COMPS} & \left\langle ..., \left[\text{INDEX} \quad \boxed{1} \right] \right\rangle \end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix} \text{RELS} & \left\langle ! \begin{bmatrix} \text{PRED} & \textit{parg\_d\_rel} \\ \text{ARG1} & \boxed{1} \\ \text{ARG2} & \boxed{2} \end{bmatrix} ! \right\rangle \\ \text{HCONS} & \left\langle ! \; ! \right\rangle \end{bmatrix} \end{bmatrix}$$

[19] Song and Bender (2011) look at translation of active/passive pairs to confirm how information structure can be used to improve transfer-based machine translation.

This method cannot capture the point that active/passive pairs are semantically equivalent, and thereby they cannot be paraphrased into each other monolingually. This is an analysis which disregards the fact that active/passive pairs are truth-conditionally equivalent, provided that the demoted argument is overt in the passive. In translation, passive sentences in English sometimes need to be translated into active sentences in other languages, such as Japanese and Korean. Moreover, using a discourse relation such as '_parg_d_rel' is redundant in that this can be replaced by an information structure value.

My alternative method is as follows. The information structure of promoted/demoted arguments is still articulated in the lexical rule which passivizes main verbs. However, the EP involving the discourse predicate (i.e. 'parg_d_rel') is removed from the lexical rule, and instead two *info-str* values are inserted into C-CONT. The TARGET value of the first element has a coreference with ARG2, and that of the second one is co-indexed with ARG1. In addition, the preposition *by* is specified as a semantically empty item. An AVM of the type responsible for passivization is presented as (48).[20] Note that the first element in SUBJ and the last element in COMPS specify their *info-str* values as *focus-or-topic* and *non-topic* respectively.

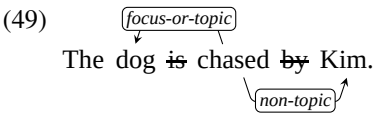
(48) *passive-verb-lex-rule* →

$$\begin{bmatrix} \text{VAL} & \begin{bmatrix} \text{SUBJ} & \left\langle \begin{bmatrix} \text{INDEX} & \boxed{2} \\ \text{ICONS-KEY} & \boxed{3}\left[\textit{focus-or-topic}\right] \end{bmatrix} \right\rangle \\ \text{COMPS} & \left\langle ..., \begin{bmatrix} \text{INDEX} & \boxed{1} \\ \text{ICONS-KEY} & \boxed{4}\left[\textit{non-topic}\right] \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix} \text{RELS} & \left\langle ! \begin{bmatrix} \text{ARG1} & \boxed{1} \\ \text{ARG2} & \boxed{2} \end{bmatrix} ! \right\rangle \\ \text{HCONS} & \left\langle ! \; ! \right\rangle \\ \text{ICONS} & \left\langle ! \; \boxed{3}\left[\text{TARGET } \boxed{2}\right], \boxed{4}\left[\text{TARGET } \boxed{1}\right] ! \right\rangle \end{bmatrix} \end{bmatrix}$$

A sample representation of a passive construction is accordingly sketched out in (49), in which the

[20] In the future when prosody information is modelled in the ERG and thereby accents for marking focus and topic are employed in the grammar, the constraints on C-CONT|ICONS in (48) should be changed. If rules for dealing with prosody are used, the rules will be responsible for introducing the ICONS elements for constraining information structure values on promoted and demoted arguments. In this case, *passive-verb-lex-rule* will have an empty C-CONT|ICONS, but it still will assign specific values to the ICONS-KEYs of the promoted arguments (*focus-or-topic* on the first element of SUBJ) and demoted arguments (*non-topic* on the last element of COMPS). Because prosodic information has not yet been used in the current ERG, tentatively (48) puts the ICONS elements into C-CONT herein.

auxiliary copula ~~*is*~~ and the preposition ~~*by*~~ are semantically and informatively empty.

(49) 

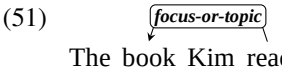
12.6 Focus/Topic Fronting

(50) exemplifies a focus/topic fronting construction in English: (50a) is the unmarked sentential form which is devoid of any specific information structure markings. On the other hand, the object *the book* in (50b) occupies the sentence-initial position, and the remaining part of the sentence has a syntactic gap for the preposed object.

(50) a. Kim reads the book.

b. The book Kim reads.

The point at issue in analyzing focus/topic fronting constructions is to determine which information structure meaning(s) the preposed argument gives. (50b) in itself sounds ambiguous between the two readings: One assigns a topic reading to *the book*, which bears a likeness to *'as for ...'* construction. The other, similarly to *it*-clefts, has a focus reading on the preposed argument. The choice between them is largely conditioned by the contextual situation that utterances prior to the current sentence create, which is actually infeasible to measure in sentence-based language processing (Kuhn, 1996). Thus, as long as we do not deploy an extra device to resolve the meaning with respect to the context, the information structure value on *the book* should be underspecified so that it can cover both meanings. The lowest supertype of both *focus* and *topic* in Figure 9.1 is *focus-or-topic*, which implies the associated constituents (e.g. *the book* in (50b)) are informatively interpreted as either focus or topic. (51) is illustrative of the schema that (50b) has.

(51) 

12.7 Dislocation

Unlike focus/topic fronting constructions, dislocation constructions do not have any syntactic gap irrespective of whether the peripheral topic is sentence-initial (i.e. left dislocation) or sentence-final

(i.e. right dislocation). The example structurally similar to (50b) is provided in (52)[21], in which (i) an intonational break at the phonological level intervenes between the left-peripheral NP *the book* and the rest of the utterance, and (ii) a resumptive pronoun *it* corresponding to *the book* satisfies the object of *reads*.

(52) a. The book, Kim reads it.

b. Kim reads it, the book.

The book in this case is an external topic that is not inside the sentence. It is regarded as containing *frame-setting* according to the cross-linguistic study offered in the previous chapters. In other words, its pragmatic role is to narrow down the domain of what is being referred to.

In the analysis of dislocation, there is one more factor to be considered; agreement between the topicalized NP and the corresponding pronoun inside the head sentence. For example, in (52) only the third singular pronoun *it* which agrees with *the book* can be resumptive. In languages which exhibit rich morphology (e.g. Italian (Cinque, 1977; Rizzi, 1997), Spanish (Rivero, 1980; Zagona, 2002; Bildhauer, 2008), German (Grohmann, 2001), Modern Greek (Alexopoulou and Kolliakou, 2002), Czech (Sturgeon, 2010), etc.) the choice of resumptive pronouns matters. The options are: (i) (clitic) left dislocations[22] and (ii) hanging topics. The resumptive pronouns in left dislocation constructions have to agree perfectly with the dislocated NP in person, number, gender, case, etc., whereas a hanging topic and its corresponding pronoun do not agree with each other. This implies that hanging topics have a looser relationship with the remaining part of the sentence than left dislocations (Frascarelli, 2000).

(53) a. [Seinen$_i$ Vater], den mag jeder$_i$.
his-ACC father RP-ACC likes everyone
'His father, everyone likes.'

[21]Commas after topicalized NPs are not obligatorily used, and are mainly attached just as a preferable writing style for the reader's convenience. On the other hand, there should be a phonetic pause between the topicalized NPs and the main sentence in speech. The pause information should be included in the typed feature structure of PHON, because information structure-based TTS (Text-To-Speech) and ASR (Automatic Speech Recognition) systems can use it to improve performance.

[22]In terms of interpretation of left dislocations, there are two different points of view. One regards left dislocations as an ordinary construction that expresses topic (Zubizarreta, 1998; Zagona, 2002; Casielles-Suárez, 2004). The other argues that left dislocations have to do with focus projection and thereby can sometimes constitute a wide focus construction (Alexopoulou and Kolliakou, 2002; Gutierrez-Bravo, 2006; Bildhauer, 2008). I claim that the former looks right because the data the previous studies analyze are all a kind of topic construction.

b. [Sein$_i$ Vater], jeder$_{*i/k}$ mag den/ihn.
his-NOM father everyone likes RP/him-ACC
'His father, everyone likes him.' [ger] (Grohmann, 2001, p. 92)

c. Honzu, toho ještě neznám.
Honza.ACC that.ACC still NEG-know.1SG
'Honza, I still don't know him.'

d. Anička? Té se nic nestalo.
Anička.NOM that.DAT REFL-CL nothing NEG-happened
'Anička? Nothing happened to her.' [cse] (Sturgeon, 2010, p. 288)

(53a-b) are examples of left dislocation and hanging topics in German, respectively. In (53a), the accusative on the dislocated NP *Seinen Vater* agrees with that on the resumptive pronoun *den*. By contrast, a hanging topic *Sein Vater* in (53b) is in nominative, which does not agree with the resumptive pronoun in accusative. The same holds for (53c-d) in Czech; both *Honza* and its resumptive pronoun *toho* in (53c) are in accusative, while there is no agreement between the lefthand NP *Anička* and *Té* in (53d).

In movement-based analyses, (clitic) left dislocations and hanging topics are regarded as being configured via two different syntactic operations: Dislocated NPs in (clitic) left dislocations are originally realized inside the sentence, and move forward leaving resumptive pronouns with the same features. Hanging topics, by contrast, are base-generated *ab initio* without any agreement with their corresponding pronoun. Hanging topics in transformation-based studies are also assumed to have several additional characteristics (Frascarelli, 2000): (i) Only one hanging topic can show up in a sentence, (ii) hanging topics can appear only sentence-initially, (iii) if a hanging topic co-occurs with other topics in a sentence, it should be followed by the other topics (i.e. hanging topic first).[23] From this point of view, Cinque (1977) distinguishes English-like languages from Italian-like languages; the former employ only hanging topics, whereas the latter have both left dislocation and hanging topics.

In HPSG-based studies, agreement between a dislocated NP and its resumptive pronoun is modelled. For instance, in the following AVM taken from Bildhauer (2008, p. 350), a coreference $\boxed{3}$

[23]Frascarelli (2000), exploiting a corpus, provides some counterexamples to these properties that hanging topics presumably possess, which implies they are tendencies, rather than strict rules.

means the dislocated NP and its corresponding clitic in the main clause should be consistent, and another coreference $\boxed{4}$ indicates that they share the same INDEX.

(54)

$$\textit{clld-phrase} \Rightarrow \begin{bmatrix} \text{COMPS} & \langle\rangle \\ \text{CLITICS} & \boxed{1} \oplus \boxed{2} \\ \text{HEAD-DTR} & \begin{bmatrix} \text{COMPS} & \langle\rangle \\ \text{CLITICS} & \boxed{1} \oplus \left\langle \begin{bmatrix} \text{HEAD} & \boxed{3} \\ \text{INDEX} & \boxed{4} \\ \text{UPV} & \boxed{5} \end{bmatrix} \right\rangle \oplus \boxed{2} \end{bmatrix} \\ \text{NON-HD-DTR} & \left\langle \begin{bmatrix} \text{HEAD} & \boxed{3} \\ \text{INDEX} & \boxed{4} \\ \text{UPV} & \boxed{5} \\ \text{SPR} & \langle\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \right\rangle \end{bmatrix}$$

Because this dissertation does not employ a rigid distinction between (clitic) left dislocations and hanging topics, all these constraints can be fully covered in the proposal of this dissertation. That is, they can be merged into just one single type that assigns *contrast-topic* to the fronted constituent.

(55) a. *contrast-topic*
The book, Kim reads it.

b. *contrast-topic*
Kim reads it, the book.

As mentioned earlier, the present work does not fully implement focus/topic-fronting and dislocation in terms of how to build up this representation compositionally. In §14.2.4, several types of dislocated constituents are partially implemented using *head-filler-phrase* in order to constrain clause-initial and clause-final foci. Future work needs to look into how the *contrast-topic* element can be added into the ICONS list.

12.8 Summary

This chapter has delved into the specific forms of expressing information structure. First, focus sensitive items are classified into two subtypes; one assigns an information structure value to itself, and the other assigns a value to its adjacent item. Second, in terms of argument optionality, unexpressed arguments always bear *non-focus* because focused items cannot be elided. Third, scrambling in Japanese and Korean was addressed. This dissertation proposes a cross-classification of three phrase structure types, which refers to a value of MKG for looking at which lexical marker is used. Fourth,

an AVM responsible for cleft constructions in the ERG was revised, which signal *focus* (i.e. a plain focus) to the pivot XP in cleft constructions. Fifth, promoted and demoted arguments in passive constructions also have specific value of *info-str*: *focus-or-topic* for the former, and *non-topic* for the latter. Lastly, focus/topic-fronting constructions and two types of dislocations (i.e. (clitic) left dislocation and hanging topics) were examined. The fronted elements in OSV sentences in English have a value of *focus-or-topic*, because they can be interpreted as either *focus* or *topic*. On the other hand, dislocated NPs are assigned *contrast-topic* along the line of the analysis of previous studies.

Chapter 13

FOCUS PROJECTION

Focus projection refers to a phenomenon of information structure in which meaning of focus associated with specifically marked word spreads into the larger phrase(s) to which the word belongs. It has been normally said that a focus domain in a sentence has to contain at least one accented word, which functions as the core of focus meaning. That implies that focus projection is basically related to how F(ocus)-marking (normally realized with a specific pattern of prosody, such as the A-accent (H*) in English) co-operates with information structure meanings. The fundamentals of focus projection, suggested by Selkirk (1984, 1995) and Büring (2006), are summarized as follows. These definitions seem true at least in English in which prosody is mainly responsible for expressing focus.

(1) a. Basic Focus Rule: An accented word is F-marked.

b. Focus of a sentence: An F-marked constituent is not dominated by any other F-marked constituent.

c. Focus Projection: either (i) F-marking of the head of a phrase licenses F-marking of the phrase, or (or both) (ii) F-marking of an internal argument of a head licenses the F-marking of the head. (Büring, 2006, p. 322–323)

In this chapter, I would like to lay the groundwork for how an analysis based on ICONS (Individual CONStraints) and MKG (MarKinG) could eventually support an analysis of focus projection. Quite a few HPSG-based works on information structure are particularly concerned with focus projection, mostly based on the Focus Projection Principle as presented in (1). The previous studies have three points in common, but these points need to be taken into account in the context of creation of a computational model: First, they provide multiple parse trees for a single sentence in which focus projection may happen. The first section (§13.1) provides a counterargument to this strategy in representation. Second, previous studies claim that assignment of focus-marking accent plays an important role in calculating the extent of focus domain (§13.2). Third, distinctions between grammatical relations, such as peripheral *vs*. non-peripheral, head *vs*. non-head, are critically

used in constraining focus projection (§13.3). The last section (§13.4) makes an illustrative analysis of one sentence in which focus projection happens.

13.1 Parse Trees

Most previous approaches in the HPSG-based study on information structure provide multiple parse trees. In fact, a sentence that potentially involves focus projection sounds ambiguous in itself. For example, (2) may have at least three parse trees following the previous approaches.

(2) [$_f$ Kim [$_f$ gives Lee [$_f$ a BOOK]]].

From the point that one sentence can have multiple readings simultaneously, this method may be not so bad. However, this kind of approach does not work well in the context of computational processing. The main problem is such that multiple parse trees for a single sentence can have an adverse effect on system performance. A large number of parse trees decreases speed while an increase in ambiguity decreases accuracy, both detrimental to the system's goals That is, several external modules that enhance feasibility of computational grammars (e.g. reranking model) do not perform actively with such a large number of intermediate results. Thus, I argue that a single parse tree that potentially covers the whole meanings that the sentence can convey should be provided. The main mechanism to facilitate this flexibility is underspecification.

13.2 F(ocus)-marking

F-marking, which crucially contributes to formation of focus projection, has been presumed to have to do exclusively with prosody, as shown in (1a). That is to say, in previous literature, a set of specific accents (e.g. the A-accent in English) has been considered tantamount to F-marking. However, it is my position that bearing a specific accent is not a necessary condition, but a sufficient condition for F-marking: F-marking does not necessarily depend on whether the word is accented or not. Across languages, there are several examples in which focus projection is triggered by non-prosodic features.

Building on the phonological rules provided in (3) (already presented in §8.3), the focus prominence rule that Bildhauer (2007) presents is constrained as follows, which means the focused constituent has to contain the Designated Terminal Element (DTE) on the level of phonological UTter-

ance (UT). In the following rules, PHP is short for PHonological Phrase, IP for Intonational Phrase, RE for Right Edge, PA for Pitch Accent, and BD for BounDary tone.

(3) a. $\left[\text{PHP} \mid \text{DTE} \quad +\right] \rightarrow \left[\text{PA} \quad \textit{tone}\right]$

b. $\left[\text{PHP} \mid \text{DTE} \quad -\right] \rightarrow \left[\text{PA} \quad \textit{none}\right]$

c. $\left[\text{IP} \mid \text{RE} \quad +\right] \rightarrow \left[\text{BD} \quad \textit{tone}\right]$

d. $\left[\text{IP} \mid \text{RE} \quad -\right] \rightarrow \left[\text{BD} \quad \textit{none}\right]$

(4) $$\begin{bmatrix} \textit{sign} & \\ \text{SYNSEM} \mid \text{CONT} & \begin{bmatrix} \textit{mrs} \\ \text{RELS } \boxed{1} \end{bmatrix} \\ \text{IS} \mid \text{FOC} & \left\langle \boxed{1} \right\rangle \end{bmatrix} \rightarrow \left[\text{PHON} \quad \left\langle \ldots \left[\text{UT} \mid \text{DTE} \quad +\right] \ldots \right\rangle\right]$$

Bildhauer claims that the schematic AVM (4) can be presumably applied to most human languages that mark focus by means of prosody. Furthermore, it can have a subtype which places a more precise constraint. For instance, given that focus prominence in Spanish has a strong tendency to fall on the last prosodic word in the PHON list of a focused sign, (4) can be altered into (5) in Spanish (Ibid. p. 191).

(5) $$\begin{bmatrix} \textit{sign} & \\ \text{SYNSEM} \mid \text{CONT} & \begin{bmatrix} \textit{mrs} \\ \text{RELS } \boxed{1} \end{bmatrix} \\ \text{IS} \mid \text{FOC} & \left\langle \boxed{1} \right\rangle \end{bmatrix} \rightarrow \left[\text{PHON} \quad \textit{list} \oplus \left\langle \left[\text{UT} \mid \text{DTE} \quad +\right] \right\rangle\right]$$

One of the advantages that this formalism provides might be that the relation between focus and prosodic prominence is restricted in a fairly straightforward manner as shown in (4). In addition, it is a significant endeavor to the HPSG framework to look into how various phonological layers interact with each other in phases and end up with focus projection.

However, these AVMs are viewed differently with this dissertation. I argue that F-marking is relevant to marking information structure. In English, prosody has a relatively straightforward relationship to information structure marking. However, this does not hold necessarily true in other languages. Instead, I argue that F-marking needs to be represented as MKG|FC in the formalism of this dissertation. In other words, [MKG|FC +] indicates that the word (or the phrase) is F(ocus)-marked. As the name itself implies, F-marking is a matter of markedness, rather than a meaning. In

brief, F-marking which triggers the spreading of focus has to be specified as a feature of MKG under CAT. There several reasons for this argument, which are discussed in the following subsections.

13.2.1 Usage of MRS

First of all, the two AVMs (4) and (5) proposed by Bildhauer (2007) has an inconsistency with the DELPH-IN formalism that this dissertation relies on. In the DELPH-IN formalism of HPSG, we cannot search a specific element included in a list unless we create pointers into RELS (like ICONS-KEY in the present work).

13.2.2 Languages without Focus Prosody

Second, as presented in Chapter 4 (§4.2), some languages do not use prosody in expressing focus (e.g. Yucatec Maya (Kügler et al., 2007), Akan (Drubig, 2003), and Catalan (Engdahl and Vallduví, 1996)). Besides, in Hausa, prosodic prominence is disallowed for focus *in situ* (Hartmann and Zimmermann, 2007; Büring, 2010) (p. 98). If focus projection always happened by means of prosody, there could be no focus projection in these languages. Yet, it is my understanding that focus projection seems to be a universal phenomenon in human language (Büring, 2006).

13.2.3 Lexical Markers

Finally and most importantly, some languages make use of lexical markers to invoke focus projection. Some previous studies regard these lexical items as comment markers or scope markers. For instance, Koreans employs *man* 'only', and this lexical item contributes to extension of focus meaning, although a specific pattern of prosody may or may not occur when an element is focused (Choe, 2002). Similarly, *ba* in Abma (Schneider, 2009) and *shì* in Mandarin Chinese (von Prince, 2012) function to extend focus meaning into the larger constituents. Thus, the main component responsible for the spreading of focus meaning in this type of languages is not necessarily prosody.

13.3 Grammatical Relations

In the previous HPSG-based studies, ARG-ST (ARGument-STructure) or a linear arrangement of dependents of verbs plays a crucial role in identifying which phrases are projected from the F(ocus)-

marked word. Engdahl and Vallduví (1996) claims that focus projection can be licensed if and only if the most oblique argument of the phrase's head is F-marked. Their INFO-STRUCT instantiation principles (for English) are as follows.

(6) a. Either if a DAUGHTER's INFO-STRUCT is instantiated, then the mother inherits this instantiation (for narrow foci, links and tails),

b. or if the most oblique DAUGHTER's FOCUS is instantiated, then the FOCUS of the mother is the sign itself (wide focus). (Engdahl and Vallduví, 1996, p. 12)

De Kuthy (2000) provides a different argument with reference to the linear order of constituents; focus projection can happen if and only if the rightmost daughter is accented. Since the rightmost daughter is not always an oblique argument, De Kuthy's focus projection rules are not the same as the claim of Engdahl and Vallduví. The main point that Chung et al. (2003) claim is that ARG-ST is the locus where focus projection happens, which is largely in line with Engdahl and Vallduví, but there are some differences. They add two more factors into the formation of focus projection. One includes modification and coordination; the focus that a modifier bears cannot spread into its modificand and the larger phrase, and none of the operands in coordination (i.e. non-headed phrases) can project focus. The other is agentivity; if the focus value of the non-agentive lowest ranking argument is instantiated in its local position, then focus projection can take place. Bildhauer (2007) is another endeavor to show how focus projection can be dealt with within the HPSG formalism. Bildhauer points out the problems with previous studies: First, looking at obliqueness would sometimes be too rigorous or sometimes too loose to identify how the focus domain is built up. Second, the previous approaches are language-specific, and thereby may not be straightforwardly applied to other languages.

There are potentially (at least) six possibilities in spreading of focus meaning in English. For instance, a ditransitive sentence *Kim sent Lee a big book yesterday.* consists of six components as shown in (7).

(7)

Kim	sent	Lee			
(i) subject	(ii) verb	(iii) non-peripheral argument			
a	big		book		yesterday.
	(iv) NP modification		(v) peripheral argument		(vi) VP modification

First, focus associated with (i) subjects cannot be projected into the larger phrase (Chung et al., 2003). Although the subject in (7) bears the A-accent (i.e. KIM), the whole sentence cannot be in the focus domain. In other words, a Q/A pair (8Q2-A2) sounds infelicitous, whereas (8A1) sounds as an appropriate reply to (8Q1).

(8) Q1: Who sent Lee a big book yesterday?
A1: [$_f$KIM] sent Lee a big book yesterday.
Q2: What happened?
A2: #[$_f$KIM sent Lee a big book yesterday.].

This is in accordance with the proposal of Selkirk (1984, 1995). The subject is neither the head of the sentence nor an internal argument of the main verb. However, when the subject is an internal argument, the focus on subjects can be projected. The subjects of unaccusative verbs (e.g. *die*) have been analyzed as not an external argument of the verbs, but an internal argument. Chung et al. (2003) argue that whether the subject is an internal argument of the verb or not can explain an asymmetry in focus projection of subjects. Since unergative verbs, such as *ran* in (9b), take their subject as an external argument, the focus cannot be projected from the subject. In contrast, *Tom* in (9a) can play the core of focus projection, because the verb *died* takes it as an internal argument.

(9) a. [$_f$ TOM died].
b. #[$_f$ TOM ran]. (Chung et al., 2003, p. 395)

Second, it has been said that focus on (ii) verbs can be projected into the larger phrases (e.g. VP and S), but Gussenhoven (1999) argues that such a projection is incompatible with intuition. That is, the following Q/A pair does not sound natural to Gussenhoven. That is to say, the focus associated with SENT cannot be projected into the VP.

(10) Q: What did she do?
A: #She SENT a book to Mary.

Third, distinction between (iii) non-peripheral argument and (v) peripheral argument with respect to focus projection has been deeply investigated. Bresnan (1971) argues that focus projection in English happens if and only if the A-accented word is the peripheral argument.

(11) a. The butler [$_f$ offered the president some COFFEE].
b. *The butler [$_f$ offered the PRESIDENT some coffee].
c. The butler offered [$_f$ the PRESIDENT some coffee]. (Chung et al., 2003, p. 388)

Fourth, modifiers (i.e. (iv) and (vi)) are incapable of extending the focus that they are associated with to their head phrases. Thus, any head cannot inherit a focus value from its adjunct.

In the following section, I narrow down the scope of analysis to the distinction between (iii) non-peripheral argument and (v) peripheral argument, and leave a deeper analysis of the full range of focus projection to future work.

13.4 An Analysis

My analysis this section provides makes use of ICONS and MKG. They are used to place a restriction on possibility of focus projection and to represent the meaning of a sentence in which focus projection can happen into a single parse tree.

13.4.1 Basic Data

A set of allosentences (i.e. close paraphrases which share truth-conditions (Lambrecht, 1996)) is presented in (12), and the difference amongst them is where the A-accent (marked as SMALL CAPS) falls on. In other words, what is focused is different in the different allosetences.

(12) a. KIM sent Lee the book.
b. Kim SENT Lee the book.
c. Kim sent LEE the book.
d. Kim sent Lee the BOOK.

