Porting Grammars between Typologically Similar Languages: Japanese to Korean

Roger KIM
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304 USA
rkim@parc.com

Mary DALRYMPLE
Dept. of Computer Science
King's College London
Strand, London WC2R 2LS UK
mary@dcs.kcl.ac.uk

Ronald M. KAPLAN
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304 USA
kaplan@parc.com

Tracy Holloway KING
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304 USA
thking@parc.com

Abstract

We report on a preliminary investigation of the difficulty of converting a grammar of one language into a grammar of a typologically similar language. In this investigation, we started with the ParGram grammar of Japanese and used that as the basis for a grammar of Korean. The results are encouraging for the use of grammar porting to bootstrap new grammar development.

1 Introduction

The Parallel Grammar project (ParGram) is an international collaboration aimed at producing broad-coverage computational grammars for a variety of languages (Butt et al., 1999; Butt et al., 2002). The grammars (currently of English, French, German, Japanese, Norwegian, and Urdu) are written in the framework of Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Dalrymple, 2001), and they are constructed using a common engineering and high-speed processing platform for LFG grammars, the XLE (Maxwell and Kaplan, 1993). These grammars, as do all LFG grammars, assign two levels of syntactic representation to the sentences of a language: a superficial phrase structure tree (called a *constituent structure* or c-structure) and an underlying matrix of features and values (the *functional structure* or f-structure). The c-structure records the order of words in a sentence and their hierarchical grouping into phrases. The f-structure encodes the grammatical functions, syntactic features, and predicate-argument relations conveyed by the sentence. F-structures are meant to encode a language universal level of analysis, allowing for cross-linguistic parallelism at this level of abstraction.

The ParGram project attempts to test the LFG formalism for its universality and coverage and to see how far parallelism can be maintained across languages. Previous ParGram work and much theoretical analysis has largely confirmed the universality claims of LFG theory. The f-structures produced by the grammars for similar constructions in each language have the same major functions and features, with minor variations across languages (e.g., the f-structures for French nouns have a grammatical gender feature but that distinction is not marked in English f-structures). This uniformity has the computational advantage that the grammars can be used in similar applications and that machine translation (Frank, 1999) can be simplified.

We have found that it takes roughly two person-years of effort to construct for a new language a grammar that approximates existing grammars in terms of coverage and accuracy (see (Riezler et al., 2002) for a discussion of the coverage and accuracy of the current English grammar). This suggests that the deep-grammar construction task is not as difficult as many people have believed, and indeed may require less effort than would be needed to produce training materials for automatic learning procedures for shallower grammars. However, it is still interesting to explore methods for reducing the linguistic effort that grammar construction requires. To that end, we report here on a preliminary investigation of the difficulty of converting a grammar of one language into a grammar of a typologically similar language. In this investigation, we started with the ParGram grammar of Japanese (Masuichi and Ohkuma, 2003) and used that as the basis for a grammar of Korean.

Typologically similar but not necessarily genetically related languages are those that not only admit of similar f-structures, as LFG theory suggests is the case with all languages, but also have similar c-structure to f-structure mappings. Whether or not Japanese and Korean are genetically related (an issue that is in some dispute; see (Sohn, 1999) for some discussion), Japanese and Korean are typologically similar in at least the following ways: they both are verb final and more generally head final, have relatively free word order, use postpositions to mark grammatical functions, and exhibit rampant pro-drop.

2 Grammar Porting: Direct Port

We found that many rules of the Japanese grammar could be used without modification in the Korean grammar. This was particularly the case for the majority of annotated phrase structure rules that produce the LFG c- and f-structures for the basic constructions of languages. In this section, we discuss the areas in which direct porting was possible.

2.1 The Japanese ParGram Grammar

The creation of the current Japanese grammar involved two person years of work at Fuji Xerox (see (Masuichi and Ohkuma, 2003) for details on the design of the grammar system and on its coverage). The grammar has broad coverage, providing parses for 97% of sentences in a large test suite with good accuracy. Ambiguity is kept to a minimum due to the integration of the ChaSen tokenizer as a preprocessor. In addition to string segmentation, the ChaSen tokenizer helps to disambiguate part of speech in the input. To give an idea of its size, the Japanese grammar has 54 rules which compile into finite-state machines with a total of 360 states, 1247 arcs, and 1830 disjuncts.