According to Bresnan (1971), amongst these allosetences, focus projection can happen only in (12d): only the most peripheral argument can be the starting point of focus projection. For example, if a *wh*-question requires an answer of *all-focus* ("an absence of the relevant presuppositions" (Lambrecht, 1996, p. 232)), only the sentence in which the most peripheral argument bears an focus-marking (e.g. the A-accent) sounds felicitous, as exemplified in (13).[1]

[1] I would rather say "focus-marked" rather than "accented", because F(ocus)-marking does not necessarily mean prosodic marking as discussed before.

(13) Q: What happened?
A1: #[$_f$ KIM sent Lee the book].
A2: #[$_f$ Kim SENT Lee the book].
A3: #[$_f$ Kim sent LEE the book].
A4: [$_f$ Kim sent Lee the BOOK].
A5: #Kim sent [$_f$ Lee the BOOK].

In addition, there are two more restrictions on occurrence of focus projection:[2] First, focus projection takes place only when the syntactic head dominates the focus-marked element. For instance, focus cannot be projected in the way presented in (13A5) in which the verb *sent* is not in the focus domain. Second, the focus-marked element should be included in the focus domain, and this is also supported by the corpus analysis (§7.1.1). For instance, the followings in which the focus-marked BOOK is out of the bracket are ill-formed.

(14) a. *[$_f$Kim] sent Lee the BOOK.
b. *[$_f$Kim sent] Lee the BOOK.
c. *[$_f$Kim sent Lee] the BOOK.
d. *Kim [$_f$ sent] Lee the BOOK.
e. *Kim [$_f$ sent Lee] the BOOK.
f. *Kim sent [$_f$ Lee] the BOOK.

13.4.2 Rules

This dissertation follows the idea Chung et al. (2003) propose that ARG-ST is the locus where focus projection takes place. That means that the main constraint on the range of spreading focus should be specified in the lexical structure of the verb (i.e. *sent* in (13A4)). I introduce extra lexical rules to manipulate the feature structure(s) under VAL for constraining such a possibility of focus projection. That is, each verbal entry has its own ARG-ST independent of focus marking, and one extra verbal node is introduced at the lexical level when constructing a parse tree. On the other hand, the lexical rules for calculating focus projection refer to F-marking specified as a value of MKG|FC of the dependents specified in the list of VAL|COMPS (and VAL|SUBJ).

I propose a ditransitive verbal entry *send* used in (13A4)) takes <NP(NOM), NP(ACC), NP(ACC)>

[2] Jae-Woong Choe, p.c.

(i.e. two elements in COMPS) as its ARG-ST.[3] The basic entry is conjugated into *sent* by inflectional rules, and the inflected element can be the daughter of the lexical rules that I employ for computing focus projection. There are two rules to look at the values in VAL|COMPS, as presented below.

(15) a.
$$\left[\begin{array}{l} \textit{no-focus-projection-rule} \\ \text{INDEX } \boxed{1} \\ \text{ICONS-KEY } \boxed{2} \\ \text{VAL} \left[\begin{array}{ll} \text{SUBJ} & \left\langle \left[\text{ICONS-KEY} \quad \textit{non-focus}\right] \right\rangle \\ \text{COMPS} & \left\langle \left[\text{MKG} \mid \text{FC} \quad +\right], \left[\begin{array}{ll} \text{MKG} \mid \text{FC} & - \\ \text{ICONS} & \langle\, !\ ! \,\rangle \end{array}\right] \right\rangle \end{array}\right] \\ \text{C-CONT} \mid \text{ICONS} \left\langle\, ! \, \boxed{2} \left[\begin{array}{l} \textit{non-focus} \\ \text{TARGET } \boxed{1} \end{array}\right] ! \,\right\rangle \\ \text{DTR } \textit{lex\_rule\_infl\_affixed} \end{array}\right]$$

b.
$$\left[\begin{array}{l} \textit{focus-projection-rule} \\ \text{CLAUSE-KEY } \boxed{1} \\ \text{VAL} \mid \text{COMPS} \left\langle \left[\begin{array}{ll} \text{MKG} \mid \text{FC} & - \\ \text{INDEX} & \boxed{2} \end{array}\right], \left[\begin{array}{ll} \text{MKG} \mid \text{FC} & + \\ \text{ICONS} & \left\langle\, ! \left[\textit{semantic-focus}\right] ! \,\right\rangle \end{array}\right] \right\rangle \\ \text{C-CONT} \mid \text{ICONS} \left\langle\, ! \left[\begin{array}{ll} \textit{non-focus} & \\ \text{TARGET} & \boxed{2} \\ \text{CLAUSE} & \boxed{1} \end{array}\right] ! \,\right\rangle \\ \text{DTR } \textit{lex\_rule\_infl\_affixed} \end{array}\right]$$

No-focus-projection-rule shown in (15a) takes a non-focus-marked element as its the last component, while *focus-projection-rule* shown in (15b) takes a focus-marked one. Focus projection in a sentence whose main verb stems from *send* can happen by using only the second one (i.e. *focus-projection-rule*), and the first one (i.e. *no-focus-projection-rule*) predicts other sentences in which the most peripheral argument (i.e. *the book* in this case) introduces no *info-str* value into ICONS. Note that *focus-projection-rule* requires one information structure value (specified as *semantic-focus*) of the last element in VAL|COMPS.

For example, (16a-b) are not compatible with each other. When *Lee* is A-accented (i.e. LEE with [FC +]), (15b) cannot take it as its complement. (15a) can take LEE as its complement, but (15a)

[3] In the ERG (English Resource Grammar, Flickinger 2000), a default form of *send* is divided into several different types, mainly depending on specification of ARG-ST, such as 'send_v1', 'send_v2', etc. I follow this strategy of enumerating verbal entries.

prevents the A-accented BOOK with [FC +] from being the second complement. In other words, *sent* in (16a) is constrained by (15a), while that in (16b) is constrained by (15b).

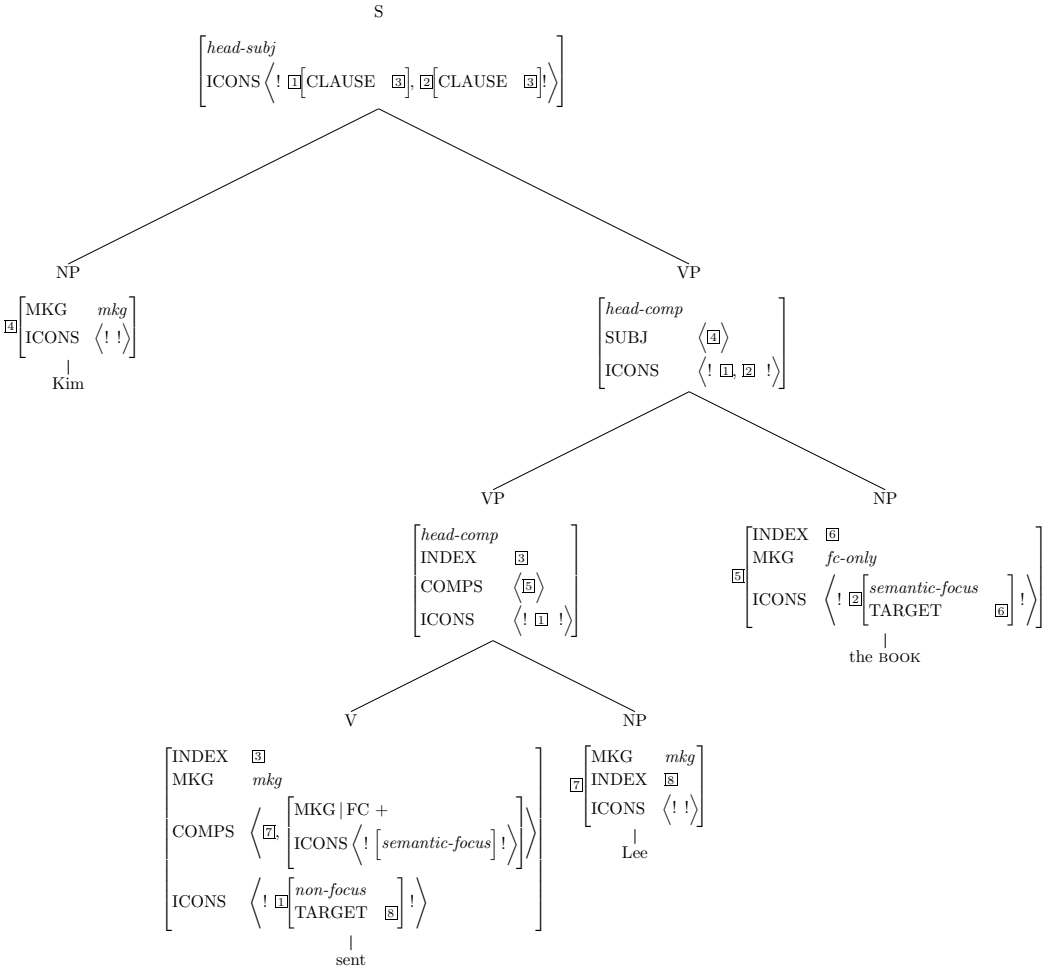
(16) a. Kim sent LEE the book.

b. Kim sent Lee the BOOK.

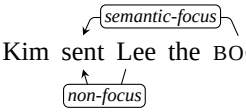
13.4.3 Representation

The main motivation of using ICONS with respect to focus projection is to provide only one single parse tree that covers all potential meanings of focus projection. The parse tree of (16b) is sketched out in (17).

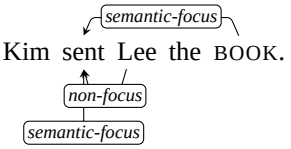
(17)

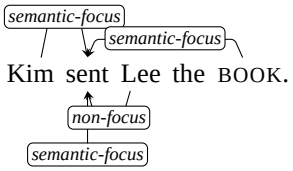


The corresponding dependency graph is provided in (18).

(18) 

In (18), there are four information structure relations. Two of them are visible in (18): One is *non-focus* between *Lee* (unmarked) and the semantic head *sent*, and the other is *semantic-focus* between BOOK (A-accented) and *sent*. In addition to them, there are two other potential relations, left underspecified in the dependency graph. One is between *Kim* and *sent*, and the other is *sent* to itself. These can be monotonically specified in further processing. That is, further constraints can be added, but only if they are consistent with what is there. This underspecified ICONS representation get further specified to VP focus or S focus. According to the graph in (18), *Lee* should not be focused, *book* should be focused, and *Kim* and *sent* may or may not be focused. When *sent* is focused, the ICONS list in the output includes three ICONS elements (i.e. VP focus). When both *sent* and *sent* are focused, the ICONS list in the output includes four ICONS elements (i.e. S focus). When they are associated with focus, the representations are sketched out in (19a-b), respectively. Note that the input representation provided in (18) subsumes (19a-b), but not *vice versa*.

(19) a. 

b. 

Especially, this representation works in generation. First, the first element in the final ICONS list given in (17) assigns *non-focus* to *Lee*, an A-accented LEE is ruled out in the generation output. Second, the second element in the final ICONS list assigns *semantic-focus* to *the book*, *the book* must be focus-marked in the generation output (i.e. *the* BOOK). Third, *Kim* and *sent* in the underspecified relations can be associated with *semantic-focus*, and thereby '-a' can be attached to them. Consequently, the following three outputs can be generated when using *Kim sent Lee the book-a.* as the input string. (20a-c) hypothetically represent NP focus, VP focus, and S focus, respectively.[4]

(20) a. Kim sent Lee the book-a.
b. Kim sent-a Lee the book-a.
c. Kim-a sent-a Lee the book-a.

[4]For ease of comparison, the other hypothetical suffix '-b' is not considered here.

Lastly, it is noteworthy that a sentence built with (15a), such as (16a), cannot be further specified in the same way. (15a) introduces an element whose value is *non-focus* into the ICONS list. Since this constraint prevents the verb *sent* from being focused, neither VP focus nor S focus can happen in the sentence. Additionally, since the subject is constrained as [ICONS-KEY *non-focus*] in (15a), an A-accented KIM cannot be the subject. In an actual processing, (21a) cannot be paraphrased as sentences in (21b-d).

(21) a. Kim sent Lee-a the book.
b. Kim sent-a Lee-a the book.
c. Kim-a sent-a Lee-a the book.
d. Kim-a sent Lee-a the book.

13.4.4 Further Question

My analysis presented thus far leaves an interesting question for future work. The EP that *sent* introduces into the RELS list is represented as (22). In the sentences given in (16), the INDEXes of *Kim*, *Lee*, and *book* have coreferences with ARG1, ARG2, and ARG3, respectively.

(22)
$$\begin{bmatrix} \text{RELS} \left\langle \begin{bmatrix} \textit{\_send\_v\_rel} \\ \text{ARG0}\ \boxed{1} \\ \text{ARG1}\ \boxed{2} \\ \text{ARG2}\ \boxed{3} \\ \text{ARG3}\ \boxed{4} \end{bmatrix} \right\rangle \end{bmatrix}$$

In accordance with (22), the ICONS lists for (16a-b) are constructed as (23a-b). They explicitly specify the information structure values on ARG2 and ARG3, but ARG1 is not included in them.

(23) a. Kim sent LEE the book.
$$\begin{bmatrix} \text{ICONS} \left\langle \begin{bmatrix} \textit{semantic-focus} \\ \text{TARGET}\ \boxed{3} \\ \text{CLAUSE}\ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

b. Kim sent Lee the BOOK.
$$\begin{bmatrix} \text{ICONS} \left\langle \begin{bmatrix} \textit{non-focus} \\ \text{TARGET}\ \boxed{3} \\ \text{CLAUSE}\ \boxed{1} \end{bmatrix}, \begin{bmatrix} \textit{semantic-focus} \\ \text{TARGET}\ \boxed{4} \\ \text{CLAUSE}\ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix}$$

The current assumption is that focus cannot spread to the role associated with the subject (here ARG1) without including the verb. For instance, (23b) cannot be interpreted in the same way as (24) via focus projection, because the ICONS in (24) does not contain an element for the verb. Note that the first ICONS element in (24) is introduced by the A-accent rule for KIM.

(24) KIM sent Lee the BOOK.

$$\left[\text{ICONS}\left\langle\begin{bmatrix}\textit{semantic-focus}\\ \text{TARGET } \boxed{2}\\ \text{CLAUSE } \boxed{1}\end{bmatrix},\begin{bmatrix}\textit{non-focus}\\ \text{TARGET } \boxed{3}\\ \text{CLAUSE } \boxed{1}\end{bmatrix},\begin{bmatrix}\textit{semantic-focus}\\ \text{TARGET } \boxed{4}\\ \text{CLAUSE } \boxed{1}\end{bmatrix}\right\rangle\right]$$

This assumption seems linguistically true in that the sentence in (24) is not an instance of S focus. The question is which mechanism technically blocks such specialization. This mechanism for focus projection has to play two functions. First, it allows the subject (here ARG1) to be associated with focus if and only if the verb is also associated with focus. Second, it serves to prevents the ICONS list in (23a) from being further specified. My further research will delve into how the mechanism works.

13.5 Summary

This chapter offered a new approach of computing focus projection in terms of sentence generation. First, the present study argues that a single parse tree of a sentence with focus projection is enough to represent the meaning of information structure and also more effective in the context of grammar engineering. Second, F-marking is not necessarily encoded by prosody. In some languages (e.g. Mandarin Chinese and Korean), some lexical markers play a role to extend the domain of focus. Thus, F-marking in this dissertation is dealt with [MKG|FC +]. Third, (at least in English), focus projection happens normally when the most peripheral item is focus-marked though there are some exceptional cases. Fourth, there are two more constraints on focus projection. One is that the the focus-marked element should be included in the focus domain. The other is that focus-marked element should be headed. In other words, the focus meaning cannot be extended to any non-head phrases. Building upon these arguments, the last section in this chapter shows how a simple ditransitive sentence is analyzed with respect to focus projection. Two lexical rules are introduced to discriminate a sentence in which focus projection happens. This is a piece of evidence to support my argument in this chapter, but a more thorough study is required in future research.

Part V

IMPLEMENTATION

Building upon the linguistic findings discussed thus far and the formalism using ICONS (Individual CONStraint), this part implements the library of information structure for the LinGO Grammar Matrix system and substantiates that the ICONS-based representation can be used to improve transfer-based machine translation. Chapter 14 dwells on the details of incorporating the information structure library into the LinGO Grammar Matrix customization system. Chapter 15 explains how multilingual machine translation works in conjunction with ICONS, and conducts a small experiment for testing multilingual machine translation.

Chapter 14

CUSTOMIZING INFORMATION STRUCTURE

Grammar customization with the LinGO Grammar Matrix (Bender and Flickinger, 2005; Drellishak, 2009; Bender et al., 2010b) takes as input the typological and lexical information that a user provides about a language. This information is provided via a web-based questionnaire which has subpages for a series of language phenomena. For each phenomenon, the questionnaire gives a basic explanation and questions designed to help the user describe an analysis of the phenomenon (`http://www.delph-in.net/matrix/customize`). After the questionnaire has been answered, the user can press a button to customize a grammar.[1] This button invokes the customization script, which takes the user's answers stored in a `choices` file, and first validates them for consistency, then articulates grammar fragments into a complete grammar for the user's language. The output is an HPSG/MRS-based grammar built automatically on the basis of specifications the user has given. If the automatic construction is successful, a compressed file (zip or tar.gz) is made available for download. The downloadable file includes all required components for HPSG/MRS-based grammar engineering within the DELPH-IN formalism, so once uncompressed, the user can try out the grammar with processors such as LKB (Copestake 2002), PET (Callmeier 2000), and other DELPH-IN software such as [incr tsdb()] (Oepen 2001).

The grammatical categories covered in the current version are listed in (1). The pages sometimes work independently, and sometimes co-operate with choices given in other subpages. For example, users can add some additional features when there is a need (e.g. animacy) on the "Other Features" page, which will then appear as an option of syntactic or semantic features in other subpages such as "Lexicon" and "Morphology". To take another example, the "Sentential Negation" page elicits information about morphosyntactic strategies of negation in the user's language, and specific forms

[1] The users can check out the feasibility of their choices on the questionnaire beforehand by using "Test by Generation", which performs the customization in the background and displays sentences realized using the grammar for generation with predefined semantic templates. The users can then refine their choices based on the quality of the results. For more information, see Bender et al. (2010a).

of negation operators can be inserted in "Lexicon" and/or "Morphology" (Crowgey, 2012). The "Information Structure" page works in a similar way.

(1) a. Word Order (Fokkens, 2010)
b. Number (Drellishak, 2009)
c. Person (Drellishak, 2009)
d. Gender (Drellishak, 2009)
e. Case (Drellishak, 2009)
f. Direct-inverse (Drellishak, 2009)
g. Tense, Aspect and Mood (Poulson, 2011)
h. Other Features (Drellishak, 2009; Poulson, 2011)
i. Sentential Negation (Crowgey, 2012)
j. Coordination (Drellishak and Bender, 2005)
k. Matrix Yes/No Questions (Bender and Flickinger, 2005)
l. Information Structure (this dissertation)
m. Argument Optionality (Saleem, 2010; Saleem and Bender, 2010)
n. Lexicon (Drellishak, 2009)
o. Morphology (Goodman, 2013)

Four more pages not directly related to grammar creation but necessary for ease of development are presented in (2) as well. In the "General Information" page, users input supplementary information, such as the ISO 639-3 code of the language, delimiters in the languages, etc. The "Import Toolbox Lexicon" page provides an interface to the Field Linguist's Toolbox[2] (`http://www-01.sil.org/computing/toolbox`), which is provided by SIL[3] (`http://www.sil.org`). Users can input test sentences in the "Test Sentences" page, which are included with the customized grammar for basic evaluation of the grammar's parsing coverage. The last one provides several options for fine-tuning the results of "Test by Generation".

[2]This page facilitates using the customization system as a documenting tool for endangered languages.

[3]This institution (a.k.a. Summer Institute of Linguistics) studies less commonly taught languages and/or endangered languages, and thereby has published linguistic literature on the languages. They provide an online collection of ethnologic data (`http://www.ethnologue.com`), which is importantly used for the typological studies of human language.

Figure 14.1: Screenshot of the questionnaire (main page)

(2) a. General Information
b. Import Toolbox Lexicon
c. Test Sentences
d. Test by Generation Options

The screenshot of the current version's main page is shown in Figure 14.1.

14.1 The Questionnaire

The first task of implementing the customization system's information structure library centers around adding an HTML-based page to the web-based questionnaire. The "Information Structure" page is comprised of four sections: namely "Focus", "Topic", "Contrastive Focus", and "Contrastive Topic". Each section, except the last one, consists of two subparts: one for syntactic positioning and

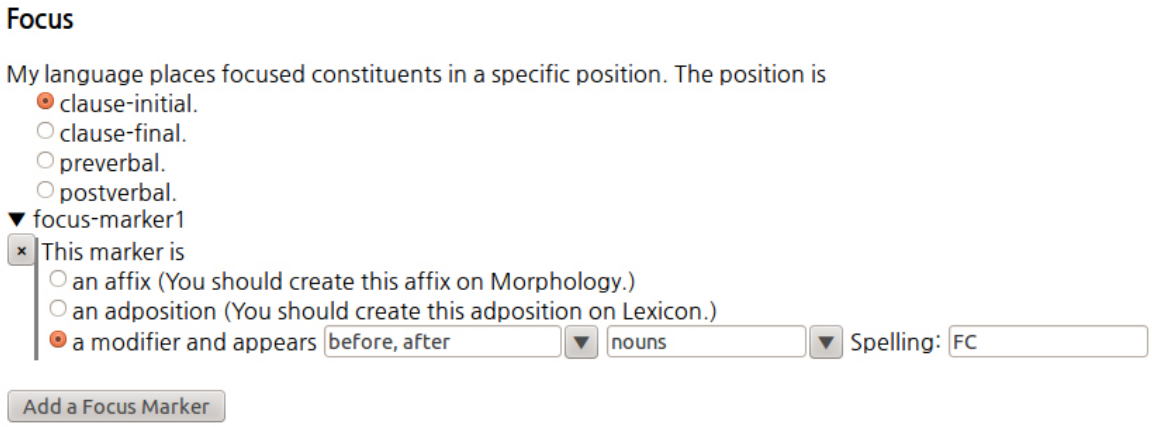


Figure 14.2: Screenshot of editing focus position/markers in the questionnaire

the other for lexical marking(s).

14.1.1 Focus

First, in the "Focus" section, users can specify the canonical position of focus in the user's language. A screenshot is shown in Figure 14.2. According to the cross-linguistic survey given in Chapter 4 (p. 75), there are four options: clause-initial, clause-final, preverbal, and postverbal. A sentence in the neutral word order is a default form in the language, which can be interpreted as conveying a range of information structure. For example, if a language basically employs SVO, [O] in SVO does not involve any specific *info-str* value even though the canonical position of focus is clause-final or postverbal in the language.

Users can add one or more focus markers. The type of a focus marker is either an affix, an adposition, or a modifier as surveyed in §4.3 (p. 67). Affixes are treated in the morphological paradigm (i.e. `irules.tdl`), while the other two are treated like a word (i.e. `lexicon.tdl`). The distinction between the last two is also discussed in §4.3: If a language employs case-marking adpositions and a lexical marker to express focus and/or topic is in complementary distribution with the case-marking adpositions, the marker is categorized as an adposition in principle. Otherwise, the marker is treated as just a modifier.[4] Specific forms for information-structure marking affixes

[4]Nonetheless, the users of the LinGO Grammar Matrix system may have the flexibility to describe what they see in

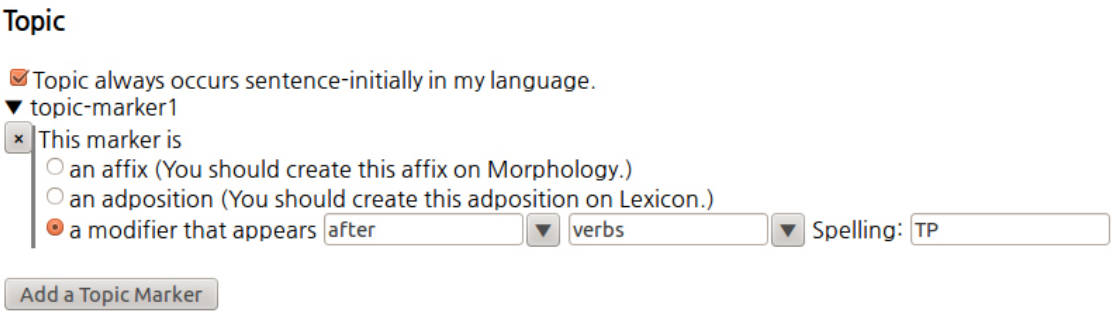


Figure 14.3: Screenshot of editing topic position/markers in the questionnaire

and adpositions are not specified in the "Information Structure" page, and instead they should be defined in the "Morphology" and "Lexicon" pages, respectively. If users select that their language has an affix or an adposition of expressing focus in the "Information Structure" page, but an affix or an adposition that involves *focus* or super/sub-types of *focus* as a value of "information-structural meaning" is not added in the "Morphology" or "Lexicon" pages, a validation error is produced. The spelling of information-structure marking modifier(s) is directly specified on the "Information Structure" page, because there is no room for such an expression (e.g. particles, clitics, etc.) in "Morphology" and "Lexicon". Users can specify more constraints on information-structure marking modifier(s), such as before and/or after nouns and/or verbs.

For instance, Figure 14.2 is illustrative of users' choices on "Focus". As mentioned earlier in §4.3.3 (p. 71), one lexical marker may be used to signal focus to both nominal and verbal items. One lexical marker may occur sometimes before focused constituents and sometimes after them. Thus, the users can take multiple options for the constraints, as presented as [before, after] in Figure 14.2.

14.1.2 Topic

Second, the "Topic" section has two choices for constraints. As for the constraint on positioning, an option for the topic-first restriction is provided for the languages in which topic always occupies the sentence-initial position. Next, one or more topic markers can be added, which operate in the same way as the "Add a Focus Marker" button discussed above. As shown in §3.3.3 (p. 46), verbal items

their language following Poulson (2011)'s meta-modeling idea.

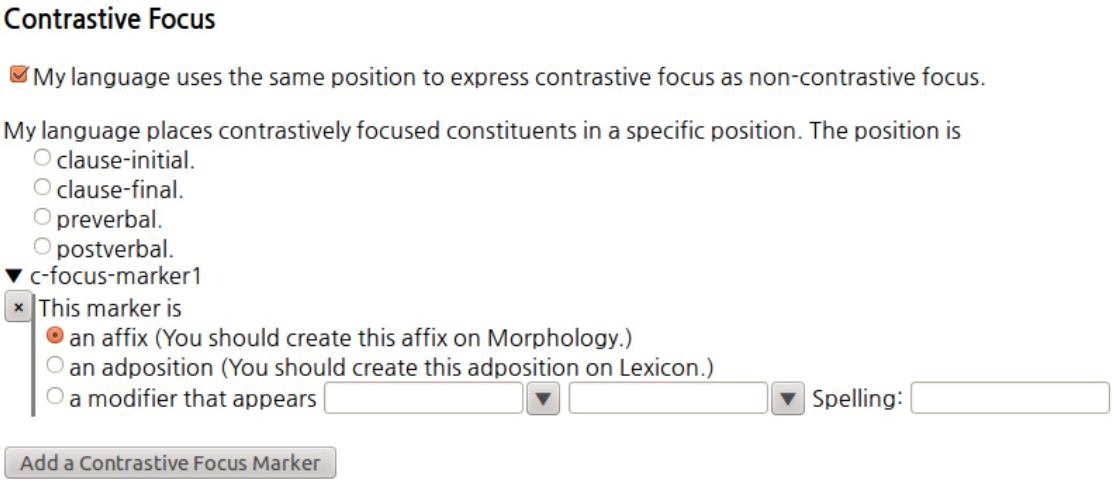


Figure 14.4: Screenshot of editing contrastive focus position/markers in the questionnaire

can be topicalized in some languages (e.g. Paumarí (Chapman, 1981)). Thus, [verbs] in Figure 14.3 is selected for illustrating the language-specific constraint.

14.1.3 Contrastive Focus

Third, contrastive focus may or may not be marked differently from non-contrastive focus, which is language-specific. If the first checkbox in Figure 14.4 (just under the title "Contrastive Focus") is not selected, there can be two types of foci: one is *semantic-focus* for non-contrastive focus, and the other is *contrast-focus* for contrastive focus. In the latter case, users have to choose a specific position for contrastive focus, such as clause-initial, clause-final, preverbal, or postverbal. If users do not choose one of them, the validation script gives an error message. Likewise, users can add a contrastive focus marker in the same manner.

14.1.4 Contrastive Topic

Finally, there is an option for "Contrastive Topic". Throughout my cross-linguistic survey (explained in §4.1), I found no language in which contrastive topics have a constraint on positioning.[5]

[5]In fact, previous literature also provides the same analysis (Choi, 1999; Erteschik-Shir, 2007; Bianchi and Frascarelli, 2010).

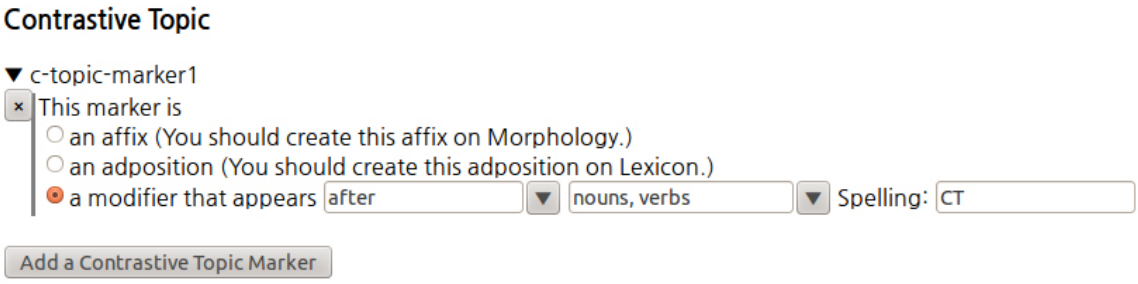


Figure 14.5: Screenshot of editing contrastive topic markers in the questionnaire

Accordingly, there is no checkbox for the positioning constraint. On the other hand, some languages employ a contrastive topic marker (e.g. *thì* in Vietnamese (Nguyen, 2006)). These can also be specified using the button "Add a Contrastive Topic Marker".

14.2 The Matrix Core

The next task was to incorporate the analysis based on ICONS (Individual CONStraints) into the Matrix core. The core TDL fragments written in `matrix.tdl` define universally useful types in widespread linguistic phenomena. Notably, integrating ICONS into the grammar requires touching lots of previously implemented types as well as adding several new types. This is because I am concerned not merely with the representations (implemented via changes to MRS and the addition of the actual type for ICONS), but also with their composition at the syntactic level. Thus, I had to revise many lexical rules and types inherited by almost all phrase structure rules and lexical rules. The details are as follows, and the TDL statements that I modified and created are attached in Appendix D.

14.2.1 Fundamentals

First of all, three type hierarchies presented in Chapter 9, such as *info-str*, *mkg*, and *sform*, were added. Then, [MKG *mkg*] was added into CAT, and CONT values were also edited as containing ICONS-related features, such as [ICONS-KEY *icons*] and [CLAUSE-KEY *event*] under *hook*, and [ICONS *diff-list*] under *mrs*. The TDL statement for representing *info-str* is presented in (3).

(3)
```
icons := avm.
info-str := icons &
  [ CLAUSE individual,
    TARGET individual ].
non-topic := info-str.
contrast-or-focus := info-str.
focus-or-topic := info-str.
contrast-or-topic := info-str.
non-focus := info-str.
focus := non-topic & contrast-or-focus & focus-or-topic.
contrast := focus-or-topic & contrast-or-focus & contrast-or-topic.
topic := non-focus & focus-or-topic & contrast-or-topic.
bg := non-topic & non-focus.
semantic-focus := focus.
contrast-focus := contrast & focus.
contrast-topic := contrast & topic.
aboutness-topic := topic.
```

ICONS were added into the basic lexical and phrasal types in `matrix.tdl` (e.g. *unary-phrase*, *binary-phrase*, *ternary-phrase*, etc.).[6] Next, I specifically inserted [C-CONT|ICONS <! !>] into phrase structure rules and lexical rules: when a lexical or phrasal type is not related to informaion-structure marking, C-CONT|ICONS is specified as an empty list.

14.2.2 Lexical Types

Regarding lexical types, the set of *icons-lex-item* presented in §10.1 (p. 196), such as *no-icons-lex-item*, *basic-icons-lex-item*, *one-icons-lex-item*, and *two-icons-lex-item*, were written as TDL statements. Lexical types for constraining ARG-ST (ARGument-STructure) inherit from one of them and impose some additional constraints on CLAUSE-KEY. For example, *intransitive-lex-item* which places a constraint on ARG-ST of intransitive verbs is defined as in (4b). Note that this type inherits from *basic-icons-lex-item* that has an empty ICONS list as shown in (4a). There is a coreference tag `#clause` in (4b), which indicates that every argument shares the value of CLAUSE-KEY with the semantic head within a single clause.

---

[6] Note that ICONS is a *diff-list* of *info-str*.

(4) a.
```
basic-icons-lex-item := lex-item &
  [ SYNSEM.LOCAL.CONT.ICONS <! !> ] ].
```
b.
```
intransitive-lex-item := basic-one-arg-no-hcons & basic-icons-lex-item &
  [ ARG-ST < [ LOCAL.CONT.HOOK [ INDEX ref-ind & #ind,
                                 ICONS-KEY.CLAUSE #clause ] ] >,
    SYNSEM [ LKEYS.KEYREL.ARG1 #ind,
             LOCAL.CONT.HOOK.CLAUSE-KEY #clause ] ].
```

Some lexical types inherently include an *info-str* element in the ICONS list. In this case, lexical types for ARG-ST impose a constraint on the element of *info-str*. For instance, according to my proposal of information structure in complement clauses (discussed in §11.1 (p. 230)), *clausal-second-arg-trans-lex-item* responsible for the ARG-ST of verb classes taking a clausal complement (e.g. *think*, *ask*, etc.) is defined as (5b). Note that the INDEX of the second argument (i.e. a clausal complement) and the TARGET of the element in the ICONS list are co-indexed (`#target`).

(5) a.
```
one-icons-lex-item := lex-item &
  [ SYNSEM.LOCAL.CONT.ICONS <! [ ] !> ] ].
```
b.
```
clausal-second-arg-trans-lex-item := basic-two-arg & one-icons-lex-item &
  [ ARG-ST < [ LOCAL.CONT.HOOK [ INDEX ref-ind & #ind,
                                 ICONS-KEY.CLAUSE #clause ] ],
             [ LOCAL.CONT.HOOK [ LTOP #larg, INDEX #target ] ] >,
    SYNSEM [ LOCAL.CONT [ HOOK.CLAUSE-KEY #clause,
                          HCONS <! qeq & [ HARG #harg, LARG #larg ] !>,
                          ICONS <! [ CLAUSE #clause, TARGET #target ] !> ],
             LKEYS.KEYREL [ ARG1 #ind, ARG2 #harg ] ] ].
```

14.2.3 Lexical Rules

There are two types of lexical rules. One introduces an element of *info-str* into C-CONT|ICONS, and the other does not. In order to support these types, I needed to change *no-ccont-rule* to include [C-CONT|ICONS <! !>] and to add a type that allows for a non-empty ICONS list. This type is called *no-rels-hcons-rule*, and constrains RELS and HCONS to be empty while leaving ICONS underspecified. These rules are also used for phrase structure rules as well.

14.2.4 Phrase Structure Rules

First, *basic-head-subj-phrase*, *basic-head-comp-phrase*, *basic-head-spec-phrase*, and *basic-bare-np-phrase* inherit from *no-ccont-rule*. Therefore, they have an empty ICONS list. Second, I edited two phrase structure rules for argument optionality: namely *basic-head-opt-subj-phrase* and *basic-*

head-opt-comp-phrase. They introduce an ICONS element that indicates the value of information structure the dropped argument has (i.e. *non-focus*) into C-CONT|ICONS. This constraint is in line with my analysis presented in §12.2 (p. 252). Third, I touched *basic-head-mod-phrase-simple* (a subtype of *head-mod-phrase*) in accordance with the AVM presented in §10.2 (p. 210): now it has an empty list in C-CONT|ICONS, and the CLAUSE-KEY of modifiers (NON-HEAD-DTR) and that of their modificands are co-indexed with each other. Finally, *head-filler-phrase* does not include [C-CONT|ICONS <! !>]. This is because its subtypes sometimes constructionally introduce an element of *info-str*. For example, clause-initial and clause-final focus in languages with a fixed word order are instances of *head-filler-phrase*, and introduce an element into C-CONT|ICONS. The remaining part of a sentence which has a syntactic gap is constrained by *basic-head-subj-nmc-phrase* or *basic-head-comp-nmc-phrase*, in which *nmc* stands for non-matrix-clause. These rules work for *head-subj-phrase* and *head-comp-phrase* that cannot be root nodes by themselves (i.e. specified as [MC –]). These phrases are supposed to be combined only with a *filler-phrase*. There is one more phrase structure rule related to *filler-phrase*: namely *nc-filler-phrase*. This rule handles a non-canonical *filler-phrase*; for example, detached constituents in right dislocation. The following section (§14.3.2) addresses the phrase structure rules for constraining such a constituent dislocated from the main clauses.

14.3 Customized Grammar Creation

The third task is to implement the Python code to customize the users' choices. The code first validates the content in the `choices` file to check whether an inconsistency happens and whether there is a missing input. If no error occurs, then the code converts the content in the `choices` file into TDL statements.

(6)
```
section=info-str
focus-pos=clause-final
  focus-marker1_type=modifier
  focus-marker1_pos=after
  focus-marker1_cat=nouns, verbs
  focus-marker1_orth=FC
topic-first=on
c-focus-pos=preverbal
c-topic-marker1_type=affix
```

The users' answers about information structure are stored in a `choices` file as exemplified as

exemplified in (6). (6) specifies that the language places a focused constituent in the clause-final position, and employs a focus marker, which is a single word spelled as '`FC`' appearing after nouns or verbs. The language is a topic-first language as indicated by '`topic-first=on`'. The language uses a different place for signaling contrastive focus. In this case, it is the preverbal position. Finally, the language has an affix responsible for conveying a meaning of contrastive topic, which should be defined in the "Morphology" page. Those choices are transmitted into the customization script for information structure.[7] This script written in Python creates information-structure related TDL statements, working out with what I have proposed hitherto in the previous chapters.

14.3.1 Lexical Markers

There are three types of lexical markers: (i) affixes, (ii) adpositions, and (iii) modifiers. Amongst them, the first one and the second one are specified in the "Morphology" and "Lexicon" pages respectively. They are handled by existing customization code, which works seamlessly with the information-structure related features and values enabled by the information structure library. I touched two existing customization libraries for the first two options, and the script for information structure (i.e. `information_structure.py`) creates only the last type of markers.

(i) Affixes are customized by `morphotactics.py`. If a lexical rule imposes a constraint on information structure meaning, the lexical rule inherits from *no-rels-hcons-rule* (explained before in (§14.2.3) and introduces an element of *info-str* into C-CONT|ICONS. Otherwise, it inherits just from *add-only-no-ccont-rule* (or other lexical rules with an empty C-CONT|ICONS). For instance, the TDL fragment presented in (7) is responsible for a focus-marking suffix, and introduces a value of *info-str* into ICONS. Note the two coreference tags in *add-icons-rule*: namely #icons and #target.

(7)
```
add-icons-rule := phrase-or-lexrule & word-or-lexrule &
  [ SYNSEM.LOCAL.CONT.HOOK [ INDEX #target,
                             ICONS-KEY #icons ],
    C-CONT.ICONS <! info-str & #icons & [ TARGET #target] !> ].

p1-lex-rule-super := add-only-no-rels-hcons-rule & infl-lex-rule &
  [ DTR noun-lex ].

r1-lex-rule := add-icons-rule & p1-lex-rule-super &
  [ SYNSEM.LOCAL [ CAT.MKG fc,
                   CONT.HOOK.ICONS-KEY focus ] ].
```

[7] `gmcs/linglib/information_structure.py`

(ii) Adpositions are dealt with by `lexical_items.py`. Likewise, an ICONS list of an adposition is constituted depending on whether an adposition has a feature that constrains the semantics related to information structure. An instance is provided in (8). In this case, the adposition lexically includes a value of *info-str* in CONT|ICONS, and the TARGET is co-indexed with the INDEX of the complement. This works in the same manner as *ga* and *wa* in Japanese as presented previously in §10.4.2 (p. 221).

(8)
```
infostr-marking-adp-lex := basic-one-arg & raise-sem-lex-item & one-icons-lex-item &
  [ SYNSEM.LOCAL [ CAT [ HEAD adp & [ MOD < > ],
                         VAL [ SPR < >,
                               SUBJ < >,
                               COMPS < #comps &
                                       [ LOCAL.CONT.HOOK.INDEX #target ] >,
                               SPEC < > ] ],
                   CONT [ HOOK.ICONS-KEY #icons,
                          ICONS <! #icons & [ TARGET #target ] !> ] ],
    ARG-ST < #comps &
             [ LOCAL.CAT [ HEAD noun,
                           VAL.SPR < > ] ] > ].
```

(iii) Finally, `information_structure.py` makes TDL statements for information-structure marking modifiers, and depending on the specific choices, the lexical types are also elaborated. For example, the choices in (6) invoke the following TDL statements given in (9). TDL statements presented in (9) define the lexical type of modifiers that mark information structure. Like information-structure marking adpositions shown above, a value of *info-str* is lexically included in CONT|ICONS, but the TARGET is co-indexed with the INDEX of its modificand.

(9) a.
```
infostr-marking-mod-lex := no-rels-hcons-lex-item & one-icons-lex-item &
  [ SYNSEM.LOCAL [ CAT [ HEAD adv &
                               [ MOD < [ LIGHT -,
                                         LOCAL.CONT.HOOK [ INDEX #target,
                                                           ICONS-KEY #icons ] ] > ],
                         VAL [ SUBJ < >, COMPS < >, SPR < >, SPEC < > ] ],
                 CONT.ICONS <! #icons & [ TARGET #target ] !> ] ].
```

b.
```
focus-marking-mod-lex := infostr-marking-mod-lex &
  [ SYNSEM.LOCAL.CAT [ MKG fc,
                       HEAD.MOD < [ L-PERIPH luk,
                                    LOCAL [ CAT.HEAD noun,
                                            CONT.HOOK.ICONS-KEY focus ] ] > ] ].
```

Since a modifier and its modificand are combined with each other by a phrase structure rule, the customization script additionally creates some TDL statements related to *head-mod-phrase*. For

example, if the language employs an information-structure marking modifier and the modifier appears after its modificand, *head-adj-int-phrase* (a subtype of *basic-head-mod-phrase-simple*) and *head-adj-int* are inserted into `mylang.tdl` and `rules.tdl`, respectively. Additionally, an entry of the information-structure marking modifier is specified in `lexicon.tdl`.

14.3.2 Syntactic Positioning

The customization script `information_structure.py` also creates grammatical fragments in TDL for constraining focus or topic in a specific position. An an initial step, the script merges the users' choices into a single type. For example, if a language places focused constituents in the clause-initial position and the language has the topic-first restriction, clause-initial constituents *ex situ* are specified as *focus-or-topic* in the language.

As mentioned in §14.2.4, languages with a fixed word order (e.g. SVO, SOV, VSO, VOS, OSV, and OVS) employ a specific type of *head-filler-phrase* for clause-initial and clause-final focus and clause-initial topic. In other words, the focused and topicalized constituents fill out the syntactic gap of the remaining part of a sentence. The remaining part of sentence is realized as non-main-clausal constituents (e.g. *head-nmc-subj-phrase* and *head-nmc-comp-phrase*), which (i) have a nonempty list in NON-LOCAL|SLASH, and flag features indicating (ii) the phrase cannot be a main clause (i.e. [MC –]), and (iii) the phrase is not peripheral (i.e. [L-PERIPH –, R-PERIPH –]). Such a phrasal type with the *nmc* prefix should be combined with phrases with [L-PERIPH +] or [R-PERIPH +] to constitute a *infostr-filler-head-phrase*. The assignment of an *info-str* value is carried out by *infostr-dislocated-phrase* presented in (10a). The gap is filled in by *infostr-filler-head-phrase* presented in (10a).[8] Since this type specifies [L-PERIPH –] on itself, no further combination to the left side is allowed.

[8] In the case of right dislocation, *infostr-head-filler-phrase* which inherits from *nc-filler-phrase* instead of *basic-head-filler-phrase* is used.

(10) a.

```
infostr-dislocated-phrase := no-rels-hcons-rule & narrow-focus &
  [ SYNSEM.LOCAL.CAT.MC +,
    C-CONT.ICONS <! info-str & #icons &
                    [ TARGET #index, CLAUSE #clause ] !>,
    HEAD-DTR.SYNSEM.LOCAL [ CAT [ MC -,
                                  HEAD verb ],
                            CONT.HOOK [ INDEX #clause,
                                        CLAUSE-KEY #clause ] ],
    NON-HEAD-DTR.SYNSEM [ LIGHT -,
                          LOCAL [ CAT.HEAD +np,
                                  CONT.HOOK [ INDEX #index,
                                              ICONS-KEY #icons ] ] ] ].
```

b.

```
infostr-filler-head-phrase := basic-head-filler-phrase &
                              infostr-dislocated-phrase & head-final &
  [ SYNSEM.L-PERIPH +,
    HEAD-DTR.SYNSEM [ L-PERIPH -, LOCAL.CAT.VAL.SUBJ < > ],
    NON-HEAD-DTR.SYNSEM.LOCAL.CONT.HOOK.ICONS-KEY semantic-focus ].
```

If the user's language employs a fixed word order, preverbal and postverbal focus is constrained not by *head-filler-phrase*, but by specific types of *head-subj-phrase* and *head-comp-phrase*. Since preverbal/postverbal foci are immediately adjoined to the verb or the verb cluster,[9] they do not behave as a syntactic filler. Such a specific phrasal type imposes [LIGHT +] on the HEAD-DTR and [ICONS-KEY *focus*] (or a subtype of *focus*) on the NON-HEAD-DTR. What is significant here is using a flag feature: namely INFOSTR-FLAG. This feature indicates whether a constituent can be used as the preverbal and postverbal focus. *Narrow-focused-phrase* presented in (11a) is a unary phrase structure rule that specifies the plus value of INFOSTR-FLAG and introduces an element into ICONS. Only constituents with [INFOSTR-FLAG +] can be narrowly focused as constrained by *head-nf-comp-phrase-super* given in (11b) (or *head-nf-subj-phrase-super*). The specific value of *info-str* (e.g. *focus*) is assigned by *nf-comp-head-phrase* (or its siblings) presented in (11c).

[9] See the Basque example presented in §10.3.2 (p. 214), in which the subject is combined with a verb plus an auxiliary.

(11) a.

```
narrow-focused-phrase := head-only & no-rels-hcons-rule &
  [ C-CONT [ HOOK #hook,
             ICONS <! focus-or-topic & #icons &
                      [ TARGET #target ] !> ],
    SYNSEM [ LIGHT -,
             INFOSTR-FLAG +,
             LOCAL [ CAT.VAL [ SPR < >, SUBJ < >, COMPS < >, SPEC < > ],
                     CONT.HOOK [ INDEX #target,
                                 ICONS-KEY #icons ] ] ],
    HEAD-DTR.SYNSEM [ LIGHT -,
                      INFOSTR-FLAG -,
                      LOCAL [ CAT.HEAD noun,
                              CONT [ HOOK #hook,
                                     ICONS <! !> ] ] ] ].
```

b.

```
head-nf-comp-phrase-super := basic-head-comp-phrase & narrow-focus &
  [ SYNSEM.LOCAL.CAT [ MC -, VAL.COMPS #comps ],
    HEAD-DTR.SYNSEM.LOCAL.CAT.VAL.COMPS < #synsem . #comps >,
    NON-HEAD-DTR.SYNSEM #synsem & [ INFOSTR-FLAG + ] ].
```

c.

```
nf-comp-head-phrase := head-nf-comp-phrase-super & head-final &
 [ SYNSEM.LOCAL.CAT.MC -,
   HEAD-DTR.SYNSEM [ LIGHT +,
                     LOCAL.CAT.MC - ],
   NON-HEAD-DTR.SYNSEM.LOCAL [ CAT.HEAD +np,
                               CONT.HOOK.ICONS-KEY focus ] ].
```

When these constraints are included in users' grammar, other ordinary phrase structure rules have an additional constraint: [INFOSTR-FLAG +] on themselves and their daughter(s).

If the word order is flexible (e.g. v-final and v-initial), no subtype of *head-filler-phrase* is introduced. Instead, *head-subj-phrase* and/or *head-comp-phrase* become twofold, depending on the positioning constraint(s). Such a twofold strategy is the same as how scrambling in Japanese and Korean is constrained with respect to information structure roles (§12.3). In this case, the flag feature INFOSTR-FLAG is also used, because arguments *ex situ* introduce an *info-str* element into ICONS while arguments *in situ* do not. INFOSTR-FLAG serves to make a distinction between them. That is, this strategy is almost the same to that in languages that employ a fixed word order and place focused constituents in the preverbal or postverbal position. The same goes for V2 languages. If a language employs the V2 word order (e.g. Yiddish), all information structure-marked constituents are dealt with in the same way as in (11). §14.5 shows how information structure in V2 languages is customized with reference to two V2 languages: Frisian and Yiddish.

There is still room for refinement, which should be studied in future work. First, there is a need for improvement of free word order languages (e.g. Russian). It is reported that word ordering

variation in in such languages largely depends on information structure (Rodionova, 2001). Grammatical modules for constraining positions of information structure components in free word order languages should be designed in tandem with a study of the full range of word order possibilities. Second, *head-filler* also predicts the possibility of long-distance dependencies, which are not fully tested in the present work. Whether or not using *head-filler* for constraining information structure causes a side effect should be borne out in future work.

14.4 Regression Testing

When developing a grammar library, regression testing using testsuites (a collection of sentences modelled for inspection of implementation) is crucial (Bender et al., 2007). Using a set of testsuites, regression testing checks if a new implementation works well with all the previous functionality in the development of software. That is to say, it should be tested to ensure the newly adapted development is not detrimental to the previous implementation. I ran the regression tests from all previous libraries in order to confirm that my library did not break anything, and then added regression tests to document this library.

14.4.1 Testsuites

The first step is to develop pseudo languages, picking up hypothetical types of languages that that show the full range of information structure marking systems, and then describe a set of testsuites for the pseudo languages. The testsuites represent abstract language types in the space defined by "Information Structure" library.

Testsuites for pseudo languages consist of pseudo words that stand for sentential configurations. Each pseudo word indicates its linguistic category, similar to glosses in interlinear annotation. For example, CN in the string stands for 'Common Noun', IV for 'Intransitive Verb', TV for 'Transitive Verb', and so on. The linear order of the elements within strings simulates the word order. For instance, "CN IV" is an instance of an intransitive sentence like *Dogs bark.* In the pseudo languages that I created for testing this library, there are several specific strings that simulate an *info-str* role. For instance, a morpheme '-FC' or a separate word (i.e. an adposition and a modifier) 'FC' (FoCus), can be used in languages that employ lexical markers to invoke focus meaning. For example, "CN-

FC IN" carries an information structure meaning similar to what DOGS *bark* conveys. Each testsuite include both grammatical pseudo sentences and ungrammatical ones. For example, "IV CN" in which the verb (IN) is inversed may or may not be grammatical depending on whether the grammar allows clause-final or postverbal focus.

The pseudo languages are created according to several factors that have an influence on information structure marking. These include (a) components of information structure (i.e. focus, topic, contrast), (b) word order, and (c) means of expression (i.e. prosody, lexical marking, syntactic configuration). For example, a pseudo language *infostr-foc-svo-initial* is a SVO language and places focused constituents in the clause-initial position.

14.4.2 Pseudo Grammars

The second step is to customize a grammar that covers the testsuites. After a language phenomenon in the testsuites is analyzed and implemented into a library, the library should be verified via regression testing. The final step is to check out if the current system works right using regression tests. Grammatical sentences should be parsed and generated, while ungrammatical ones should not. A parse tree and its MRS (Copestake et al. 2005) representation should indicate information structure roles correctly. This step also includes checking the resulting semantic representations by hand, which then become the gold standard for future runs to check against.

I created 46 pseudo grammars (i.e. 46 choices files) for regression tests of the "Information Structure" library. These grammars are representative of a range of information structure marking in human language. The list of the *choices* files are provided in Appendix E, and some *choices* files are representatively given in Appendix F.

First, I referred to the choices of word order and focus position. There are nine options of word order, excluding free word order: namely SVO, SOV, VSO, VOS, OSV, OVS, v-final, v-initial, and v2. On the other hand, there are four options of focus positions: namely clause-initial, clause-final, preverbal, and postverbal. Thus, using these two factors, logically we can have 36 grammars (9×4). Amongst them, I excluded four grammars which I doubted if such types authentically exist in natural languages. For instance, if a language employs the v-final word order, NPs cannot canonically occur after the verb. If NPs canonically appear in the clause-final or postverbal position, we cannot say that

the language is a genuine v-final language. All human languages presumably have right dislocation constructions (Lambrecht, 1996), but they are non-canonical at least in v-final and v-initial languages. Note that the present work does not use *head-filler-phrase* for these languages. For example, Korean is a v-final language and employs right dislocation (Kim, 2011b), but the constructions does not seem to be *head-filler-phrase*. The excluded ones include `infostr-foc-vf-final`, `infostr-foc-vf-postv`, `infostr-foc-vi-initial`, and `infostr-foc-vi-prev`. Thus, I developed 32 grammars. This subgroup is called TYPE A. Second, three grammars in which multiple positions are used for different components of information structure were added. This subgroup is called TYPE B. The other subgroups in which lexical markers are chosen are TYPE C. Third, three types of lexical markers (affixes, adpositions, and modifiers) that express focus are separately chosen in the creation of pseudo grammars (TYPE C-1). Fourth, the other three components (topic, contrastive focus, contrastive topic) are selected with an option of modifiers (TYPE C-2). Fifth, the categorical choices (e.g. nouns, verbs, and both) and positioning choices are also considered. This provided five more grammars (TYPE C-3).

14.4.3 Processing

The third step is running the regression tests.[10] I ran the whole previous choices files created for regression test without considering ICONS. After getting 100% of matches using the previous testsuites, then I created new gold profiles with ICONS using ACE (`http://sweaglesw.org/linguistics/ace`). The newly created profiles were manually checked to make sure the ICONS were properly computed.

14.5 Testing with Language CoLLAGE

Language CoLLAGE (Collection of Language Lore Amassed through Grammar Engineering) is a repository of student-created grammars on the basis of the LinGO Grammar Matrix system (`http://www.delph-in.net/matrix/language-collage`). This collection of grammars covers a variety of language types in different language families, and a linguistic survey of them could offer

[10]The processor for the regression test was LKB previously, but I reproduced the script to run with ACE. Because there were some minor mismatches in representation between LKB and ACE, some gold profiles used in the regression test were altered.

Table 14.1: Customized grammars with information structure in 2013

name	ISO 639-3	language family
Classical Chinese	[lzh]	Sino-Tibetan
(Northern) Frisian†	[frr]	Indo-European
Halkomelem	[hur]	Salish
Lakota†	[lkt]	Siouan
Miyako†	[mvi]	Japonic
Penobscot	[aaq-pen]	Algic
Yiddish†	[ydd]	Indo-European

valuable insights into language phenomena in human language. Language CoLLAGE currently (Dec. 14, 2013) provides a set of grammars, `choices` files, and testsuites for five languages, and there are many other languages to be curated later. This language resource is readily available under the MIT license.

The grammars have been created in fulfillment of a grammar engineering course in the Department of Linguistics at the University of Washington, Linguistics 567 (`http://courses.washington.edu/ling567`, Bender 2007). In 2013, information structure in seven languages was explored and customized using the initial version of the information structure library in this course. These seven languages are listed in Table 14.1. Of these, there are four languages for which the respective grammar's author gave full permission for the grammar to be used in Language CoLLAGE. They are marked with † in Table 14.1: (Northern) Frisian, Lakota, Miyako, and Yiddish.

After the course concluded, I refined the grammar library for information structure based on the results of the customized grammars and the feedback of their authors. Thus, in the spirit of regression testing, it was necessary to check if the updated library still worked with the students' grammars. I tested whether the newer version provided better representation of information structure, and also did not have an adverse effect on grammar configuration. Moreover, it is also significant to examine how information structure in these languages is articulated and represented. Saleem (2010) makes use of three types of languages for evaluating her "Argument Optionality" library: namely pseudo languages, illustrative languages, and held-out languages. Pseudo languages are hypothetical languages (i.e. not human languages) that indicate the major properties of language phenomenon that the library developer has a keen interest in. Illustrative languages are actual languages whose analysis was considered during the development of the Grammar Matrix library. This contrasts with

held-out languages (i.e. natural languages used in the evaluation of the library only). Thus, the four languages (Frisian, Lakota, Miyako, and Yiddish) in this testing play a similar role to illustrative languages. One difference is that the four grammars used here were already constructed with specifications on information structure properties by the initial library. I used their `choices` files in order to compare the two results produced by the initial library and the newer library.

14.5.1 Languages

The four languages typologically differ from each other, and also employ different strategies of marking information structure. (i) Northern Frisian (spoken in Schleswig-Holstein, Germany) is a V2 language. That is, verbs in Frisian have to appear in the second position in the word order, and the first position can be occupied by subjects or objects. According to the `choices` file created by the developers, Frisian makes use of the preverbal position to indicate focus, and contrastive and non-contrastive focus share this position. Accordingly, the preverbal objects in Frisian are assigned a plain *focus* (a supertype of both *semantic-focus* and *contrast-focus*). (ii) Yiddish is also a V2 language, and employs focus/topic-fronting. That is, focused and topicalized constituents occur sentence-initially.[11] Thus, fronted constituents in Yiddish are assigned *focus-or-topic*. (iii) Miyako (a Ryukyuan language spoken in Okinawa, Japan) is very similar to Japanese. It makes use of information-structure marking adpositions. There are three adpositions of expressing information structure. Two of them signal topic, but they are different in case assignment (i.e. *a* for nominatives *vs*. *baa* for accusatives). The other one spelled as *du* signals focus, and can be used for both nominatives and accusatives. (iv) Lakota (a Siouan language spoken around North and South Dakota, US) uses a specific definite determiner *k'uŋ* to signal contrastive topic.[12] The information-structure related fragments taken from the `choices` files are presented in (12).

---

[11] As surveyed before in §5.2 (p. 92), if focus and topic contest for the sentence-initial position, topic normally wins. However, I have not yet verified if this generalization is straightforwardly applied to V2 languages.

[12] According to the lab notes written by the developers, this article has dual roles as both a quantifier and a topic marker.

(12) a. Northern Frisian

```
section=info-str
focus-pos=preverbal
c-focus=on
```

b. Yiddish

```
section=info-str
focus-pos=clause-initial
topic-first=on
```

c. Miyako

```
section=info-str
  focus-marker1_type=adp
  topic-marker1_type=adp
...
  adp6_orth=a
  adp6_order=after
    adp6_feat1_name=information-structure marking
    adp6_feat1_value=tp
    adp6_feat2_name=information-structure meaning
    adp6_feat2_value=topic
    adp6_feat3_name=case
    adp6_feat3_value=nom
  adp7_orth=du
  adp7_order=after
    adp7_feat1_name=information-structure marking
    adp7_feat1_value=fc
    adp7_feat2_name=information-structure meaning
    adp7_feat2_value=focus
  adp8_orth=baa
  adp8_order=after
    adp8_feat1_name=information-structure marking
    adp8_feat1_value=tp
    adp8_feat2_name=information-structure meaning
    adp8_feat2_value=topic
    adp8_feat3_name=case
    adp8_feat3_value=acc
```

d. Lakota

```
det11_name=def-pst
  det11_stem1_orth=k'uŋ
  det11_stem1_pred=_def-pst_q_rel
  det11_feat1_name=information-structure meaning
  det11_feat1_value=contrast-topic
```

14.5.2 Testsuites

The numbers of sentences in each testsuite for the four languages are shown in Table 14.2. Note that testsuites consist of both grammatical sentences and ungrammatical sentences. Each testsuite also includes test items that represent how information structure is configured for the language.

Table 14.2: # of test items

language	# of total items	# of grammatical items	# of information-structure related items
Frisian	164	109	6
Yiddish	228	150	6
Miyako	102	71	6
Lakota	168	100	2

14.5.3 Comparison

The data set of Language CoLLAGE includes the final grammar and the `choices` file, other than the testsuite. Using the `choices` file, I created two different versions of grammars. One was customized by the previous library, and the other was customized by the new library. These two versions of grammars are represented as 'old' and 'new' respectively hereafter. I ran these two grammars plus the final grammar ('final') provided by each developer to see the coverage and the number of parse trees. Using the LKB and [incr tsdb()], I parsed all test items in the testsuites for each language, and then examined how many sentences were covered by each grammar (i.e. coverage) and how many readings were produced (i.e. number of parse trees).

First, coverage of these three types of grammars are compared. The grammars created only using the `choices` file include the main linguistic modules that can be fully created on the LinGO Grammar Matrix customization system, while the final grammars ('final') contain more elaborated types and rules that developers manually edited. Accordingly, the final grammars always yield better coverage than the other two versions of grammars. Regarding 'old', and 'new', ideally, the coverage between the grammars created by the old library and those created by the new library should be the same. That is to say, the distinction between handling grammatical sentences and ungrammatical sentences must not be changed. The coverage that each grammar produced were calculated as shown in Table 14.3. As indicated in the third and fourth columns of Table 14.3, there was no difference in coverage between the two versions of grammars.

Second, the number of parse trees (i.e. readings) may or may not be changed. This is because I elaborated phrase structure rules that place constraints on syntactic positioning of marking information structure. In particular, one of the main components that I refined in the newer version is a routine that deals with narrow foci in V2 languages. In fact, the old version had vulnerability in con-

Table 14.3: Coverage (%)

language	final	old	new
Frisian	70.6	45.0	45.0
Yiddish	60.0	32.0	32.0
Miyako	77.5	38.0	38.0
Lakota	91.0	60.0	60.0

Table 14.4: # of Readings

language	final	old	new
Frisian	178	195	209
Yiddish	118	97	98
Miyako	80	34	34
Lakota	103	62	62

straining narrow foci in V2 languages, and syntactic composition did not work well. As shown in the third and fourth column of Table 14.4, the numbers of parse trees produced by the grammars in Miyako and Lakota are the same, while those in V2 languages increase in the new versions. I manually checked whether the newly produced parse trees were properly constructed and their semantic representations were correct. That implies that the newer version performs better.

14.5.4 Information Structure in the Four Languages

Finally, I checked out how information-structure related test items, whose numbers are given in the last column of Table 14.2, were parsed and represented in the ICONS list. I found that the customized grammars had complete coverage over these items and returned correct analyses.

Frisian, a V2 language, is specified as placing focused constituents in the preverbal position irrespective of contrastiveness. As discussed before in §14.3.2, this language includes *head-nf-comp-phrase-super*, *nf-comp-head-phrase*, and *narrow-focused-phrase*. The value of *info-str* that preverbal foci have is *focus* which can be used for both *semantic-focus* and *contrast-focus*.

Yiddish employs focus/topic-fronting. The grammar for Yiddish also includes *head-nf-comp-phrase-super*, *nf-comp-head-phrase*, and *narrow-focused-phrase* like Frisian, and the value of *info-str* that fronted constituents involve is constrained as *focus-or-topic*.

Three adpositions that mark information structure in Miyako were also inspected. For example, the nominative topic marker *a* in Miyako is customized as follows.

(13)
```
top-marker := case-marking-adp-lex &
  [ STEM < "a" >,
    SYNSEM.LOCAL [ CAT [ VAL.COMPS < [ LOCAL.CONT.HOOK.INDEX #target ] >,
                         HEAD.CASE nom,
                         MKG tp ],
                   CONT [ ICONS <! #icons & info-str &
                                   [ TARGET #target ] !>,
                          HOOK.ICONS-KEY #icons & topic ] ] ].
```

The adpositions introduce an *info-str* element into ICONS, and the value is successfully copied up the trees.

The topic-marking determiner *k'uŋ* in Lakota is an instance of *def-pst-determiner-lex* in the grammar, and the type is described as follows.

(14)
```
infostr-marking-determiner-lex := basic-determiner-lex & one-icons-lex-item &
  [ SYNSEM.LOCAL [ CAT.VAL.SPEC.FIRST.LOCAL.CONT.HOOK [ INDEX #target,
                                                        ICONS-KEY #icons ],
                   CONT.ICONS <! info-str & #icons & [ TARGET #target] !> ] ].

def-pst-determiner-lex := determiner-lex & infostr-marking-determiner-lex &
  [ SYNSEM.LOCAL.CAT.VAL.SPEC.FIRST.LOCAL.CONT.HOOK.ICONS-KEY contrast-topic ].
```

The *infostr-marking-determiner-lex* type includes an element in CONT|ICONS (i.e. *one-icons-lex-item*), and *def-pst-determiner-lex* constrains the value as *contrast-topic*. This value comes from the user's choice given in (12d).

14.5.5 Summary

This section substantiates whether my newer version of the information structure library works well using four grammars and `choices` provided in Language CoLLAGE (Frisian, Lakota, Miyako, and Yiddish). I customized four old versions of grammars as well as four new versions of grammars using the `choices` files. Exploiting the testsuites also included in Language CoLLAGE, I ran the grammars to see if there was no change in coverage, and how many parse trees were produced. Notably, I recognized that the newer version yielded better performance in manipulating information structure in V2 languages (Frisian and Yiddish). Additionally, I verified that information structure values were properly constrained and the values were incrementally augmented in the ICONS list. In summary, I confirmed that the newer version correctly operated at least with the four languages.

14.6 Live-site

All the components of the information structure library (e.g. (i) web-based questionnaire, (ii) the Matrix-core in TDL, and (iii) the Python code for validation and customization) were successfully implemented in the LinGO Grammar Matrix system. The library for information structure was added in the live site of the customization system (`http://www.delph-in.net/matrix/`

customize), so the functionality of the information structure library is now available for all users of the Grammar Matrix customization system.

14.7 Download

The source code is downloadable in the subversion repository of the LinGO Grammar Matrix system (svn://lemur.ling.washington.edu/shared/matrix/trunk). The specific version that this dissertation describes is separately provided, and can be obtained from svn://lemur.ling.washington.edu/shared/matrix/branches/sanghoun. This version is also independently served in another web page, whose url is http://depts.washington.edu/uwcl/sanghoun/matrix.cgi.

Chapter 15

MULTILINGUAL MACHINE TRANSLATION

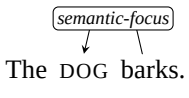
Using information structure for multilingual machine translation can improve performance. The main contribution that using information structure makes is that a machine translation system informed by information structure is capable of reducing the number of infelicitous translations dramatically. This reduction has two effects on the performance of transfer-based machine translation (Song and Bender, 2011): First, the processing burden of the machine translation component which ranks the translations and selects only suitable results can be greatly lightened, which should improve translation speed. Second, although it is still necessary to employ a re-ranking model for choosing translations, we can start from a refined set of translations, which should improve translation accuracy.

§15.1 offers an explanation of how ICONS (Individual CONStraints) operates in transfer-based machine translation. §15.2 addresses the processor the current work employs for testing machine translation. §15.3 conducts an evaluation to examine how many infelicitous translations are filtered out by means of ICONS.

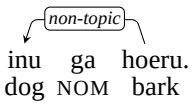
15.1 Basic Machinery

A graph presented in (1a) represents an English sentence in which the subject *the* DOG bears the A-accent, thereby plays the role of *semantic-focus*. The second graph in (1b) represents the Japanese translation, in which the subject *inu* 'dog' is combined with the nominative marker *ga* that signals *non-topic*. That is to say, although they are proper translations to each other, information is differently structured.

(1) a.



b.



Nonetheless, it is noticeable that *non-topic* is a supertype of *semantic-focus* in the type hierarchy of

info-str given in Figure 9.1 (p. 174). This ability to partially specify information structure allows us to reduce the range of outputs in translation while still capturing all legitimate possibilities.

Two hypothetical suffixes '-a' and '-b' are employed for testing hereafter, and they represent the A and B accents in English (Bolinger, 1961; Jackendoff, 1972) respectively. Note that the '-b' suffix cannot be attached to the verb *barks*, because verbs presumably cannot be marked via B-accent for the information structure role of *topic* in English. *The dog barks* without any information structure marking logically can be interpreted as six types of sentences (3×2).

(2) dog dog: [ICONS: < >]
dog-a: [ICONS: < e2 semantic-focus x4 >]
dog-b: [ICONS: < e2 contrast-or-topic x4 >]
bark barks: [ICONS: < >]
barks-a: [ICONS: < e2 semantic-focus e2 >]

However, if we apply ICONS to generation, we can filter out sentences which are not informatively equivalent to the input sentence. For example, if the input sentences are *The* DOG *barks* and *The* ***dog*** *barks* in which the subject bears the A and B accents respectively, they can be monolingually paraphrased as (3). That is, four infelicitous sentences from each set of sentences can be removed. Two sentences in (3a-i) and (3a-iii) cannot be generated because the subject does not include any value in ICONS. In other words, information structure-marked constituents in the source cannot be generated as an unmarked constituent in the target. Two sentences in (3a-v) and (3a-vi) cannot be generated, either. This is because the B-accented subject conveys *contrast-or-topic* which is incompatible with *semantic-focus*. The same goes for (3b): Since only the last two sentences are compatible with the information structure meaning that the input sentence conveys, the others cannot be paraphrased with respect to information structure.

(3) a. The dog-**a** barks [ICONS: < e2 **semantic-focus** x4 >]
(i) ~~The dog barks~~
(ii) The dog-a barks
(iii) ~~The dog barks-a~~
(iv) The dog-a barks-a
(v) ~~The dog-b barks~~
(vi) ~~The dog-b barks-a~~

b. The dog-**b** barks [ICONS: < e2 **contrast-or-topic** x4 >]
   (i) ~~The dog barks~~
   (ii) ~~The dog-a barks~~
   (iii) ~~The dog barks-a~~
   (iv) ~~The dog-a barks-a~~
   (v) The dog-b barks
   (vi) The dog-b barks-a

The same goes for Japanese in which lexical markers signal information structure. There are at least three Japanese translations (i.e. case-marking, *wa*-marking, and null-marking) corresponding to *The dog barks*, but case-marked NPs cannot be paraphrased into *wa*-marked NPs within our *info-str* hierarchy given in Figure 9.1, and vice versa. Note that null-marked items in Japanese (e.g. *inu* in (4a-iii) and (4b-iii)) are assigned *non-focus* (Yatabe, 1999), which is compatible with both *non-topic* in (4a) and *contrast-or-topic* (4b). Thus, *inu-ga/wa hoeru* can be paraphrased into *inu hoeru*.

(4) a. inu **ga** hoeru [ICONS: < e2 **non-topic** x4 >]
   (i) inu ga hoeru
   (ii) ~~inu wa hoeru~~
   (iii) inu hoeru

b. inu **wa** hoeru [ICONS: < e2 **contrast-or-topic** x4 >]
   (i) ~~inu ga hoeru~~
   (ii) inu wa hoeru
   (iii) inu hoeru

Translating across languages is constrained in the same manner. An English sentence (5a) cannot be translated into (5a-ii) and (5a-iii), because the *semantic-focus* role that DOG involves is incompatible with the *contrast-or-topic* role that *wa* assigns and the *non-focus* role that the null marker (indicated by Ø) involves. On the other hand, a Japanese sentence (5b) can be translated into only (5b-ii) and (5b-iv), because *non-topic*, which comes from the nominative marker *ga*, is contradictory to *contrast-or-topic* that the B-accent signals in English and the constituent corresponding to the *ga*-marked subject should introduce an *info-str* element into ICONS.

(5) a. The dog-**a** barks [ICONS: < e2 **semantic-focus** x4 >]

(i) inu ga hoeru

(ii) ~~inu wa hoeru~~

(iii) ~~inu hoeru~~

b. inu **ga** hoeru [ICONS: < e2 **non-topic** x4 >]

(i) ~~The dog barks~~

(ii) The dog-a barks

(iii) ~~The dog barks-a~~

(iv) The dog-a barks-a

(v) ~~The dog-b barks~~

(vi) ~~The dog-b barks-a~~

15.2 Processor

The processor the present work uses for the purpose of evaluation is ACE (`http://sweaglesw.org/linguistics/ace`). ACE, using DELPH-IN grammars, such as Grammar Matrix grammars created by the customization system (Bender and Flickinger, 2005; Drellishak, 2009; Bender et al., 2010b) and resource grammars (e.g. the ERG (English Resource Grammar, Flickinger 2000)), parses the sentences of natural languages, and generates sentences based on the MRS (Minimal Recursion Semantics, Copestake et al. 2005) representation that the parser creates. ACE is the first DELPH-IN processor to specifically handle ICONS as part of the MRS.[1]

When creating the data file of ACE, ACE refers to parameters described in *ace/config.tdl*. In the configuration file, grammar users can choose whether or not ICONS is used in MRS representation. The snippet that enables ICONS to be included in MRS representation is as follows.

(6)
```
enable-icons := yes.
mrs-icons-list := ICONS LIST.
icons-left := CLAUSE.
icons-right := TARGET.
```

ACE carries out ICONS-based generation via subsumption check, using the type hierarchy *info-str* (presented in Figure 9.1 (p. 174)). ACE generates all potential sentences that logically fit in the input MRS not considering the constraints on ICONS beforehand. After that, if the data file of the

[1] *agree* (Slayden 2012) also uses ICONS for constraining information structure in parsing and generation.

grammar for generation is compiled with the parameter given in (6), ACE start postprocessing the intermediate results. Depending on the subsumption relationship of information structure meanings, sentences mismatching the values in the ICONS list are filtered out in this step. For example, if *semantic-focus* is assigned to a specific individual in the source MRS, only outputs that provide an ICONS element for that individual can be produced. The *info-str* value an individual has in the output should be the same as that in the input (i.e. *semantic-focus*) or its supertypes (e.g. *focus*, *non-topic*, etc.). For instance, an A-accented constituent in English (e.g. DOG) contributes an ICONS element whose value is *semantic-focus*, and this element is translated as a *ga*-marked constituent (e.g. *inu ga*) whose value is monolingually *non-topic* in Japanese. Note that *non-topic* subsumes *semantic-focus* in the type hierarchy presented in Figure 9.1. A completely underspecified output for each ICONS element is not acceptable in generation. For instance, an A-accented DOG that introduces an ICONS element cannot be paraphrased as an unaccented *dog* that does not contribute any ICONS element. By contrast, the opposite direction is acceptable. If a constituent introduces no ICONS element in the input, the output can include an information-structure marked constituent. For instance, an unaccented *dog* can be paraphrased as an A-accented DOG in generation.

15.3 Evaluation

15.3.1 Illustrative Grammars

As stated earlier in Chapter 2, for verifying a linguistic hypothesis with reference to a computational grammar, it would be better to use a compact grammar presenting the fundamental rules in a precise manner. Illustrative grammars are constructed for this purpose. The illustrative languages used here are English, Japanese, and Korean. These languages are chosen, because the resource grammars for each of the language (ERG (English Resource Grammar, Flickinger 2000) for English, Jacy (Siegel and Bender 2002) for Japanese, and KRG (Korean Resource Grammar, Kim et al. 2011) for Korean) will be the main concern in my further study. The information structure properties each language has are summarized in the following subsections.

English

As is well-known, English employs prosody for expressing information structure. Without consideration of the prosodic patterns, we could not draw the basic picture of information-structure related phenomena in English.[2] There are quite a few previous studies on how prosody is realized with respect to information structure (Jackendoff, 1972; Steedman, 2000; Kadmon, 2001; Büring, 2003; Hedberg, 2006), but there seems to be no clear consensus (as surveyed earlier in §4.2). The illustrative grammar for English makes use of just the traditional distinction of the A and B accents (Bolinger, 1958). In other words, the two hypothetical suffixes '-a' and '-b' are used. However, the meanings that the accents take charge of are differently represented from the traditional approach. The information structure meanings that '-a' and '-b' convey are marked following Hedberg's argument: '-a' for *semantic-focus* and '-b' for *contrast-or-topic*. The AVMs are already presented in §10.4.1 (p. 219).

Korean

The illustrative grammar for Korean includes two kinds of grammatical sets of constraints for expressing information structure. The first one employs lexical markers, such as *i/ka* and (*l*)*ul* for case marking, (*n*)*un* for topic marking, and Ø for null marking. The AVMs for these markers are presented in §10.4.2 (p. 224). The second fragment aims to handle scrambling. The AVMs for constraining scrambling constructions are provided in §12.3 (p. 257). These AVMs use different rules instantiating *head-subj-phrase* and *head-comp-phrase* with reference to lexical markings of daughters (i.e. MKG).

Japanese

As mentioned before, this dissertation respects the traditional ways of dealing with lexical markers in Japanese and Korean from different points of view. While lexical markers in Korean are dealt with as suffixes (Kim and Yang, 2004), those in Japanese are treated as adpositions (Siegel, 1999, among

[2]English also makes use of some constructional means to configure focus and topic. These include focus/topic fronting, clefting, etc. Nonetheless, they have to do with various grammatical components. For example, implementing grammatical modules for cleft constructions necessitates lots of TDL statements for relative clauses as an essential prerequisite. This involves too many complexity for an illustrative grammar to cover. For this reason, the illustrative grammar for English in this evaluation is exclusively concerned with prosody.

Table 15.1: # of outputs without ICONS

	eng	jpn	kor
eng	144	126	126
jpn	990	180	180
kor	1080	198	198

Table 15.2: # of outputs with ICONS

	eng	jpn	kor
eng	53	39	39
jpn	150	120	150
kor	140	115	154

others). Other than the difference, the illustrative grammar for Japanese has the same configuration as that for Korean explained above. Notably, the null marker in Japanese is constrained by a lexical rule in the current work (p. 221), which is different from previous HPSG-based suggestion about so-called case-ellipsis (Yatabe, 1999; Sato and Tam, 2012).

15.3.2 Testsuites

The testsuites (i.e. a collection of sentence to be modelled) for this multilingual machine translation testing are provided in Appendix G and the set of English sentences are as follows.

(7) [1] The dog barks.
[2] The dog-a barks.
[3] The dog barks-a.
[4] The dog-b barks.
[5] The dog-b barks-a.
[6] The dog-a barks-a.
[7] Kim reads the book.
[8] Kim-a reads the book.
[9] Kim reads-a the book.
[10] Kim reads the book-a.
[11] Kim-b reads-a the book.
[12] Kim-b reads the book-a.

There are one intransitive sentence and one transitive sentence in English, and they are encoded with two hypothetical suffixes and differentiated as allosentences.

15.3.3 An Experiment

All test items presented in (7) and their translations in Japanese and Korean are parsed, transferred, and generated. Table 15.1 and Table 15.2 show the number of translation results in each translation

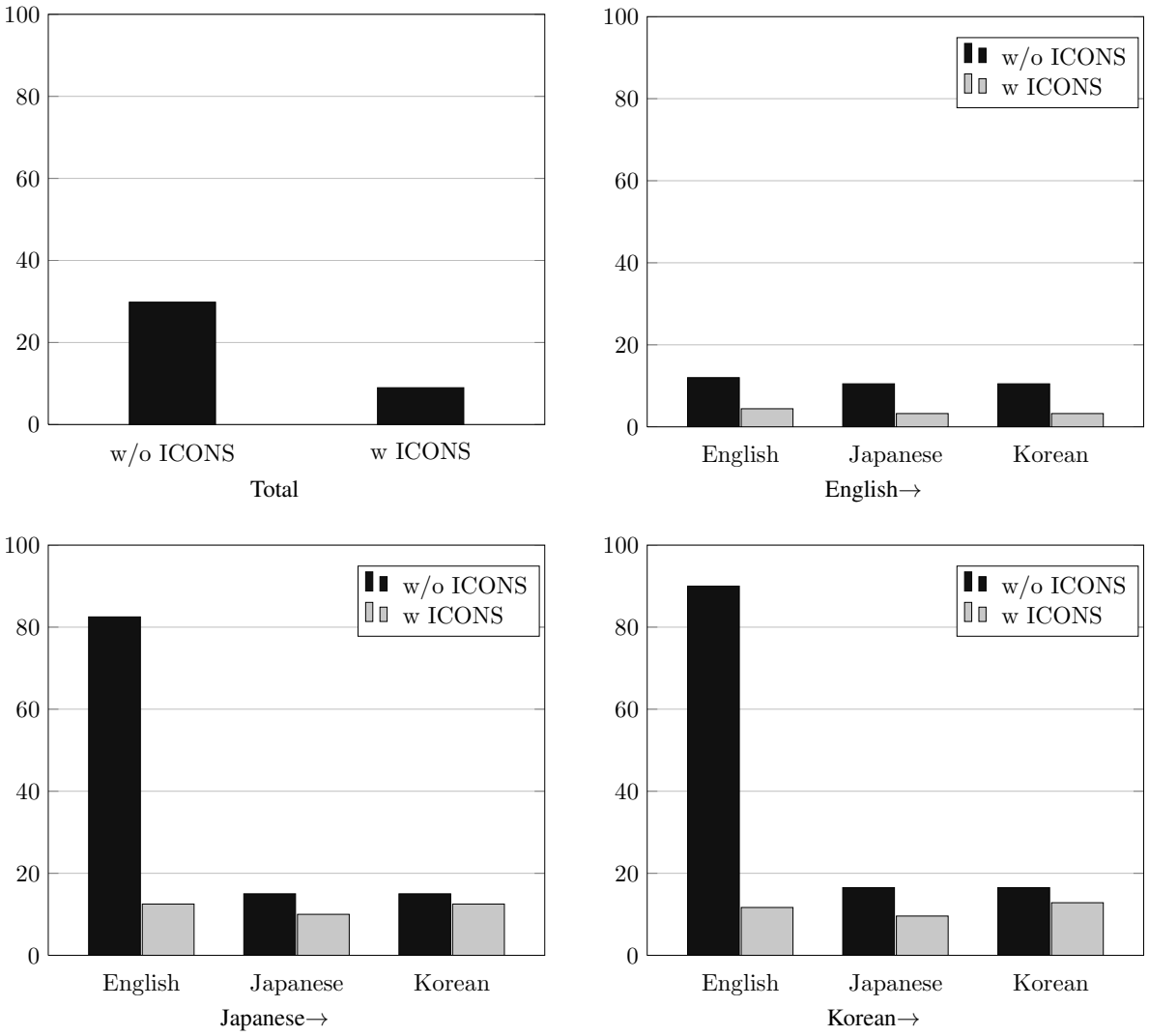


Figure 15.1: Average # of outputs

pair. The first column in each table indicates the source language, and the first row indicates the target language. For example, English→Japanese produces 126 translations when not using ICONS, and 39 translations when using ICONS.

As indicated in the tables, the number of generated sentences dramatically decreases when using ICONS. The total number of translation outputs in Table 15.1 is 3,222, while that in Table 15.2

is merely 960.[3] That means approximately 70% translations are filtered out in total when using ICONS.

The four charts in Figure 15.1 compare the average number of outputs in total and in each translation pair. The decrease indicated in each bar chart shows that information structure can be used to filter out inappropriate sentences from the translation result.

These charts show that when translating Japanese and Korean to English many outputs are filtered out. The main reason for the dramatic decrease in Japanese→English and Korean →English is that the illustrative grammar for English includes a lexical rule to mark focus on verbal items, while the illustrative grammars for Japanese and Korean do not. When a verb is focused in the English grammar, the lexical rule introduces a *focus* element into ICONS. In contrast, verbs cannot involve any *info-str* value in the current illustrative grammars for Japanese and Korean. Thus, the huge difference in Japanese→English and Korean→English is largely caused by the different marking system for verbal items.

Finally, I verified 108 sets of translation outputs (9 directions × 12 test items) by hand. The same problem was also found. When an English item includes an A-accented verb (e.g. *barks-a* and *reads-a*), the item cannot be translated into Japanese and Korean. This suggest that there might be a problem with the strategy of requiring some information structure marking in the output for an item if there is some in the input. In future work there will be a way to relativize that to the possibility of any marking on that item. Other than this difference, the translation outputs were legit and felicitous. I also sampled the filtered translations to verify that they were all infelicitous ones, and found that this information structure-based model works for them, too.

15.4 Summary

It is my firm opinion that translating should mean reshaping the ways in which information is conveyed, not simply changing words and reordering phrases. In almost all human languages, articulation of sentential constituents is conditioned by information structure. Because the means of expressing information structure differs amongst languages, identifying how information is struc-

[3]When the source language is not English and the target language is English, the numbers are rather big in Table 15.1. This is because English employs number and COG-ST features, while Japanese and Korean do not. For example, *inu* in Japanese can be translated into at least four NP types in English: *a dog*, *the dog*, *the dogs*, and *dogs*.

tured in a given set of languages plays a key role in achieving felicity in a machine translation between them. Hence, information structure is of great help to multilingual machine translation in that information structure facilitates more felicitous translations. This chapter conducted a small experiment to prove this hypothesis. I created three illustrative grammars for English, Japanese, and Korean following my ICONS-based analyses presented thus far. In the test of transfer-based machine translation, I found that using information structure served to filter out infelicitous translations dramatically. This testing should be further elaborated in future work using resource grammars, such as the ERG (Flickinger 2000), Jacy (Siegel and Bender 2002), and the KRG (Kim et al. 2011).

Part VI

CONCLUSION

This part concludes my dissertation. I present a summary review of the main content of my dissertation, and then enumerate the contributions this study makes to theoretical linguistics and language processing. I conclude with several proposals for future work in keeping with the current research interests.

Chapter 16

CONCLUSION

16.1 Summary

This dissertation begins with key motivations laid out Chapter 1 for the creation of a computational model of information structure. Chapter 2 offers background knowledge for understanding the main proposals of this dissertation, including an overview of HPSG (Pollard and Sag 1994) and MRS (Copestake et al. 2005).

Part II scrutinizes meanings and markings of information structure from a cross-linguistic standpoint. Information structure is composed of four components: focus, topic, contrast, and background. Focus identifies that which is important and/or new in an utterance, which cannot be removed from the sentence. Topic can be understood as what the speaker is speaking about, and does not always appear in a sentence (unlike focus). Contrast applies to a set of alternatives, which can be realized as either focus or topic. Lastly. background is defined as that which is neither focus nor topic. There are three means of expressing information structure; prosody, lexical markers, and syntactic positioning. There are three lexical types responsible for marking information structures; morphemes, adpositions, and modifiers (e.g. clitics). Canonical positions of focus include clause-initial, clause-final, preverbal, and postverbal. Building upon these fundamental notions, Chapter 5 looks into several cases in which discrepancies in form-meaning mapping of information structure happen.

Part III describes data compilation for exploiting a naturally occurring text in four languages. The annotated text consists of a translation set of the first 100 sentences in *The Adventure of the Speckled Band*. The annotation follows guidelines provided in Dipper et al. (2007), with the addition of dropped elements and a distinction between inner frames and outer frames within a discourse. From the annotated data, several intriguing properties across four languages can be observed. These help establish the HPSG/MRS-based formalism in Part IV.

Part IV proposes using ICONS (Individual CONStraints) for representing information structure

in MRS (Copestake et al. 2005). This is motivated by three factors. First, information structure markings should be distinguished from information structure meanings in order to solve the apparent mismatches between them. Second, the representation of information structure should be underspecifiable, because there are many sentences whose information structure cannot be conclusively identified in the context of sentence-level, text-based processing. Third, information structure should be represented as a binary relation between an individual and a clause. In other words, information structure roles should be filled out as a relationship with the clause a constituent belongs to, rather than as a property of a constituent itself. In order to meet these requirements, three type hierarchies are suggested; *mkg*, *sform*, and most importantly *info-str*. In addition to them, two types of flag features, such as L/R-PERIPH and LIGHT, are used for configuring focus and topic. Using hierarchies and features, the remaining chapters of Part IV scrutinize multiclausal utterances and specific forms of expressing information structure and also calculate focus projection via ICONS.

Part V creates a customization system for implementing information structure within the LinGO Grammar Matrix (Bender et al. 2010b) and examines how information structure improves transfer-based multilingual machine translation. Building on cross-linguistic and corpus-based findings, a large part of HPSG/MRS-based constraints presented in Part IV is implemented in TDL. A web-based questionnaire is designed in order to allow users to implement information structure constraints within the `choices` file. Common constraints across languages are added into the Matrxi core (`matrix.tdl`), and language-specific constraints which depend on the users' choices are processed by Python scripts and stored into the customized grammar. Evaluations of this library using regression tests and Language CoLLAGE (`http://www.delph-in.net/matrix/language-collage`) show that this library works well with various types of languages. Finally, an experiment of multilingual machine translation bears out that information structure can be used to reduce the number of infelicitous translations dramatically.

16.2 Contributions

This dissertation holds particular significance for general theoretic studies of the grammar of information structure. First of all, quite a few languages are surveyed to capture cross-linguistic generalizations about information structure meanings and markings, which can serve as an impor-

tant milestone for typological research on information structure. Second, the data collection in the corpus study, despite its small size, offers distributional findings on information structure. In particular, the data set presents its own implications given that it is comprised of a set of parallel annotated sentences, and that the set of included languages presents a diverse set of information structure properties. Since the data set is readily available, it is my hope that other linguists exploit the compiled data for their own research goods.

This dissertation also makes a contribution to HPSG/MRS-based studies by enumerating strategies for representing meanings and markings of information structure within the formalism in a comprehensive and fine-grained way. Notably, this dissertation establishes a single formalism for representation and applies this formalism to various types of forms in a straightforward and cohesive manner. Moreover, this dissertation addresses how information structure can be articulated within the HPSG/MRS framework and implemented within a computational system in the context of grammar engineering.

This dissertation also shows that information structure can be used to produce better performance in natural language processing systems. My firm opinion is that information structure contributes to multilingual processing; languages differ from each other not merely in the words and phrases employed but in the structuring information. It is my expectation that this study will inspire future studies in computational linguistics to pay more attention to information structure.

Last but most importantly, this dissertation makes a contribution to the LinGO Grammar Matrix library. The actual library makes it easy for other developers to adopt and build on my analyses of information structure. Moreover, the methodology of creating libraries I take in this study can be used for other libraries in the system. In order to construct the model in a fine-grained way, I collected cross-linguistic findings about information structure markings and exploited a multilingual parallel text in four languages. These two methods are essential in further advancements in the LinGO framework.

16.3 Future Work

This dissertation closes with a brief look at directions for improvement in the future.

First, it is necessary to examine other types of particles responsible for marking information

structure. Not all focus sensitive items are entirely implemented in TDL in this dissertation even for English. Japanese and Korean employ a variety of lexical markers for expressing focus and topic, which are presented in Hasegawa (2011) and Lee (2004). A few focus markers in some languages have a positional restrictions. For example, as shown in Chapter 4 (§4.3), the clitic *tvv* in Cherokee signals focus and the focused constituent with *tvv* should be followed by other constituents in the sentence. That is, two means of marking information structure operate at the same time. These kinds of additional constraints will be interestingly investigated in the future.

Second, a few more types of constructions related to information structure will be studied in future work. The constructions include echo questions (§7.1.4), *Yes*/*No*-questions (King, 1995), coordinated clauses (Heycock, 2007), double nominative constructions (Kim and Sells, 2007; Choi, 2012), floating quantifiers (Yoshimoto et al., 2006; Kim, 2011a), pseudo clefts (Kim, 2007), and *it*-clefts in other languages in the DELPH-IN grammars (e.g. Japanese (Hiraiwa and Ishihara, 2002; Kizu, 2005) and Korean (Kim and Yang, 2009)).

Third, the method for computing focus projection in this dissertation also needs to be more thoroughly examined. There are various constraints on how focus can be spread to larger constituents. These are not addressed in this dissertation, which looks at the focus projection of only simple sentences in English. The method the present work employs for handling focus projection could be reinforced in further studies.

Fourth, it would be interesting for future work to delve into how scopal interpretation can be dealt with within the framework that this dissertation proposes. Topic has an influence on scopal interpretation in that topic has the widest scope in a sentence (Büring, 1997; Portner and Yabushita, 1998; Erteschik-Shir, 2007). MRS employs HCONS (Handle CONStraints) in order to resolve scope ambiguity. Further work can confirm whether HCONS+ICONS is able to handle the relationship between topic and scope resolution.

Finally, the evaluation of multilingual machine translation will be extended with a large number of test suites. More grammatical fragments related to ICONS will be incorporated into the DELPH-IN resource grammars, such as ERG (English Resource Grammar, Flickinger 2000), Jacy (Siegel and Bender 2002), KRG (Korean Resource Grammar, Kim et al. 2011), SRG (Spanish Resource Grammar, Marimon 2012), and so forth.

Bibliography

Abeillé, Anne and Godard, Daniele. 2001. A Class of "Lite" Adverbs in French. In Joaquim Camps and Caroline R. Wiltshire (eds.), *Romance Syntax, Semantics and L2 Acquisition: Selected papers from the 30th Linguistic Symposium on Romance Languages, Gainesville, Florida, February 2000*, pages 9–26, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Alexopoulou, Theodora and Kolliakou, Dimitra. 2002. On Linkhood, Topicalization and Clitic Left Dislocation. *Journal of Linguistics* 38(2), 193–245.

Alonso-Ovalle, Luis, Fernández-Solera, Susana, Frazier, Lyn and Clifton, Charles Jr. 2002. Null vs. Overt Pronouns and the Topic-Focus Articulation in Spanish. *Italian Journal of Linguistics* 14, 151–170.

Ambar, Manuela. 1999. Aspects of the Syntax of Focus in Portuguese. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 23–54, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Arnold, Jenifer E. 2010. How Speakers Refer: The Role of Accessibility. *Language and Linguistics Compass* 4(4), 187–203.

Arregi, Karlos. 2000. Tense in Basque (Ms.).

Arregi, Karlos. 2003. Clitic Left Dislocation is Contrastive Topicalization. *U. Penn Working Papers in Linguistics* 9(1), 31–44.

Avgustinova, Tania and Zhang, Yi. 2010. Conversion of a Russian Dependency Treebank into HPSG Derivations. In *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories*, Tartu, Estonia.

Baldwin, Timothy. 1998. *The Analysis of Japanese Relative Clauses*. Doctoral Dissertation, Tokyo Institute of Technology.

Baldwin, Timothy, Beavers, John, Bender, Emily M., Flickinger, Dan, Kim, Ara and Oepen, Stephan. 2005. Beauty and the Beast: What Running a Broad-coverage Precision Grammar over

the BNC Taught us about the Grammar — and the Corpus. In Stephan Kepser and Marga Reis (eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, pages 49–70, Berlin: Mouton de Gruyter.

Beaver, David, Clark, Brady Zack, Flemming, Edward, Jaeger, T Florian and Wolters, Maria. 2007. When Semantics Meets Phonetics: Acoustical Studies of Second-occurrence Focus. *Language* 83(2), 245–276.

Beaver, David I. and Clark, Brady Z. 2008. *Sense and Sensitivity: How Focus Determines Meaning*. Malden, MA: Wiley-Blackwell.

Bender, Emily M. 2007. Combining Research and Pedagogy in the Development of a Crosslinguistic Grammar Resource. In *Proceedings of the Workshop on Grammar Engineering across Frameworks (GEAF07)*, Stanford, CA.

Bender, Emily M. 2008. Grammar Engineering for Linguistic Hypothesis Testing. In Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer and Elias Ponvert (eds.), *Proceedings of the Texas Linguistics Society X Conference: Computational Linguistics for Less-Studied Languages*, pages 16–36, Stanford, CA: CSLI Publications.

Bender, Emily M. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology. Special Issue on Interaction of Linguistics and Computational Linguistics* 6(3), 1–26.

Bender, Emily M., Drellishak, Scott, Fokkens, Antske, Goodman, Michael Wayne, Mills, Daniel P., Poulson, Laurie and Saleem, Safiyyah. 2010a. Grammar Prototyping and Testing with the LinGO Grammar Matrix Customization System. In *Proceedings of the ACL 2010 System Demonstrations*, pages 1–6, Uppsala, Sweden: Association for Computational Linguistics.

Bender, Emily M., Drellishak, Scott, Fokkens, Antske, Poulson, Laurie and Saleem, Safiyyah. 2010b. Grammar Customization. *Research on Language & Computation* 8(1), 23–72.

Bender, Emily M. and Flickinger, Dan. 2005. Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core. In *Proceedings of the 2nd Interna-*

tional Joint Conference on Natural Language Processing IJCNLP-05: Posters/Demos, Jeju Island, Korea.

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

Bender, Emily M. and Goss-Grubbs, David. 2008. Semantic Representations of Syntactically Marked Discourse Status in Crosslinguistic Perspective. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 17–29, Association for Computational Linguistics.

Bender, Emily M., Poulson, Laurie, Drellishak, Scott and Evans, Chris. 2007. Validation and Regression Testing for a Cross-linguistic Grammar Resource. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 136–143, Prague, Czech Republic: Association for Computational Linguistics.

Bianchi, Valentina and Frascarelli, Mara. 2010. Is Topic a Root Phenomenon? *Iberia* 2(1), 43–88.

Bildhauer, Felix. 2007. *Representing Information Structure in an HPSG Grammar of Spanish*. Doctoral Dissertation, Universität Bremen.

Bildhauer, Felix. 2008. Clitic Left Dislocation and Focus Projection in Spanish. In Stefan Müller (ed.), *Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar*, pages 346–357, Stanford, CA: CSLI Publications.

Bildhauer, Felix and Cook, Philippa. 2010. German Multiple Fronting and Expected Topichood. In Stefan Müller (ed.), *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, pages 68–79, Stanford, CA: CSLI Publications.

Bjerre, Anne. 2011. Topic and Focus in Local Subject Extractions in Danish. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 270–288, Stanford, CA: CSLI Publications.

Bolinger, Dwight Le Merton. 1958. A Theory of Pitch Accent in English. *Word* 14, 109–149.

Bolinger, Dwight Le Merton. 1961. Contrastive Accent and Contrastive Stress. *Language* 37(1), 83–96.

Bolinger, Dwight Le Merton. 1977. *Meaning and Form*. London: Longman.

Bonami, Olivier and Delais-Roussarie, Elisabeth. 2006. Metrical Phonology in HPSG. In Stefan Müller (ed.), *Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 39–59, Stanford, CA: CSLI Publications.

Bond, Francis. 2005. *Translating the Untranslatable: A Solution to the Problem of Generating English Determiners*. Stanford, CA: CSLI Publications.

Bond, Francis, Fujita, Sanae and Tanaka, Takaaki. 2006. The Hinoki Syntactic and Semantic Treebank of Japanese. *Language Resources and Evaluation* 40(3–4), 253–261.

Bond, Francis, Isahara, Hitoshi, Fujita, Sanae, Uchimoto, Kiyotaka, Kuribayashi, Takayuki and Kanzaki, Kyoko. 2009. Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources*, Singapore.

Bouma, Gerlof, Øvrelid, Lilja and Kuhn, Jonas. 2010. Towards a Large Parallel Corpus of Cleft Constructions. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC10)*, pages 3585–3592, Valletta, Malta.

Branco, António and Costa, Francisco. 2010. A Deep Linguistic Processing Grammar for Portuguese. In *Computational Processing of the Portuguese Language*, volume LNAI6001 of *Lecture Notes in Artificial Intelligence*, pages 86–89, Berlin: Springer.

Bresnan, Joan. 1971. Sentence Stress and Syntactic Transformations. *Language* 47(2), 257–281.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Malden, MA: Blackwell Publisher Inc.

Bresnan, Joan and Mchombo, Sam A. 1987. Topic, Pronoun, and Agreement in Chicheŵa. *Language* 63(4), 741–782.

Brunetti, Lisa, Bott, Stefan, Costa, Joan and Vallduví, Enric. 2011. A Multilingual Annotated Corpus for the Study of Information Structure. In *Grammatik und Korpora 2009*, pages 305–327, Tübingen: Gunter Narr.

Büring, Daniel. 1997. The Great Scope Inversion Conspiracy. *Linguistics and Philosophy* 20(2), 175–194.

Büring, Daniel. 1999. Topic. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 142–165, Cambridge, UK: Cambridge University Press.

Büring, Daniel. 2003. On D-trees, Beans, and B-accents. *Linguistics and Philosophy* 26(5), 511–545.

Büring, Daniel. 2006. Focus Projection and Default Prominence. In Valéria Molnár and Susanne Winkler (eds.), *The Architecture of Focus*, pages 321–346, Berlin: Mouton de Gruyter.

Büring, Daniel. 2010. Towards a Typology of Focus Realization. In Malte Zimmermann abd Caroline Féry (ed.), *Information Structure*, pages 177–205, Oxford, UK: Oxford University Press.

Burnard, Lou. 2000. User Reference Guide for the British National Corpus. Technical Report, Oxford University Computing Services.

Byron, Donna K., Gegg-Harrison, Whitney and Lee, Sun-Hee. 2006. Resolving Zero Anaphors and Pronouns in Korean. *Traitement Automatique des Langues* 46, 91–114.

Calhoun, Sasha, Nissim, Malvina, Steedman, Mark and Brenier, Jason. 2005. A Framework for Annotating Information Structure in Discourse. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 45–52, Association for Computational Linguistics.

Callmeier, Ulrich. 2000. PET – a Platform for Experimentation with Efficient HPSG Processing Techniques. *Natural Language Engineering* 6(1), 99–107.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics* 22(2), 249–254.

Carreras, Xavier, Chao, Isaac, Padró, Lluıs and Padró, Muntsa. 2004. Freeling: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th LREC*.

Casielles-Suárez, Eugenia. 2003. On the Interaction between Syntactic and Information Structures in Spanish. *Bulletin of Hispanic Studies* 80(1), 1–20.

Casielles-Suárez, Eugenia. 2004. *The Syntax-Information Structure Interface: Evidence from Spanish and English*. New York and London: Routledge.

Cecchetto, Carlo. 1999. A Comparative Analysis of Left and Right Dislocation in Romance. *Studia Linguistica* 53(1), 40–67.

Chafe, Wallace L. 1976. Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View in Subject and Topic. In Charles N. Li (ed.), *Subject and Topic*, pages 25–55, New York, NY: Academic Press.

Chang, Suk-Jin. 2002. Information Unpackaging: A Constraint-based Grammar Approach to Topic-Focus Articulation. *Japanese/Korean Linguistics* 10, 451–464.

Chapman, Shirley. 1981. *Prominence in Paumarí*, volume 153 of *Archivo Linguistico*. Brasilia: Summer Institute of Linguistics.

Chen, Chen and Ng, Vincent. 2013. Chinese Zero Pronoun Resolution: Some Recent Advances. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1360–1365, Seattle, WA, USA: Association for Computational Linguistics.

Choe, Jae-Woong. 2002. Extended Focus: Korean Delimiter *man*. *Language Research* 38(4), 1131–1149.

Choi, Hye-Won. 1999. *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI Publications.

Choi, Incheol. 2012. Sentential Specifiers in the Korean Clause Structure. In Stefan Müller (ed.), *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar*, pages 75–85, Stanford, CA: CSLI Publications.

Chomsky, Noam. 1993. *Lectures on Government and Binding: The Pisa Lectures*. Berlin/New York: Walter de Gruyter.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, UK: Cambridge Univ Press.

Chung, Chan and Kim, Jong-Bok. 2009. Inverted English Concessive Constructions: A Construction-Based Approach. *Studies in Modern Grammar* 58, 39–58.

Chung, Chan, Kim, Jong-Bok and Sells, Peter. 2003. On the Role of Argument Structure in Focus Projections. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 39, pages 386–404, Chicago Linguistic Society.

Churng, Sarah. 2007. The Prosody of Topic and Focus: Explained away in Phases. *UW Working Papers in Linguistics* 26.

Cinque, Guglielmo. 1977. The Movement Nature of Left Dislocation. *Linguistic Inquiry* 8(2), 397–412.

Clech-Darbon, Anne, Rebuschi, Georges and Rialland, Annie. 1999. Are There Cleft Sentences in French. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 83–118, John Benjamins Publishing Company.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1), 37–46.

Comrie, Bernard. 1984. Some Formal Properties of Focus in Modern Eastern Armenian. *Annual of Armenian Linguistics* 5, 1–21.

Constant, Noah. 2012. English Rise-Fall-Rise: A Study in the Semantics and Pragmatics of Intonation. *Linguistics and Philosophy* 35(5), 407–442.

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

Copestake, Ann. 2007. Semantic Composition with (Robust) Minimal Recursion Semantics. In *Proceedings of the Workshop on Deep Linguistic Processing*, pages 73–80, Association for Computational Linguistics.

Copestake, Ann. 2009. Slacker Semantics: Why Superficiality, Dependency and Avoidance of Commitment can be the Right Way to Go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 1–9, Athens, Greece: Association for Computational Linguistics.

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

Croft, William. 2002. *Typology and Universals*. Cambridge, UK: Cambridge University Press.

Crowgey, Joshua. 2012. *The Syntactic Exponence of Negation: A Model for the LinGO Grammar Matrix*. M.A. Thesis, University of Washington.

Crowgey, Joshua and Bender, Emily M. 2011. Analyzing Interacting Phenomena: Word Order and Negation in Basque. In Stefan Müller (ed.), *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*, pages 46–59, Stanford: CSLI Publications.

Crysmann, Berthold. 2003. On the Efficient Implementation of German Verb Placement in HPSG. In *Proceedings of RANLP 2003*, pages 112–116, Borovets, Bulgaria.

Crysmann, Berthold. 2005a. Relative Clause Extraposition in German: An Efficient and Portable Implementation. *Research on Language & Computation* 3(1), 61–82.

Crysmann, Berthold. 2005b. Syncretism in German: a Unified Approach to Underspecification, Indeterminacy, and likeness of Case. In *Proceedings of the 12th International Conference on Head-driven Phrase Structure Grammar (HPSG)*, pages 91–107, Stanford: CSLI Publications.

De Kuthy, Kordula. 2000. *Discontinuous NPs in German – A Case Study of the Interaction of Syntax, Semantics and Pragmatics*. CSLI publications.

De Kuthy, Kordula and Meurers, Detmar. 2011. Integrating GIVENness into a structured meaning approach in HPSG. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 289–301, Stanford, CA: CSLI Publications.

De Saussure, Ferdinand. 1931. *Grundfragen der allgemeinen Sprachwissenschaft*. Berlin: Walter de Gruyter & Company.

Diesing, Molly. 1992. Bare Plural Subjects and the Derivation of Logical Representations. *Linguistic Inquiry* 23(3), 353–380.

Dipper, Stefanie, Goetze, Michael and Skopeteas, Stavros. 2007. *Information Structure in Cross-linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. Universitätsverlag Potsdam.

Drellishak, Scott. 2009. *Widespread but Not Universal: Improving the Typological Coverage of the Grammar Matrix*. Doctoral Dissertation, University of Washington.

Drellishak, Scott and Bender, Emily M. 2005. A Coordination Module for a Crosslinguistic Grammar Resource. In Stefan Müller (ed.), *The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, pages 108–128, Stanford, CA: CSLI Publications.

Drubig, Hans Bernhard. 2003. Toward a Typology of Focus and Focus Constructions. *Linguistics* 41(1), 1–50.

É. Kiss, Katalin. 1998. Identificational Focus versus Information Focus. *Language* 74(2), 245–273.

É. Kiss, Katalin. 1999. The English Cleft Construction as a Focus Phrase. *Boundaries of Morphology and Syntax* pages 217–229.

Emonds, Joseph. 1979. Appositive Relatives Have No Properties. *Linguistic Inquiry* 10(2), 211–243.

Emonds, Joseph. 2004. Unspecified Categories as the Key to Root Constructions. In David Adger, Cécile de Cat and Georges Tsoulas (eds.), *Peripheries: Syntactic Edges and their Effects*, pages 75–120, Dordrecht: Kluwer Academic Publishers.

Engdahl, Elisabet and Vallduví, Enric. 1996. Information Packaging in HPSG. *Edinburgh Working Papers in Cognitive Science* 12, 1–32.

Erteschik-Shir, Nomi. 1999. Focus Structure and Scope. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 119–150, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Erteschik-Shir, Nomi. 2007. *Information Structure: The Syntax-Discourse Interface*. Oxford, UK: Oxford University Press.

Eveleigh, Vaughan. 2010. *Efficient Operations on Semantic Dependency Structures*. MPhil Thesis, University of Cambridge.

Fabb, Nigel. 1990. The Difference between English Restrictive and Nonrestrictive Relative Clauses. *Journal of Linguistics* 26(1), 57–77.

Féry, Caroline and Ishihara, Shinichiro. 2009. The Phonology of Second Occurrence Focus. *Journal of Linguistics* 45(2), 285–313.

Féry, Caroline and Krifka, Manfred. 2008. Information Structure: Notional Distinctions, Ways of Expression pages 123–136.

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge, UK: Cambridge University Press.

Flickinger, Dan. 1987. *Lexical Rules in the Hierarchical Lexicon*. Doctoral Dissertation, Stanford University.

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

Fokkens, Antske. 2010. Documentation for the Grammar Matrix Word Order Library. Technical Report, Saarland University.

Frascarelli, Mare. 2000. *The Syntax-Phonology Interface in Focus and Topic Constructions in Italian*. Dordrecht/Boston: Kluwer Academic Publishers.

Frota, Sónia. 2000. *Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation*. New York, NY: Garland Publishing Inc.

Gell-Mann, Murray and Ruhlen, Merritt. 2011. The Origin and Evolution of Word Order. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 108, pages 17290–17295, National Acad Sciences.

Givón, Talmy. 1991. Isomorphism in the Grammatical Code: Cognitive and Biological Considerations. *Studies in Language* 15(1), 85–114.

Goodman, Michael Wayne. 2013. Generation of Machine-Readable Morphological Rules with Human Readable Input. *UW Working Papers in Linguistics* 30.

Gracheva, Varvara. 2013. *Markers of Contrast in Russian: A Corpus-based Study*. M.A. Thesis, University of Washington.

Grewendorf, Günther. 2001. Multiple *Wh*-Fronting. *Linguistic Inquiry* 32(1), 87–122.

Grishina, Elena. 2006. Spoken Russian in the Russian National Corpus (RNC). In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 121–124.

Grohmann, Kleanthes K. 2001. On Predication, Derivation and Anti-Locality. *ZAS Papers in Linguistics* 26, 87–112.

Gryllia, Styliani. 2009. *On the Nature of Preverbal Focus in Greek: a Theoretical and Experimental Approach*. Doctoral Dissertation, Leiden University.

Gundel, Jeanette K. 1977. Where Do Cleft Sentences Come From? *Language* 53(3), 543–559.

Gundel, Jeanette K. 1983. *The Role of Topic and Comment in Linguistic Theory*. New York, NY: Garland.

Gundel, Jeanette K. 1985. Shared Knowledge and Topicality. *Journal of Pragmatics* 9, 83–107.

Gundel, Jeanette K. 1988. Universals of Topic-Comment Structure. *Studies in Syntactic Typology* 17, 209–239.

Gundel, Jeanette K. 1999. On Different Kinds of Focus. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 293–305, Cambridge, UK: Cambridge University Press.

Gundel, Jeanette K. 2002. Information Structure and the Use of Cleft Sentences in English and Norwegian. In H. Hasselgrd, S. Johansson, B. Behrens and C. Fabricius-Hansen (eds.), *Information Structure in a Cross-Linguistic Perspective*, pages 113–128, Amsterdam: Rodopi.

Gundel, Jeanette K. 2003. Information Structure and Referential Givenness/Newness: How Much Belongs in the Grammar? In Stefan Müller (ed.), *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, pages 122–142, Stanford, CA: CSLI Publications.

Gunji, Takao. 1987. *Japanese Phrase Structure Grammar: a Unification-Based Approach*. Kluwer Academic Publishers Group.

Gunlogson, Christine. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Doctoral Dissertation, University of California at Santa Cruz.

Gussenhoven, Carlos. 1999. On the Limits of Focus Projection in English. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 43–55, Cambridge, UK: Cambridge University Press.

Gussenhoven, Carlos. 2007. Types of Focus in English. In Chungmin Lee, Matthew Gordon and Daniel Búring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, pages 83–100, Dordrecht: Kluwer Academic Publishers.

Gutierrez-Bravo, Rodrigo. 2006. *Structural Markedness and Syntactic Structure*. New York/London: Routledge.

Haegeman, Liliane. 2004. Topicalization, CLLD and the Left Periphery. In *ZAS Papers in Linguistics 35: Proceedings of the Dislocated Elements Workshop*, pages 157–192.

Haiman, John. 1978. Conditionals are Topics. *Language* 54(3), 564–589.

Haji-Abdolhosseini, Mohammad. 2003. A Constraint-Based Approach to Information Structure and Prosody Correspondence. In Stefan Müller (ed.), *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, pages 143–162, Stanford, CA: CSLI Publications.

Halliday, Michael Alexander Kirkwood. 1967. Notes on Transitivity and Theme in English: Part 2. *Journal of Linguistics* 3(2), 199–244.

Halliday, Michael Alexander Kirkwood. 1970. *A Course in Spoken English: Intonation*. Oxford: Oxford University Press.

Han, Chung-Hye and Hedberg, Nancy. 2008. Syntax and Semantics of *it*-clefts: a Tree Adjoining Grammar Analysis. *Journal of Semantics* 25(4), 345–380.

Han, Na-Rae. 2006. *Korean Zero Pronouns: Analysis and Resolution*. Doctoral Dissertation, University of Pennsylvania.

Hangyo, Masatsugu, Kawahara, Daisuke and Kurohashi, Sadao. 2013. Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 9241–934, Seattle, WA, USA: Association for Computational Linguistics.

Hartmann, Katharina and Zimmermann, Malte. 2007. Exhaustivity Marking in Hausa: A Reanalysis of the Particle *nee/cee*. In Enoch Oladé Aboh, Katharina Hartmann and Malte Zimmermann (eds.), *Focus Strategies in African Languages: The Interaction of Focus and Grammar in Niger-Congo and Afro-Asiatic*, pages 241–263, Berlin: Moutin de Gruyter.

Hasegawa, Akio. 2011. *The Semantics and Pragmatics of Japanese Focus Particles*. Doctoral Dissertation, State University of New York at Buffalo.

Hasegawa, Akio and Koenig, Jean-Pierre. 2011. Focus Particles, Secondary Meanings, and Lexical Resource Semantics: The Case of Japanese *shika*. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 81–101, Stanford, CA: CSLI Publications.

Hedberg, Nancy. 2006. Topic-Focus Controversies. In Susanne Winkler Valéria Molnár (ed.), *The Architecture of Focus*, pages 373–397, Berlin: Walter de Gruyter.

Hedberg, Nancy and Sosa, Juan M. 2007. The Prosody of Topic and Focus in Spontaneous English Dialogue. In Chungmin Lee, Matthew Gordon and Daniel Búring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, pages 101–120, Dordrecht: Kluwer Academic Publishers.

Hellan, Lars. 2005. Implementing Norwegian Reflexives in an HPSG Grammar. In *Proceedings of the 12th International Conference on Head-driven Phrase Structure Grammar (HPSG)*, pages 519–539, Stanford: CSLI Publications.

Heycock, Caroline. 1994. Focus Projection in Japanese. In *Proceedings of North East Linguistic Society*, pages 157–171.

Heycock, Caroline. 2007. Embedded Root Phenomena. In Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, pages 174–209, Wiley Online Library.

Hiraiwa, Ken and Ishihara, Shinichiro. 2002. Missing links: Cleft, Sluicing, and "*No da*" Construction in Japanese. *MIT Working Papers in Linguistics* 43, 35–54.

Hong, Jeen-Pyo and Cha, Jeong-Won. 2008. A New Korean Morphological Analyzer using the *Eojeol* Pattern Dictionary. In *Proceedings of the KCC08*.

Hooper, Joan and Thompson, Sandra. 1973. On the Applicability of Root Transformations. *Linguistic Inquiry* 4(4), 465–497.

Huang, C.-T. James. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Doctoral Dissertation, Massachusetts Institute of Technology.

Huang, C.-T. James. 1984. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry* 15(4), 531–574.

Huang, C.-T. James, Li, Y.-H. Audrey and Li, Yafei. 2009. *The Syntax of Chinese*. Cambridge, UK: Cambridge University Press.

Iatridou, Sabine. 1991. *Topics in Conditionals*. Doctoral Dissertation, Massachusetts Institute of Technology.

Iatridou, Sabine. 2000. The Grammatical Ingredients of Counterfactuality. *Linguistic Inquiry* 31(2), 231–270.

Ishihara, Shinichiro. 2001. Stress, Focus, and Scrambling in Japanese. *MIT Working Papers in Linguistics* 39, 142–175.

İşsever, Selçuk. 2003. Information Structure in Turkish: the Word Order-Prosody Interface. *Lingua* 113(11), 1025–1053.

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar.*. Cambridge, MA: The MIT Press.

Jacobs, Neil G. 2005. *Yiddish: A Linguistic Introduction*. New York: Cambridge University Press.

Jiang, Zixin. 1991. *Some Aspects of the Syntax of Topic and Subject in Chinese*. Doctoral Dissertation, University of Chicago.

Johansson, Mats. 2001. Clefts in Contrast: a Contrastive Study of it Clefts and wh Clefts in English and Swedish Texts and Translations. *Linguistics* 39(3), 547–582.

Joshi, Aravind K. and Schabes, Yves. 1997. Tree-Adjoining Grammars. In Grzegorz Rozenberg and Arto Salomaa (eds.), *Handbook of Formal Languages*, pages 69–123, Berlin: Springer.

Jun, Sun-Ah, Kim, Hee-Sun, Lee, Hyuck-Joon and Kim, Jong-Bok. 2007. An Experimental Study on the Effect of Argument Structure on VP Focus. *UCLA Working Papers in Phonetics* 105, 66–84.

Jun, Sun-Ah and Lee, Hyuck-Joon. 1998. The Phonetics and Phonology of Korean Prosody in Korean. In *International Conference on Spoken Language Processing*, pages 1295–1298, Sydney, Australia.

Kadmon, Nirit. 2001. *Formal Pragmatics*. Malden, MA: Blackwell Publisher Inc.

Kaiser, Elsi. 2009. Investigating Effects of Structural and Information-structural Factors on Pronoun Resolution. In Malte Zimmermann and Caroline Féry (eds.), *Information Structure: Theoretical, Typological, and Experimental Perspectives*, pages 332–354, Oxford, UK: Oxford University Press.

Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic*. London: Kluwer Academic Publishers.

Kiefer, Ferenc. 1967. *On Emphasis and Word Order in Hungarian*. Bloomington: Indiana University Press.

Kihm, Alain. 1999. Focus in Wolof. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 245–273, John Benjamins Publishing Company.

Kim, Jieun. 2012a. How is 'Contrast' Imposed on *-Nun*? *Language and Information* 16(1), 1–24.

Kim, Jong-Bok. 2007. Syntax and Semantics of English It-Cleft Constructions: A Constraint-Based Analysis. *Studies in Modern Grammar* 48, 217–235.

Kim, Jong-Bok. 2011a. Floating Numeral Classifiers in Korean: A Thematic-Structure Perspective. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 302–313, Stanford, CA: CSLI Publications.

Kim, Jong-Bok. 2012b. On the Syntax of the It-Cleft Construction: A Construction-based Perspective. *Linguistic Research* 29(1), 45–68.

Kim, Jong-Bok and Park, Byung-Soo. 2000. Grammatical Interfaces in Korean Relatives. In Ronnie Cann, Claire Grover and Philip Miller (eds.), *Grammatical Interfaces in HPSG*, pages 153–168, Stanford, CA: CSLI Publications.

Kim, Jong-Bok and Sells, Peter. 2007. Two Types of Multiple Nominative Construction: A Constructional Approach. In Stefan Müller (ed.), *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 364–372, Stanford, CA: CSLI Publications.

Kim, Jong-Bok and Sells, Peter. 2008. *English Syntax: An Introduction*. Stanford, CA: CSLI publications.

Kim, Jong-Bok and Yang, Jaehyung. 2004. Projections from Morphology to Syntax in the Korean Resource Grammar: Implementing Typed Feature Structures. *Lecture Notes in Computer Science* 2945, 13–24.

Kim, Jong-Bok and Yang, Jaehyung. 2009. Processing Three Types of Korean Cleft Constructions in a Typed Feature Structure Grammar. *Korean Journal of Cognitive Science* 20(1), 1–28.

Kim, Jong-Bok, Yang, Jaehyung, Song, Sanghoun and Bond, Francis. 2011. Deep Processing of Korean and the Development of the Korean Resource Grammar. *Linguistic Research* 28(3), 635–672.

Kim, Taeho. 2011b. An Empirical Study of Postposing Constructions in Korean. *Linguistic Research* 28(1), 223–238.

King, Tracy Holloway. 1995. *Configuring Topic and Focus in Russian*. Stanford, CA: CSLI publications.

King, Tracy Holloway. 1997. Focus Domains and Information-Structure. In Butt Miriam and Tracy Holloway King (eds.), *Proceedings of the LFG97 Conference*, University of California, San Diego.

King, Tracy Holloway and Zaenen, Annie. 2004. F-structures, Information Structure, and Discourse Structure. In Butt Miriam and Tracy Holloway King (eds.), *Proceedings of the LFG04 Conference*, University of Canterbury, New Zealand.

Kizu, Mika. 2005. *Cleft Constructions in Japanese Syntax*. New York, NY: Palgrave Macmillan.

Klein, Ewan. 2000. Prosodic Constituency in HPSG. In Ronnie Cann, Claire Grover and Philip Miller (eds.), *Grammatical Interfaces in HPSG*, pages 169–200, Stanford, CA: CSLI Publications.

Ko, Kil Soo. 2008. Korean Postpositions as Weak Syntactic Heads. In Stefan Müller (ed.), *Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar*, pages 131–151, Stanford, CA: CSLI Publications.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT summit*, Phuket, Thailand.

Komagata, Nobo N. 1999. *A Computational Analysis of Information Structure Using Parallel Expository Texts in English and Japanese*. Doctoral Dissertation, University of Pennsylvania.

Kong, Fang and Ng, Hwee Tou. 2013. Exploiting Zero Pronouns to Improve Chinese Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 278–288, Seattle, WA, USA: Association for Computational Linguistics.

Krifka, Manfred. 2008. Basic Notions of Information Structure. *Acta Linguistica Hungarica* 55(3), 243–276.

Kügler, Frank, Skopeteas, Stavros and Verhoeven, Elisabeth. 2007. Encoding Information Structure in Yucatec Maya: on the Interplay of Prosody and Syntax. *Interdisciplinary Studies on Information Structure* 8, 187–208.

Kuhn, Jonas. 1996. An Underspecified HPSG Representation for Information Structure. In *Proceedings of the 16th conference on Computational Linguistics*, volume 2, pages 670–675, Association for Computational Linguistics.

Kuno, Susumu. 1973. *The Structure of the Japanese Language*. Cambridge, MA: The MIT Press.

Kuno, Susumu. 1976. Subject, Theme and Speaker's Empathy: A Reexamination of Relativization Phenomena. In Charles N. Li (ed.), *Subject and Topic*, pages 417–444, New York, NY: Academic Press.

Kuroda, S.-Y. 1972. The Categorical and the Thetic Judgment: Evidence from Japanese Syntax. *Foundations of Language* 9(2), 153–185.

Lambrecht, Knud. 1986. *Topic, Focus, and the Grammar of Spoken French*. Doctoral Dissertation, University of California, Berkeley.

Lambrecht, Knud. 1996. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge, UK: Cambridge University Press.

Lambrecht, Knud. 2001. A Framework for the Analysis of Cleft Constructions. *Linguistics* 39(3), 463–516.

Law, Ann. 2003. Right Dislocation in Cantonese as a Focus-Marking Device. *UCL Working Papers in Linguistics* 15, 243–275.

Lecarme, Jacqueline. 1999. Focus in Somali. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 1–22, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Lee, Chungmin. 2003b. Contrastive Topic and Proposition Structure. In Anna Maria Di Sciullo (ed.), *Asymmetry in Grammar: Syntax and Semantics*, pages 345–372, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Lee, Youngjoo. 2004. *The Syntax and Semantics of Focus Particles*. Doctoral Dissertation, Massachusetts Institute of Technology.

Li, Charles N. and Thompson, Sandra. 1976. Subject and Topic: A New Typology of Language. In Charles N. Li (ed.), *Subject and Topic*, pages 457–490, New York: Academic Press.

Li, Kening. 2009. *The Information Structure of Mandarin Chinese: Syntax and Prosody*. Doctoral Dissertation, University of Washington.

Lim, Dong-Hoon. 2012. Korean Particle 'un/nun' and Their Syntagmatic, Paradigmatic Relations [In Korean]. *Korean Linguistics* 64, 217–271.

Maki, Hideki, Kaiser, Lizanne and Ochi, Masao. 1999. Embedded Topicalization in English and Japanese. *Lingua* 109(1), 1–14.

Man, Fung Suet. 2007. *TOPIC and FOCUS in Cantonese: An OT-LFG Account*. M.A. Thesis, University of Hong Kong.

Marimon, Montserrat. 2012. The Spanish DELPH-IN Grammar. *Language Resources and Evaluation* 47(2), 371–397.

Matsui, Tomoko. 1999. Approaches to Japanese Zero Pronouns: Centering and Relevance. In *Proceedings of the Workshop on the Relation of Discourse/Dialogue Structure and Reference*, pages 11–20.

Megerdoomian, Karine. 2011. Focus and the Auxiliary in Eastern Armenian, talk presented at the 37th Annual Meeting of the Berkeley Linguistics Society (BLS), Special session on Languages of the Caucasus.

Mereu, Lunella. 2009. Universals of Information Structure. In Lunella Mereu (ed.), *Information Structure and its Interfaces*, pages 75–104, Berlin/New York: Mouton de Gruyter.

Mitkov, Ruslan. 1999. Multilingual Anaphora Resolution. *Machine Translation* 14(3), 281–299.

Mitkov, Ruslan, Choi, Sung-Kwon and Sharp, Randall. 1995. Anaphora Resolution in Machine Translation. In *Proceedings of the 6th International conference on Theoretical and Methodological issues in Machine Translation*.

Miyao, Yusuke and Tsujii, Jun'ichi. 2008. Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics* 34(1), 35–80.

Molnár, Valéria. 2002. Contrast – from a Contrastive Perspective. In H. Hasselgrd, S. Johansson, B. Behrens and C. Fabricius-Hansen (eds.), *Information Structure in a Cross-Linguistic Perspective*, pages 147–162, Amsterdam, Netherland: Rodopi.

Montgomery-Anderson, Brad. 2008. *A Reference Grammar of Oklahoma Cherokee*. Ann Arbor, MI: ProQuest LLC.

Nagaya, Naonori. 2007. Information Structure and Constituent Order in Tagalog. *Language and Linguistics* 8, 343–372.

Nakaiwa, Hiromi and Shirai, Satoshi. 1996. Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 812–817, Association for Computational Linguistics.

Nakanishi, Kimiko. 2007. Prosody and Information Structure in Japanese: A Case Study of topic Marker wa. In Chungmin Lee, Matthew Gordon and Daniel Büring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, pages 177–193, Dordrecht: Kluwer Academic Publishers.

Neeleman, Ad and Titov, Elena. 2009. Focus, Contrast, and Stress in Russian. *Linguistic Inquiry* 40(3), 514–524.

Nelson, Gerald, Wallis, Sean and Aarts, Bas. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Philadelphia: John Benjamins Publishing Co.

Nguyen, Hoai Thu Ba. 2006. *Contrastive Topic in Vietnamese: with Reference to Korean*. Doctoral Dissertation, Seoul National University.

Nichols, Eric, Bond, Francis, Appling, Darren Scott and Matsumoto, Yuji. 2010. Paraphrasing Training Data for Statistical Machine Translation. *Journal of Natural Language Processing* 17(3), 101–122.

Nichols, Johanna. 2011. *Ingush Grammar*. Berkeley, CA: University of California Press.

Ning, Chunyan. 1993. *The Overt Syntax of Relativization and Topicalization in Chinese*. Doctoral Dissertation, University of California, Irvine.

Och, Franz Josef and Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1), 19–51.

Oepen, Stephan. 2001. [incr tsdb()] — Competence and Performance Laboratory. User Manual. Technical Report, Computational Linguistics, Saarland University.

Oepen, Stephan, Flickinger, Dan, Toutanova, Kristina and Manning, Christoper D. 2004. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. *Research on Language & Computation* 2(4), 575–596.

Oepen, Stephan, Velldal, Erik, Lønning, Jan T., Meurer, Paul, Rosén, Victoria and Flickinger, Dan. 2007. Towards Hybrid Quality-Oriented Machine Translation – On Linguistics and Probabilities in MT. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde, Sweden.

Ohtani, Akira and Matsumoto, Yuji. 2004. Japanese Subjects and Information Structure: A Constraint-based Approach. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 93–104, Tokyo, Japan.

Ortiz de Urbina, Jon. 1999. Focus in Basque. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 311–333, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Osenova, Petya. 2011. Localizing a Core HPSG-based Grammar for Bulgarian. In Hanna Hedeland, Thomas Schmidt and Kai Worner (eds.), *Multilingual Resources and Multilingual Applications, Proceedings of German Society for Computational Linguistics and Language Technology (GSCL)*, pages 175–180, Hamburg.

Oshima, David Y. 2008. Morphological *vs*. Phonological Contrastive Topic Marking. In *Proceedings of Chicago Linguistic Society (CLS) 41*, pages 371–383.

Oshima, David Y. 2009. On the So-Called Thematic Use of *Wa*: Reconsideration and Reconciliation. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, pages 405–414, City University of Hong Kong, Hong Kong.

Ouhalla, Jamal. 1999. Focus and Arabic Clefts. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 335–359, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Paggio, Patrizia. 1996. *The Treatment of Information Structure in Machine Translation*. Doctoral Dissertation, University of Copenhagen.

Paggio, Patrizia. 2009. The Information Structure of Danish Grammar Constructions. *Nordic Journal of Linguistics* 32(01), 137–164.

Partee, Barbara H. 1991. Topic, Focus and Quantification. *Cornell Working Papers in Linguistics* 10, 159–187.

Partee, Barbara H. 1999. Focus, Quantification, and Semantics-Pragmatics Issues. In Peter Bosch and Rob van der Sandt (eds.), *Focus: Linguistic, Cognitive, and Computational Perspectives*, pages 213–231, Cambridge, UK: Cambridge University Press.

Paul, Waltraud and Whitman, John. 2008. *Shi ... de* Focus Clefts in Mandarin Chinese. *The Linguistic Review* 25(3-4), 413–451.

Pedersen, Ted. 2008. Empiricism is Not a Matter of Faith. *Computational Linguistics* 34(3), 465–470.

Petronio, Karen. 1993. *Clause Structure in American Sign Language*. Doctoral Dissertation, University of Washington.

Pollard, Carl and Sag, Ivan A. 1987. *Information-Based Syntax and Semantics*. Stanford, CA: CSLI Publications.

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL: The University of Chicago Press.

Portner, Paul and Yabushita, Katsuhiko. 1998. The Semantics and Pragmatics of Topic Phrases. *Linguistics and Philosophy* 21(2), 117–157.

Poulson, Laurie. 2011. Meta-modeling of Tense and Aspect in a Cross-linguistic Grammar Engineering Platform. *UW Working Papers in Linguistics* 28.

Pozen, Zinaida. 2013. *Using Lexical and Compositional Semantics to Improve HPSG Parse Selection*. M.S. Thesis, University of Washington.

Press, Ian J. 1986. *A Grammar of Modern Breton*. Berlin/New York/Amsterdam: Mouton de Gruyter.

Prince, Ellen F. 1984. Topicalization and Left-Dislocation: A Functional Analysis. *Annals of the New York Academy of Sciences* 433(1), 213–225.

Ramsay, Violetta. 1987. The Functional Distribution of Preposed and Postposed 'if' and 'when' Clauses in Written Discourse. In Russell S. Tomlin (ed.), *Coherence and Grounding in Discourse*, pages 383–408, Amsterdam: John Benjamins.

Rebuschi, Georges and Tuller, Laurice. 1999. The Grammar of Focus: An Introduction. In Georges Rebuschi and Laurice Tuller (eds.), *The Grammar of Focus*, pages 1–22, John Benjamins Publishing Company.

Reinhart, Tanya. 1981. Pragmatics and Linguistics: An Analysis of Sentence Topics. *Philosophica* 27(1), 53–94.

Rivero, María-Luisa. 1980. On Left-dislocation and Topicalization in Spanish. *Linguistic Inquiry* 11(2), 363–393.

Rizzi, Luigi. 1997. The Fine Structure of the Left Periphery. In Liliane Haegeman (ed.), *Elements of Grammar: Handbook in Generative Syntax*, pages 281–337, Dordrecht: Kluwer Academic Publishers.

Roberts, Craige. 2011. Topics. In Claudia Maienborn, Klaus von Heusinger and Paul Portner (eds.), *Semantics: An International Handbook of Natural Language Meaning*, volume 2, pages 1908–1934, Berlin, New York: Mouton de Gruyter.

Rochemont, Michael S. 1986. *Focus in Generative Grammar*. Amsterdam: John Benjamins Publishing Company.

Rodionova, Elena V. 2001. *Word Order and Information Structure in Russian Syntax*. M.A. Thesis, University of North Dakota.

Roh, Ji-Eun and Lee, Jong-Hyeok. 2003. An Empirical Study for Generating Zero Pronoun in Korean based on Cost-based Centering Model. In *Proceedings of Australasian Language Technology Association*, pages 90–97.

Rooth, Mats. 1992. A Theory of Focus Interpretation. *Natural Language Semantics* 1(1), 75–116.

Sag, Ivan A., Wasow, Thomas, and Bender, Emily M. 2003. *Syntactic Theory: A Formal Introduction*. Stanford, CA: CSLI Publications.

Saleem, Safiyyah. 2010. *Argument Optionality: A New Library for the Grammar Matrix Customization System*. M.A. Thesis, University of Washington.

Saleem, Safiyyah and Bender, Emily M. 2010. Argument Optionality in the LinGO Grammar Matrix. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1068–1076, Beijing, China: Coling 2010 Organizing Committee.

Sato, Yo and Tam, Wai Lok. 2012. Ellipsis of Case-markers and Information Structure in Japanese. In Stefan Müller (ed.), *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar*, pages 442–452, Stanford, CA: CSLI Publications.

Schachter, Paul. 1973. Focus and Relativization. *Language* 49(1), 19–46.

Schafer, Amy, Carter, Juli, Jr, Charles Clifton and Frazier, Lyn. 1996. Focus in Relative Clause Construal. *Language and Cognitive Processes* 11(1/2), 135–163.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processingq*, Manchester, UK.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Schneider, Cynthia. 2009. Information Structure in Abma. *Oceanic Linguistics* 48(1), 1–35.

Searle, John R. 1976. A Classification of Illocutionary Acts. *Language in Society* 5(1), 1–23.

Selkirk, Elisabeth O'Brian. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: The MIT Press.

Selkirk, Elisabeth O'Brian. 1995. Sentence Prosody: Intonation, Stress, and Phrasing. In John A. Goldsmith (ed.), *The Handbook of Phonological Theory*, Cambridge: Blackwell Publishers.

Siegel, Melanie. 1999. The Syntactic Processing of Particles in Japanese Spoken Language. In Jhing-Fa Wang and Chung-Hsien Wu (eds.), *Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation*, pages 313–320.

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

Skopeteas, Stavros and Fanselow, Gisbert. 2010. Focus in Georgian and the Expression of Contrast. *Lingua* 120(6), 1370–1391.

Slayden, Glenn C. 2012. *Array TFS Storage for Unification Grammars*. M.S. Thesis, University of Washington.

Sohn, Ho-Min. 2001. *The Korean Language*. Cambridge, UK: Cambridge University Press.

Song, Sanghoun and Bender, Emily M. 2011. Using Information Structure to Improve Transfer-based MT. In Stefan Müller (ed.), *Proceedings of the 18th International Conference on Head-Driven Phrase Structure Grammar*, pages 348–368, Stanford, CA: CSLI Publications.

Song, Sanghoun and Bender, Emily M. 2012. Individual Constraints for Information Structure. In Stefan Müller (ed.), *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar*, pages 329–347, Stanford, CA: CSLI Publications.

Song, Sanghoun and Bond, Francis. 2009. Checking the Feasibility of the Sejong Bilingual Corpus for Statistical Machine Translation. *Language and Linguistics* 46, 53–84.

Song, Sanghoun, Kim, Jong-Bok, Bond, Francis and Yang, Jaehyung. 2010. Development of the Korean Resource Grammar: Towards Grammar Customization. In *Proceedings of the 8th Workshop on Asian Language Resources*, Beijing, China.

Steedman, Mark. 2000. Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry* 31(4), 649–689.

Steedman, Mark. 2001. *The Syntactic Process*. Cambridge, MA: The MIT press.

Strawson, Peter F. 1964. Identifying Reference and Truth-Values. *Theoria* 30(2), 96–118.

Sturgeon, Anne. 2010. The Discourse Function of Left Dislocation in Czech. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, volume 31.

Szendrői, Kriszta. 1999. A Stress-Driven Approach to the Syntax of Focus. *UCL Working Papers in Linguistics* 11, 545–573.

Szendrői, Kriszta. 2001. *Focus and the Syntax-Phonology Interface*. Doctoral Dissertation, University College London.

Tamrazian, Armine. 1991. Focus and Wh-movement in Armenian. *University College London Working Papers in Linguistics* 3, 101–121.

Tamrazian, Armine. 1994. *The Syntax of Armenian: Chains and the Auxiliary*. Doctoral Dissertation, University College London.

Taylor, Heather L. 2007. Movement from IF-clause Adjuncts. *University of Maryland Working Papers in Linguistics* 15, 192–206.

Traat, Maarika and Bos, Johan. 2004. Unificational Combinatory Categorial Grammar: Combining Information Structure and Discourse Representations. In *Proceedings of the 20th International Conference on Computational Linguistics*, Association for Computational Linguistics.

Tragut, Jasmine. 2009. *Armenian: Modern Eastern Armenian*. Amsterdam: John Benjamins Publishing Company.

Ueyama, Motoko and Jun, Sun-Ah. 1998. Focus Realization in Japanese English and Korean English Intonation. *Japanese/Korean Linguistics* 7, 629–645.

Valentine, Randy. 2001. *Nishnaabemwin Reference Grammar*. University of Toronto Press.

Vallduví, Enric. 1990. *The Informational Component*. Doctoral Dissertation, University of Pennsylvania.

Vallduví, Enric. 1992. Focus Constructions in Catalan. In Christiane Laeufer and Terrell A. Morgan (eds.), *Theoretical Analyses in Romance Linguistics*, pages 457–479, Amsterdam/Philadelphia: John Benjamins.

Vallduví, Enric and Vilkuna, Maria. 1998. On Rheme and Kontrast. *Syntax and Semantics* 29, 79–108.

van Valin, Robert D. 2005. *Exploring the Syntax-Semantics Interface*. Cambridge, UK: Cambridge Univ Press.

Velleman, Dan, Beaver, David, Destruel, Emilie, Bumford, Dylan, Onea, Edgar and Coppock, Liz. 2012. It-clefts are IT (Inquiry Terminating) Constructions. In *Proceedings of Semantics and Linguistic Theory 22*, pages 441–460.

Vermeulen, Reiko. 2009. On the Syntactic Typology of Topic Marking: A Comparative Study of Japanese and Korean. *UCL Working Papers in Linguistics* 21, 335–363.

von Fintel, Kai. 2004. Would You Believe It? The King of France is Back! (Presuppositions and Truth-value Intuitions). In Marga Reimer and Anne Bezuidenhout (eds.), *Descriptions and Beyond*, pages 315–341, Oxford, UK: Oxford University Press.

von Prince, Kilu. 2012. Predication and Information Structure in Mandarin Chinese. *Journal of East Asian Linguistics* 21(4), 329–366.

Webelhuth, Gert. 2007. Complex Topic-Comment Structures in HPSG. In Stefan Müller (ed.), *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 306–322, Stanford, CA: CSLI Publications.

Wee, Hae-Kyung. 2001. *Sentential Logic, Discourse and Pragmatics of Topic and Focus*. Doctoral Dissertation, Indiana University.

Wilcock, Graham. 2005. Information Structure and Minimal Recursion Semantics. *Inquiries into Words, Constraints and Contexts: Festschrift for Kimmo Koskenniemi on his 60th Birthday* pages 268–277.

Yang, Charles D. 2002. *Knowledge and Learning in Natural Language*. Oxford, UK: Oxford University Press.

Yatabe, Shûichi. 1999. Particle Ellipsis and Focus Projection in Japanese. *Language, Information, Text* 6, 79–104.

Yeh, Ching-Long and Chen, Yi-Chun. 2004. Zero Anaphora Resolution in Chinese with Shallow Parsing. *Journal of Chinese Language and Computing* .

Yoo, Hyun-kyung, An, Yeri and Yang, Su-hyang. 2007. The Study on the Principles of Selecting Korean Particle 'Ka' and 'Nun' Using Korean-English Parallel Corpus [In Korean]. *Language and Information* 11(1), 1–23.

Yoshimoto, Kei. 2000. A Bistratal Approach to the Prosody-Syntax Interface in Japanese. In Ronnie Cann, Claire Grover and Philip Miller (eds.), *Grammatical Interfaces in HPSG*, pages 267–282, Stanford, CA: CSLI Publications.

Yoshimoto, Kei, Kobayashi, Masahiro, Nakamura, Hiroaki and Mori, Yoshiki. 2006. Processing of Information Structure and Floating Quantifiers in Japanese. *Lecture Notes in Computer Science* 4012, 103–110.

Yu, Kun, Miyao, Yusuke, Wang, Xiangli, Matsuzaki, Takuya and Tsujii, Junichi. 2010. Semi-automatically Developing Chinese HPSG Grammar from the Penn Chinese Treebank for Deep Parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1417–1425, Association for Computational Linguistics.

Zagona, Karen. 2002. *The Syntax of Spanish*. Cambridge, UK: Cambridge University Press.

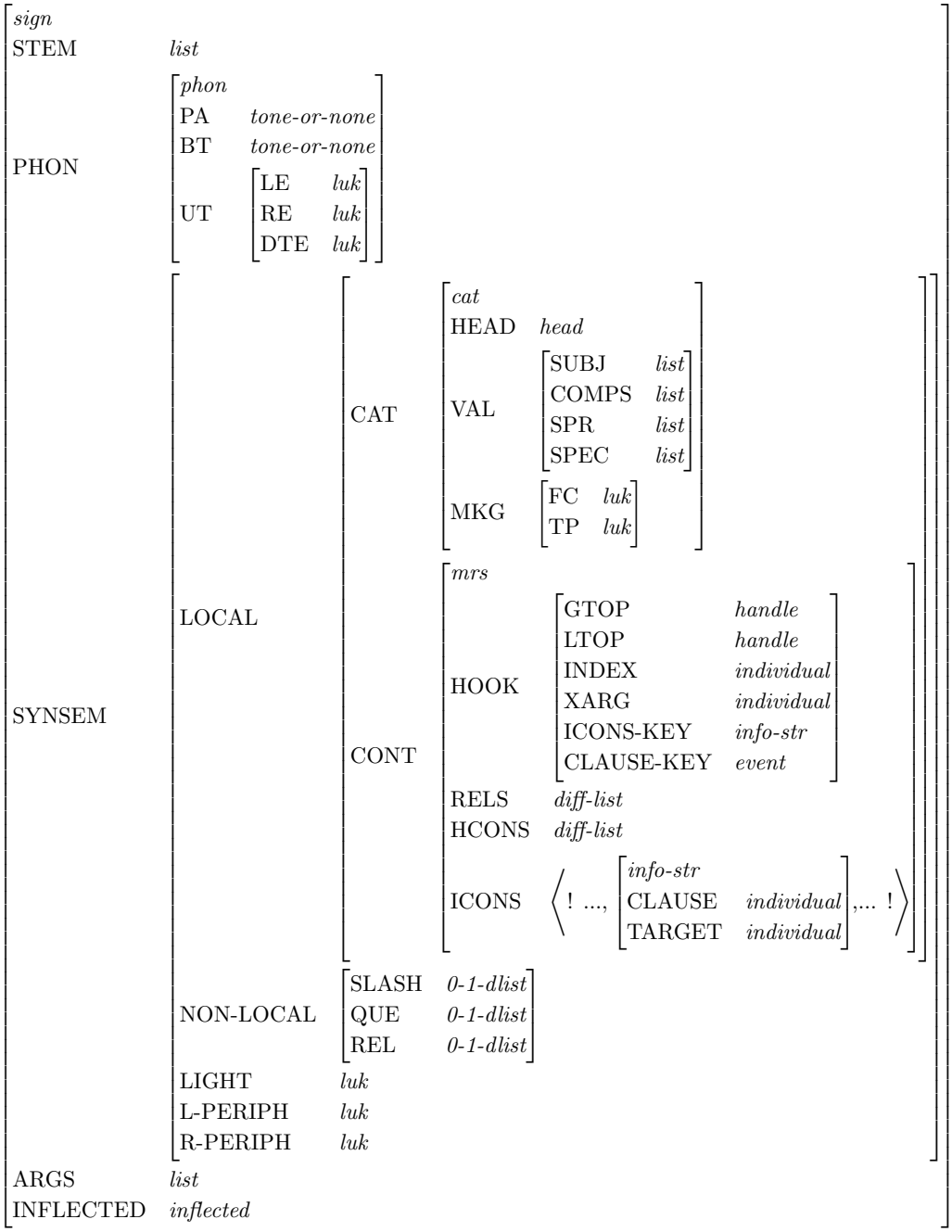
Zeevat, Henk. 1987. Combining Categorial Grammar and Unification. In Uwe Reyle and Christian Rohrer (eds.), *Natural Language Parsing and Linguistic Theories*, pages 202–229, Dordrecht: D. Reidel Publishing Company.

Zhao, Shanheng and Ng, Hwee Tou. 2007. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech.

Zubizarreta, Maria Luisa. 1998. *Prosody, Focus, and Word Order*. Cambridge, MA: The MIT Press.

Appendix A

AVM OF *SIGN*



Appendix B

CATALOGUE OF LANGUAGES

name	ISO 639-3	language family
Abma	app	Austronesian/Oceanic
Akan	aka	Niger-Congo/Kwa
Armenian	hye	Indo-European
Basque	eus	unknown
Bosnian Croatian Serbian	hbs	Indo-European/Slavic
Breton	bre	Indo-European/Celtic
Buli	bwu	Niger-Congo/Gur
Cantonese	yue	Sino-Tibetan
Catalan	cat	Indo-European/Romance
Cherokee	chr	Iroquoian
Chichewa	nya	Niger-Congo/Bantu
Czech	ces	Indo-European/Slavic
Danish	dan	Indo-European/Germanic
Ditammari	tbz	Niger-Congo/Gur
(Northern) Frisian	frr	Indo-European/Germanic
French	fra	Indo-European/Romance
Georgian	kat	Kartvelian
German	ger	Indo-European/Germanic
Greek	ell	Indo-European/Hellenic
Hausa	hau	Afro-Asiatic/Chadic
Hungarian	hun	Uralic
Ilonggo	hil	Austronesian/Philippine
Ingush	inh	Ingush
Ishkashimi	isk	Indo-European/Iranian
Italian	ita	Indo-European/Romance
Japanese	jpn	unknown
Korean	kor	unknown
Lakota	lkt	Siouan
Mandarin Chinese	cmn	Sino-Tibetan/Chinese
Miyako	mvi	Japonic
Moroccan Arabic	ary	Afro-Asiatic/Semitic

name	ISO 639-3	language family
Navajo	nav	Athabaskan
Ngizim	ngi	Afro-Asiatic/Chadic
Nishnaabemwin	ojg/otw	Algic
Norwegian	nor	Indo-European/Germanic
Paumarí	pad	Arauan
Portuguese	por	Indo-European/Romance
Rendile	rel	Afro-Asiatic/Cushitic
Russian	rus	Indo-European/Slavic
Spanish	spa	Indo-European/Romance
Standard Arabic	arb	Afro-Asiatic/Semitic
Tangale	tan	Afro-Asiatic/Chadic
Turkish	tur	Turkic
Vietnamese	vie	Austro-Asiatic/Vietic
Wolof	wol	Niger-Congo/Senegambian
Yiddish	ydd	Indo-European/Germanic

Appendix C

CATALOGUE OF EXAMPLES

language	chapter	(number, page)
Abma		
	4	(11, p. 71)
Akan		
	4	(10, p. 70)
	11	(13, p. 236)
Armenian		
	4	(27, p. 80)
	11	(5, p. 247)
Basque		
	4	(18, p. 74) (24, p. 78) (25, p. 79)
	10	(29, p. 214) (31, p. 214)
Bosnian Croatian Serbian		
	4	(23, p. 78) (30, p. 82)
	5	(13, p. 96)
Breton		
	4	(16, p. 73)
Buli		
	4	(12, p. 71)
Cantonese		
	4	(8, p. 69) (36, p. 84)
	5	(9, p. 94)
Cherokee		
	4	(9, p. 69)
Czech		
	12	(53, p. 272)
Danish		
	3	(37, p. 53)
	4	(31, p. 82)
	8	(4, p. 148)
	11	(9, p. 234)

language	chapter	(number, page)
Ditammari		
	4	(13, p. 71)
English		
	1	(1, p. 2) (3, p. 7)
	3	(4, p. 10) (5, p. 10) (6, p. 34) (7, p. 34)
		(10, p. 36) (11, p. 37) (12, p. 38) (14, p. 40)
		(15, p. 42) (16, p. 42) (22, p. 45) (23, p. 46)
		(25, p. 47) (34, p. 52) (39, p. 54) (45, p. 56)
		(48, p. 58) (49, p. 58) (51, p. 59)
	4	(2, p. 63) (3, p. 64) (4, p. 65) (14, p. 71)
		(21, p. 76) (34, p. 83)
	5	(1, p. 89) (6, p. 92) (7, p. 93) (8, p. 93)
	7	(6, p. 134) (10, p. 135) (11, p. 135) (12, p. 135)
		(6, p. 134) (13, p. 137) (15, p. 138) (18, p. 141)
		(19, p. 141) (20, p. 142)
	8	(3, p. 148)
	9	(4, p. 172) (5, p. 173) (8, p. 177)
		(9, p. 180) (30, p. 192)
	10	(7, p. 201) (10, p. 202) (15, p. 204) (17, p. 205)
		(18, p. 206) (22, p. 208) (23, p. 208) (35, p. 218)
	11	(1, p. 229) (14, p. 238) (15, p. 238) (16, p. 239)
	12	(3, p. 246) (8, p. 249) (9, p. 249) (11, p. 250)
		(24, p. 259) (41, p. 266) (29, p. 261) (33, p. 262)
		(37, p. 264) (41, p. 266) (45, p. 268)
		(50, p. 271) (52, p. 271)
	13	(2, p. 277) (7, p. 280) (8, p. 281)
		(9, p. 281) (10, p. 281) (11, p. 281)
		(12, p. 282) (13, p. 282) (14, p. 283)
		(16, p. 285) (20, p. 286) (21, p. 287)
	15	(2, p. 316) (3, p. 316) (5, p. 317) (7, p. 321)
French		
	4	(36, p. 84)
	12	(39, p. 265)
Georgian		
	4	(37, p. 85)
German		
	3	(17, p. 42) (20, p. 44)
	12	(53, p. 272)

language	chapter	(number, page)
Greek		
	3	(43, p. 56)
	5	(14, p. 97)
	9	(3, p. 170)
Hausa		
	4	(20, p. 76)
	5	(15, p. 98)
Hungarian		
	3	(4, p. 31)
	4	(26, p. 79)
	12	(26, p. 260)
Ilonggo		
	5	(4, p. 91)
	11	(13, p. 236)
Ingush		
	4	(19, p. 75) (38, p. 87)
	5	(5, p. 92)
Italian		
	3	(42, p. 55)
	10	(3, p. 198)
	11	(1, p. 229)
Japanese		
	1	(2, p. 3)
	3	(18, p. 43) (27, p. 47) (35, p. 52) (36, p. 53)
	4	(7, p. 68) (17, p. 73) (32, p. 82)
	9	(19, p. 188)
	10	(40, p. 221)
	11	(11, p. 235) (20, p. 241)
	12	(16, p. 253) (17, p. 254) (18, p. 254) (19, p. 255)
	15	(4, p. 317) (5, p. 317)
Korean		
	3	(9, p. 35) (21, p. 44) (24, p. 46) (28, p. 50) (40, p. 54) (41, p. 55) (44, p. 56) (47, p. 57) (50, p. 59)
	4	(6, p. 68) (33, p. 83) (35, p. 83)
	5	(2, p. 90) (3, p. 90)
	6	(6, p. 126)
	7	(2, p. 130) (3, p. 131) (4, p. 131) (5, p. 132) (8, p. 134) (9, p. 135) (8, p. 134) (14, p. 137) (16, p. 138)
	8	(9, p. 155) (11, p. 156) (16, p. 163)

language	chapter	(number, page)
Korean		
	9	(13, p. 182) (18, p. 188)
	10	(40, p. 221)
	11	(5, p. 232)
	12	(7, p. 249)
Madarin Chinese		
	3	(1, p. 28) (9, p. 35)
	9	(26, p. 191)
	11	(12, p. 235)
	12	(40, p. 265)
Moroccan Arabic		
	3	(30, p. 50) (35, p. 263)
Navajo		
	3	(46, p. 57)
Norwegian		
	12	(34, p. 262)
Portuguese		
	3	(31, p. 51)
	4	(28, p. 80) (29, p. 81)
Rendile		
	4	(5, p. 68)
Russian		
	3	(9, p. 35) (32, p. 51)
	4	(22, p. 77)
	5	(11, p. 95) (12, p. 96)
	7	(7, p. 134) (7, p. 134) (14, p. 137)
	9	(2, p. 169)
	10	(26, p. 211) (47, p. 226)
	12	(36, p. 263) (38, p. 265)
Spanish		
	3	(9, p. 35)
	7	(21, p. 143)
	12	(13, p. 252)
Standard Arabic		
	3	(30, p. 50)
	12	(27, p. 260) (35, p. 263)
Vietnamese		
	3	(29, p. 50)
	9	(1, p. 169)
Wolof		
	12	(35, p. 263)
Yiddish		
	4	(15, p. 72)
	10	(28, p. 213)

Appendix D

TDL FRAGMENTS

```
lex-or-phrase-synsem := synsem-min &
  [ LIGHT luk,
    L-PERIPH luk,
    R-PERIPH luk ].
cat := cat-min &
  [ HEAD head-min,
    VAL valence-min,
    MC luk,
    MKG mkg,
    HC-LIGHT luk,
    POSTHEAD bool ].
mrs := mrs-min &
  [ HOOK hook,
    RELS diff-list,
    HCONS diff-list,
    ICONS diff-list ].
hook := avm &
  [ GTOP handle,
    LTOP handle,
    INDEX individual,
    XARG individual,
    ICONS-KEY icons,
    CLAUSE-KEY event ].
icons := avm.
info-str := icons &
  [ CLAUSE individual,
    TARGET individual ].
non-topic := info-str.
contrast-or-focus := info-str.
focus-or-topic := info-str.
contrast-or-topic := info-str.
non-focus := info-str.
focus := non-topic & contrast-or-focus & focus-or-topic.
contrast := focus-or-topic & contrast-or-focus & contrast-or-topic.
topic := non-focus & focus-or-topic & contrast-or-topic.
bg := non-topic & non-focus.
semantic-focus := focus.
contrast-focus := contrast & focus.
contrast-topic := contrast & topic.
aboutness-topic := topic.
```

```
mkg := avm & [ FC luk, TP luk ].
fc := mkg & [ FC + ].
non-tp := mkg & [ TP - ].
tp := mkg & [ TP + ].
non-fc := mkg & [ FC - ].
fc-only := fc & non-tp.
fc-+-tp := tp & fc.
unmkg := non-tp & non-fc.
tp-only := tp & non-fc.
sform := basic-binary-headed-phrase.
focality := sform & [ SYNSEM.LOCAL.CAT.MKG fc-only ].
topicality := sform.
narrow-focus := focality & [ SYNSEM.LIGHT - ].
wide-focus := focality.
topicless := topicality &
 [ HEAD-DTR.SYNSEM.LOCAL.CAT.MKG non-tp,
   NON-HEAD-DTR.SYNSEM.LOCAL.CAT.MKG non-tp ].
topic-comment := topicality &
 [ SYNSEM.LOCAL.CAT.MKG tp,
   NON-HEAD-DTR.SYNSEM.LOCAL.CAT.MKG tp ].
focus-bg := narrow-focus & topicless.
all-focus := wide-focus & topicless &
  [ HEAD-DTR.SYNSEM.LOCAL.CAT.MKG fc ].
frame-setting := topic-comment &
  [ SYNSEM.L-PERIPH +,
    NON-HEAD-DTR.SYNSEM.L-PERIPH + ].
non-frame-setting := topic-comment &
  [ HEAD-DTR.SYNSEM.LOCAL.CAT.MKG non-tp ].
no-icons-lex-item := lex-item &
  [ SYNSEM.LOCAL [ CAT.MKG [ FC na, TP na ],
                   CONT.ICONS <! !> ] ].
basic-icons-lex-item := lex-item &
  [ SYNSEM.LOCAL.CONT.ICONS <! !> ].
one-icons-lex-item := lex-item &
  [ SYNSEM.LOCAL.CONT.ICONS <! info-str & [ ] !> ].
two-icons-lex-item := lex-item &
  [ SYNSEM.LOCAL.CONT.ICONS <! info-str & [ ], info-str & [ ] !> ].
no-icons-rule := phrase-or-lexrule & word-or-lexrule &
  [ C-CONT.ICONS <! !> ].
no-rels-hcons-rule := phrase-or-lexrule & word-or-lexrule &
  [ C-CONT [ RELS <! !>, HCONS <! !> ] ].
no-ccont-rule := no-rels-hcons-rule & no-icons-rule.
headed-phrase := phrase &
  [ SYNSEM.LOCAL [ CAT.HEAD head & #head,
                   AGR #agr,
                   COORD -,
                   COORD-REL #crel ],
    C-CONT.HOOK.ICONS-KEY #icons,
    HEAD-DTR.SYNSEM.LOCAL local &
                 [ CAT.HEAD #head,
                   CONT.HOOK.ICONS-KEY #icons,
                   AGR #agr,
                   COORD -,
                   COORD-REL #crel ] ].
```

```
basic-non-rel-clause := clause & head-compositional &
  [ SYNSEM.NON-LOCAL.REL 0-dlist,
    HEAD-DTR.SYNSEM [ LOCAL.CONT.HOOK [ INDEX #index,
                                        ICONS-KEY.CLAUSE #index,
                                        CLAUSE-KEY #index ],
                      NON-LOCAL [ QUE 0-dlist,
                                  REL 0-dlist ] ],
    C-CONT [ RELS <! !>,
             HCONS <! !> ] ].
basic-head-subj-phrase := head-valence-phrase & head-compositional &
                          basic-binary-headed-phrase & no-ccont-rule &
  [ SYNSEM phr-synsem &
                   [ LOCAL.CAT [ POSTHEAD +,
                                 HC-LIGHT -,
                                 VAL [ SUBJ < >,
                                       COMPS #comps,
                                       SPR #spr ] ] ],
        HEAD-DTR.SYNSEM.LOCAL.CAT.VAL [ SUBJ < #synsem >,
                                        COMPS #comps,
                                        SPR #spr ],
        NON-HEAD-DTR.SYNSEM #synsem & canonical-synsem &
                   [ LOCAL [ CAT [ VAL [ SUBJ olist,
                                         COMPS olist,
                                         SPR olist ] ] ],
                        NON-LOCAL [ SLASH 0-dlist & [ LIST < > ],
                                    REL 0-dlist,
                                    QUE 0-dlist ] ]].
basic-head-opt-subj-phrase := head-valence-phrase & head-only &
  [ INFLECTED #infl,
    SYNSEM canonical-synsem &
              [ LOCAL.CAT [ VAL [ SUBJ < >,
                                  COMPS #comps,
                                  SPR #spr,
                                  SPEC #spec ],
                            POSTHEAD #ph ],
                MODIFIED #mod ],
    HEAD-DTR [ INFLECTED #infl & infl-satisfied,
               SYNSEM [ LOCAL [ CAT [ VAL [ SUBJ < unexpressed-reg &
                                                   [ OPT +,
                                                     LOCAL.CONT.HOOK
                                                       [ INDEX #index &
                                                           [ COG-ST in-foc ],
                                                         ICONS-KEY #ikey,
                                                         CLAUSE-KEY #ckey ] ] >,
                                          COMPS #comps & < >,
                                          SPR #spr,
                                          SPEC #spec ],
                                    POSTHEAD #ph ],
                              CONT.HOOK.INDEX event ],
                        MODIFIED #mod ] ],
    C-CONT [ RELS <! !>,
             HCONS <! !>,
             ICONS <! #ikey & non-focus & [ TARGET #index,
                                            CLAUSE #ckey ] !> ] ].
```

```
basic-head-subj-nmc-phrase := head-valence-phrase &
                              basic-binary-headed-phrase &
                              declarative-clause & no-ccont-rule &
  [ SYNSEM phr-synsem &
                  [ LOCAL.CAT [ MC -,
                                POSTHEAD +,
                                HC-LIGHT -,
                                VAL [ SUBJ < >,
                                      COMPS < >,
                                      SPR #spr ] ] ],
    HEAD-DTR.SYNSEM
           [ LOCAL.CAT.VAL [ SUBJ < #synsem >,
                             COMPS < [ LOCAL #slash ] >,
                             SPR #spr ],
             NON-LOCAL.SLASH 1-dlist & [ LIST < #slash > ] ],
    NON-HEAD-DTR.SYNSEM #synsem & canonical-synsem &
           [ LOCAL [ CAT [ VAL [ SUBJ olist,
                                 COMPS olist,
                                 SPR olist ] ] ],
             NON-LOCAL [ SLASH 0-dlist & [ LIST < > ],
                         REL 0-dlist,
                         QUE 0-dlist ] ] ].
basic-head-comp-phrase := head-valence-phrase & head-compositional &
                          binary-headed-phrase & no-ccont-rule &
  [ SYNSEM phr-synsem-min &
                  [ LOCAL.CAT [ VAL [ SUBJ #subj,
                                      SPR #spr ],
                                POSTHEAD #ph,
                                HC-LIGHT #light ],
                    LIGHT #light ],
    HEAD-DTR.SYNSEM [ LOCAL.CAT [ VAL [ SUBJ #subj,
                                        SPR #spr ],
                                  HC-LIGHT #light,
                                  POSTHEAD #ph ]],
    NON-HEAD-DTR.SYNSEM canonical-synsem ].
```

```
basic-head-opt-comp-phrase := head-valence-phrase & head-only &
                              head-compositional &
  [ INFLECTED #infl,
    SYNSEM canonical-synsem &
              [ LOCAL.CAT [ VAL [ SUBJ #subj,
                                  COMPS #comps,
                                  SPR #spr,
                                  SPEC #spec ],
                            MC #mc,
                            POSTHEAD #ph ],
                MODIFIED #mod ],
    HEAD-DTR [ INFLECTED #infl & infl-satisfied,
               SYNSEM [ LOCAL [ CAT [ VAL [ SUBJ #subj,
                                            COMPS < unexpressed &
                                              [ OPT +,
                                                OPT-CS #def,
                                                LOCAL.CONT.HOOK
                                                 [ INDEX #index & [ COG-ST #def ],
                                                   ICONS-KEY #ikey,
                                                   CLAUSE-KEY #ckey ] ] . #comps >,
                                            SPR #spr,
                                            SPEC #spec ],
                                      MC #mc,
                                      POSTHEAD #ph ],
                                 CONT.HOOK.INDEX event ],
                        MODIFIED #mod ] ],
    C-CONT [ RELS <! !>,
             HCONS <! !>,
             ICONS <! #ikey & non-focus & [ TARGET #index,
                                            CLAUSE #ckey ] !> ] ].
basic-head-comp-nmc-phrase := head-valence-phrase & head-compositional &
                              binary-headed-phrase & no-ccont-rule &
  [ SYNSEM phr-synsem &
           [ LOCAL.CAT [ MC -,
                         VAL [ SUBJ < >,
                               SPR #spr ],
                         POSTHEAD #ph,
                         HC-LIGHT #light ],
             LIGHT #light ],
    HEAD-DTR.SYNSEM
           [ LOCAL.CAT [ VAL [ SUBJ < [ LOCAL #slash ] >,
                               SPR #spr ],
                         HC-LIGHT #light,
                         POSTHEAD #ph ],
             NON-LOCAL.SLASH 1-dlist & [ LIST < #slash > ] ],
    NON-HEAD-DTR.SYNSEM canonical-synsem ].
```

```
basic-head-mod-phrase-simple := head-mod-phrase & binary-headed-phrase &
  [ SYNSEM [ NON-LOCAL [ SLASH [ LIST #first, LAST #last ],
                         REL 0-dlist ] ],
    HEAD-DTR.SYNSEM
           [ LOCAL [ CAT [ HEAD #head,
                           VAL #val,
                           POSTHEAD #ph,
                           MC #hmc,
                           HC-LIGHT - ],
                     CONT.HOOK #hdhook & [ ICONS-KEY.CLAUSE #clause ],
                     AGR #agr ],
             NON-LOCAL #nonloc &
                   [ SLASH [ LIST #middle, LAST #last ] ],
             LIGHT #light,
             MODIFIED #modif ],
    NON-HEAD-DTR.SYNSEM
           [ LOCAL [ CAT [ HEAD [ MOD < [ LOCAL local &
                                             [ CAT [ HEAD #head,
                                                     VAL #val,
                                                     POSTHEAD #ph,
                                                     MC #hmc ],
                                               AGR #agr,
                                               CONT.HOOK #hdhook ],
                                          NON-LOCAL #nonloc,
                                          LIGHT #light,
                                          MODIFIED #modif ] > ],
                           VAL [ COMPS olist, SPR olist ] ],
                     CONT.HOOK [ ICONS-KEY.CLAUSE #clause,
                                 CLAUSE-KEY #clause ] ],
             NON-LOCAL [ SLASH [ LIST #first, LAST #middle ],
                         QUE 0-dlist & [ LIST null ] ] ],
    C-CONT [ RELS <! !>, ICONS <! !> ] ].
nc-filler-phrase := binary-phrase & phrasal &
  [ SYNSEM [ LOCAL [ CAT [ VAL [ COMPS < >,
                                 SPR < > ],
                                 POSTHEAD + ] ],
             NON-LOCAL.SLASH 0-dlist ],
    ARGS <  [ SYNSEM [ LOCAL.CAT [ VAL.COMPS olist ],
                       NON-LOCAL [ SLASH 1-dlist &
                                         [ LIST [ FIRST #slash,
                                                  REST < > & #last ],
                                           LAST #last ],
                                   QUE 0-dlist,
                                   REL 0-dlist ] ] ],
            [ SYNSEM [ LOCAL #slash & local &
                                 [ CAT.VAL [ SUBJ olist,
                                             COMPS olist,
                                             SPR olist ],
                                   CTXT.ACTIVATED + ],
                       NON-LOCAL.SLASH 0-dlist ] ] > ].
```

Appendix E

LIST OF THE CHOICES FILES FOR REGRESSTION TESTING

TYPE A

```
infostr-foc-sov-initial
infostr-foc-sov-final
infostr-foc-sov-prev
infostr-foc-sov-postv

infostr-foc-svo-initial
infostr-foc-svo-final
infostr-foc-svo-prev
infostr-foc-svo-postv

infostr-foc-osv-initial
infostr-foc-osv-final
infostr-foc-osv-prev
infostr-foc-osv-postv

infostr-foc-ovs-initial
infostr-foc-ovs-final
infostr-foc-ovs-prev
infostr-foc-ovs-postv

infostr-foc-vso-initial
infostr-foc-vso-final
infostr-foc-vso-prev
infostr-foc-vso-postv

infostr-foc-vos-initial
```

infostr-foc-vos-final
infostr-foc-vos-prev
infostr-foc-vos-postv

infostr-foc-vf-initial
infostr-foc-vf-prev

infostr-foc-vi-final
infostr-foc-vi-postv

infostr-foc-v2-initial
infostr-foc-v2-final
infostr-foc-v2-prev
infostr-foc-v2-postv

TYPE B

infostr-foc-initial-topic-first
infostr-foc-final-topic-first-cf-prev
infostr-foc-final-topic-on-cf-prev

TYPE C-1

infostr-foc-affix-after-noun
infostr-foc-adp-after-noun
infostr-foc-mod-after-noun

TYPE C-2

infostr-top-mod-after-noun
infostr-cf-mod-after-noun
infostr-ct-mod-after-noun

TYPE C-3

infostr-foc-mod-after-verb

```
infostr-foc-mod-before-noun
infostr-foc-mod-both-noun
infostr-foc-mod-after-both
infostr-foc-mod-both-both
```

Appendix F

SAMPLES OF CHOICES FILE

TYPE A

```
version=28

section=general
language=infostr-foc-svo-initial

section=word-order
word-order=svo
has-dets=yes
noun-det-order=det-noun
has-aux=no

section=number
  number1_name=sing
  number2_name=plural

section=person
person=1-2-3
first-person=none

section=gender
  gender1_name=masc
  gender2_name=fem

section=case
case-marking=none
```

```
section=direct-inverse

section=tense-aspect-mood

section=other-features

section=sentential-negation

section=coordination

section=matrix-yes-no

section=info-str
focus-pos=clause-initial

section=arg-opt

section=lexicon
  noun1_name=common-noun
  noun1_det=opt
    noun1_stem1_orth=CN
    noun1_stem1_pred=_dog_n_rel
  noun2_name=proper-noun
  noun2_det=imp
    noun2_stem1_orth=PN
    noun2_stem1_pred=_Kim_n_rel
  noun3_name=pronoun
  noun3_det=imp
    noun3_stem1_orth=PRO
    noun3_stem1_pred=pronoun_q_rel
  verb1_name=intransitive-verb
  verb1_valence=intrans
    verb1_stem1_orth=IV
```

```
    verb1_stem1_pred=_bark_v_rel
  verb2_name=transitive-verb
  verb2_valence=trans
    verb2_stem1_orth=TV
    verb2_stem1_pred=_chase_v_rel
  det1_name=det
    det1_stem1_orth=DET
    det1_stem1_pred=exist_q_rel

section=morphology

section=toolbox-import

section=test-sentences
  sentence1_orth=CN IV
  sentence2_orth=IV CN
  sentence2_star=on
  sentence3_orth=PN TV CN
  sentence4_orth=PN CN TV
  sentence5_orth=CN PN TV
  sentence6_orth=CN TV PN
  sentence7_orth=TV PN CN
  sentence7_star=on
  sentence8_orth=TV CN PN
  sentence8_star=on

section=gen-options

section=ToolboxLexicon
```

TYPE B

```
version=28
```

```
section=general
language=infostr-foc-initial-topic-first

section=word-order
word-order=svo
has-dets=yes
noun-det-order=det-noun
has-aux=no

section=number
  number1_name=sing
  number2_name=plural

section=person
person=1-2-3
first-person=none

section=gender
  gender1_name=masc
  gender2_name=fem

section=case
case-marking=none

section=direct-inverse

section=tense-aspect-mood

section=other-features

section=sentential-negation

section=coordination
```

```
section=matrix-yes-no

section=info-str
focus-pos=clause-initial
topic-first=on

section=arg-opt

section=lexicon
  noun1_name=common-noun
  noun1_det=opt
    noun1_stem1_orth=CN
    noun1_stem1_pred=_dog_n_rel
  noun2_name=proper-noun
  noun2_det=imp
    noun2_stem1_orth=PN
    noun2_stem1_pred=_Kim_n_rel
  noun3_name=pronoun
  noun3_det=imp
    noun3_stem1_orth=PRO
    noun3_stem1_pred=pronoun_q_rel
  verb1_name=intransitive-verb
  verb1_valence=intrans
    verb1_stem1_orth=IV
    verb1_stem1_pred=_bark_v_rel
  verb2_name=transitive-verb
  verb2_valence=trans
    verb2_stem1_orth=TV
    verb2_stem1_pred=_chase_v_rel
  det1_name=det
    det1_stem1_orth=DET
    det1_stem1_pred=exist_q_rel
```

```
section=morphology

section=toolbox-import

section=test-sentences
  sentence1_orth=CN IV
  sentence2_orth=IV CN
  sentence3_orth=PN TV CN
  sentence4_orth=PN CN TV
  sentence5_orth=CN PN TV
  sentence6_orth=CN TV PN
  sentence7_orth=TV PN CN
  sentence8_orth=TV CN PN

section=gen-options

section=ToolboxLexicon
```

TYPE C-1

```
version=28

section=general
language=infostr-foc-affix-after-noun

section=word-order
word-order=svo
has-dets=yes
noun-det-order=det-noun
has-aux=no

section=number
  number1_name=sing
```

```
  number2_name=plural

section=person
person=1-2-3
first-person=none

section=gender
  gender1_name=masc
  gender2_name=fem

section=case
case-marking=nom-acc
nom-acc-nom-case-name=nom
nom-acc-acc-case-name=acc

section=direct-inverse

section=tense-aspect-mood

section=other-features

section=sentential-negation

section=coordination

section=matrix-yes-no

section=info-str
  focus-marker1_type=affix

section=arg-opt

section=lexicon
```

```
  noun1_name=common-noun
  noun1_det=opt
    noun1_stem1_orth=CN
    noun1_stem1_pred=_dog_n_rel
  noun2_name=proper-noun
  noun2_det=imp
    noun2_stem1_orth=PN
    noun2_stem1_pred=_Kim_n_rel
  noun3_name=pronoun
  noun3_det=imp
    noun3_stem1_orth=PRO
    noun3_stem1_pred=pronoun_q_rel
  verb1_name=intransitive-verb
  verb1_valence=intrans
    verb1_stem1_orth=IV
    verb1_stem1_pred=_bark_v_rel
  verb2_name=transitive-verb
  verb2_valence=trans
    verb2_stem1_orth=TV
    verb2_stem1_pred=_chase_v_rel
  det1_name=det
    det1_stem1_orth=DET
    det1_stem1_pred=exist_q_rel

section=morphology
  noun-pc1_name=p1
  noun-pc1_order=suffix
  noun-pc1_inputs=noun
    noun-pc1_lrt1_name=r1
      noun-pc1_lrt1_feat1_name=information-structure marking
      noun-pc1_lrt1_feat1_value=fc
      noun-pc1_lrt1_feat2_name=information-structure meaning
      noun-pc1_lrt1_feat2_value=focus
```

```
      noun-pc1_lrt1_lri1_inflecting=yes
      noun-pc1_lrt1_lri1_orth=-FC

section=toolbox-import

section=test-sentences
  sentence1_orth=CN IV
  sentence2_orth=IV CN
  sentence3_orth=PN TV CN
  sentence4_orth=PN CN TV
  sentence5_orth=CN PN TV
  sentence6_orth=CN TV PN
  sentence7_orth=TV PN CN
  sentence8_orth=TV CN PN

section=gen-options

section=ToolboxLexicon
```

TYPE C-2

```
version=28

section=general
language=infostr-top-mod-after-noun

section=word-order
word-order=svo
has-dets=yes
noun-det-order=det-noun
has-aux=no

section=number
  number1_name=sing
```

```
  number2_name=plural

section=person
person=1-2-3
first-person=none

section=gender
  gender1_name=masc
  gender2_name=fem

section=case
case-marking=nom-acc
nom-acc-nom-case-name=nom
nom-acc-acc-case-name=acc

section=direct-inverse

section=tense-aspect-mood

section=other-features

section=sentential-negation

section=coordination

section=matrix-yes-no

section=info-str
  topic-marker1_type=modifier
  topic-marker1_pos=after
  topic-marker1_cat=nouns
  topic-marker1_orth=TP
```

```
section=arg-opt

section=lexicon
  noun1_name=common-noun
  noun1_det=opt
    noun1_stem1_orth=CN
    noun1_stem1_pred=_dog_n_rel
  noun2_name=proper-noun
  noun2_det=imp
    noun2_stem1_orth=PN
    noun2_stem1_pred=_Kim_n_rel
  noun3_name=pronoun
  noun3_det=imp
    noun3_stem1_orth=PRO
    noun3_stem1_pred=pronoun_q_rel
  verb1_name=intransitive-verb
  verb1_valence=intrans
    verb1_stem1_orth=IV
    verb1_stem1_pred=_bark_v_rel
  verb2_name=transitive-verb
  verb2_valence=trans
    verb2_stem1_orth=TV
    verb2_stem1_pred=_chase_v_rel
  det1_name=det
    det1_stem1_orth=DET
    det1_stem1_pred=exist_q_rel

section=morphology

section=toolbox-import

section=test-sentences
  sentence1_orth=CN IV
```

```
  sentence2_orth=CN TP IV
  sentence3_orth=TP CN IV
  sentence3_star=on
  sentence4_orth=PN TV CN
  sentence5_orth=PN TP TV CN
  sentence6_orth=TP PN TV CN
  sentence6_star=on
  sentence7_orth=PN TP TV CN TP

section=gen-options

section=ToolboxLexicon
```

TYPE C-3

```
version=28

section=general
language=infostr-foc-mod-both-both

section=word-order
word-order=svo
has-dets=yes
noun-det-order=det-noun
has-aux=no

section=number
  number1_name=sing
  number2_name=plural

section=person
person=1-2-3
first-person=none
```

```
section=gender
  gender1_name=masc
  gender2_name=fem

section=case
case-marking=nom-acc
nom-acc-nom-case-name=nom
nom-acc-acc-case-name=acc

section=direct-inverse

section=tense-aspect-mood

section=other-features

section=sentential-negation

section=coordination

section=matrix-yes-no

section=info-str
  focus-marker1_type=modifier
  focus-marker1_pos=before, after
  focus-marker1_cat=nouns, verbs
  focus-marker1_orth=FC

section=arg-opt

section=lexicon
  noun1_name=common-noun
  noun1_det=opt
    noun1_stem1_orth=CN
```

```
    noun1_stem1_pred=_dog_n_rel
  noun2_name=proper-noun
  noun2_det=imp
    noun2_stem1_orth=PN
    noun2_stem1_pred=_Kim_n_rel
  noun3_name=pronoun
  noun3_det=imp
    noun3_stem1_orth=PRO
    noun3_stem1_pred=pronoun_q_rel
  verb1_name=intransitive-verb
  verb1_valence=intrans
    verb1_stem1_orth=IV
    verb1_stem1_pred=_bark_v_rel
  verb2_name=transitive-verb
  verb2_valence=trans
    verb2_stem1_orth=TV
    verb2_stem1_pred=_chase_v_rel
  det1_name=det
    det1_stem1_orth=DET
    det1_stem1_pred=exist_q_rel

section=morphology

section=toolbox-import

section=test-sentences
  sentence1_orth=CN IV
  sentence2_orth=CN FC IV
  sentence3_orth=FC CN IV
  sentence4_orth=CN IV FC
  sentence5_orth=CN FC IV FC

section=gen-options
```

Appendix G

TESTSUITE FOR MULTILINGUAL MACHINE TRANSLATION

English

[1] The dog barks
[2] The dog-a barks
[3] The dog barks-a
[4] The dog-b barks
[5] The dog-b barks-a
[6] The dog-a barks-a
[7] Kim reads the book
[8] Kim-a reads the book
[9] Kim reads-a the book
[10] Kim reads the book-a
[11] Kim-b reads-a the book
[12] Kim-b reads the book-a

Japanese

[1] 犬 吠える
[2] 犬 が 吠える
[3] 犬 吠える
[4] 犬 は 吠える
[5] 犬 は 吠える
[6] 犬 が 吠える
[7] キム 本 読む
[8] キム が 本 読む
[9] キム 本 読む
[10] キム 本 を 読む
[11] キム は 本 読む
[12] キム は 本 を 読む

Korean

[1] 개 짖다
[2] 개가 짖다
[3] 개 짖다
[4] 개는 짖다
[5] 개는 짖다
[6] 개가 짖다
[7] 김 책 읽다
[8] 김이 책 읽다
[9] 김 책 읽다
[10] 김 책을 읽다
[11] 김은 책 읽다
[12] 김은 책을 읽다