2.2 Grammar Rules

One set of rules that could be ported directly from the Japanese grammar to the Korean one were those characterizing clausal word-order possibilities. These rules encode basic verb final order with free ordering of preceding arguments and adjuncts but also include some markings for prefered word orders (e.g., subject preceding object). Sample orders covered by the grammars are shown in (1) and (2).

(1) a. Ayuko ga gakusei ni hon wo ageta.
Ayuko NOM student DAT book ACC gave
'Ayuko gave the student a book.' (Japanese)

b. gakusei ni hon wo Ayuko ga ageta.

c. hon wo Ayuko ga gakusei ni ageta.

(2) a. Myungwoni ga haksaeng ehgeh chaek ul juuttda.
Myungwoni NOM student DAT book ACC gave
'Myungwoni gave the student a book.' (Korean)

b. haksaeng ehgeh chaek ul Myungwoni ga juuttda.

c. chaek ul Myungwoni ga haksaeng ehgeh juuttda.

The structures for the Korean sentence in (2a) is shown in Figure 1. The corresponding Japanese structure for (1a) is shown in Figure 2.

Similarly, the rules for topicalization could also be ported without modification. In Japanese, noun phrases are marked as topicalized by the postposition *ha*. Topicalized noun phrases may have certain postpositions before the final *ha*, as in (3a). However, nominals with postposition case markers such as

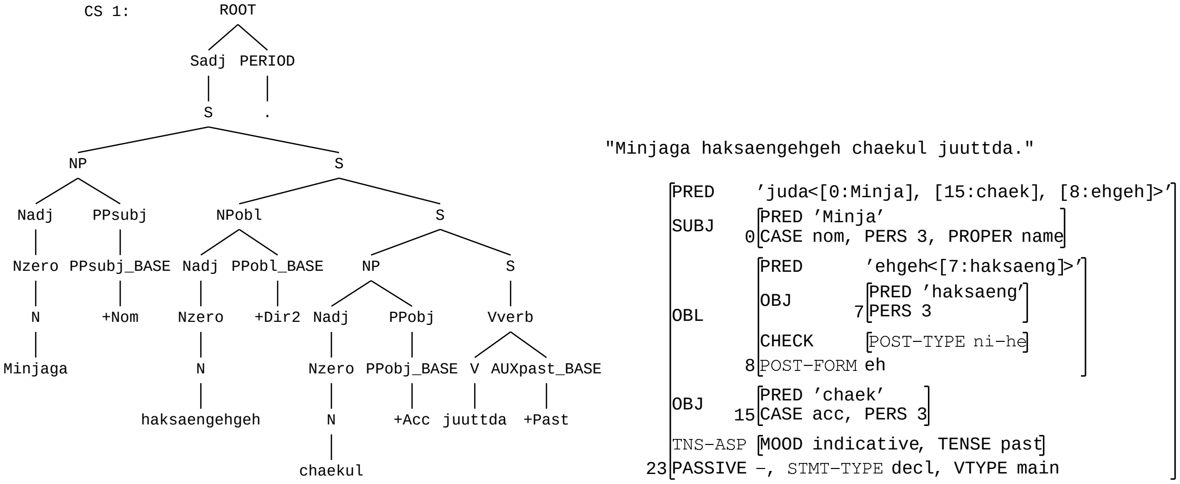

CS 1:
ROOT
Sadj PERIOD
S .
NP S
Nadj PPsubj NPobl S
Nzero PPsubj_BASE Nadj PPobl_BASE NP S
N +Nom Nzero +Dir2 Nadj PPobj Vverb
Minjaga N Nzero PPobj_BASE V AUXpast_BASE
haksaengehgeh N +Acc juuttda +Past
chaekul
"Minjaga haksaengehgeh chaekul juuttda."
PRED 'juda<[0:Minja], [15:chaek], [8:ehgeh]>'
SUBJ 0 PRED 'Minja' CASE nom, PERS 3, PROPER name
OBL 8 PRED 'ehgeh<[7:haksaeng]>' OBJ 7 PRED 'haksaeng' PERS 3 CHECK POST-TYPE ni-he POST-FORM eh
OBJ 15 PRED 'chaek' CASE acc, PERS 3
TNS-ASP MOOD indicative, TENSE past
23 PASSIVE -, STMT-TYPE decl, VTYPE main

Figure 1: Korean c-structure and f-structure for (2a)

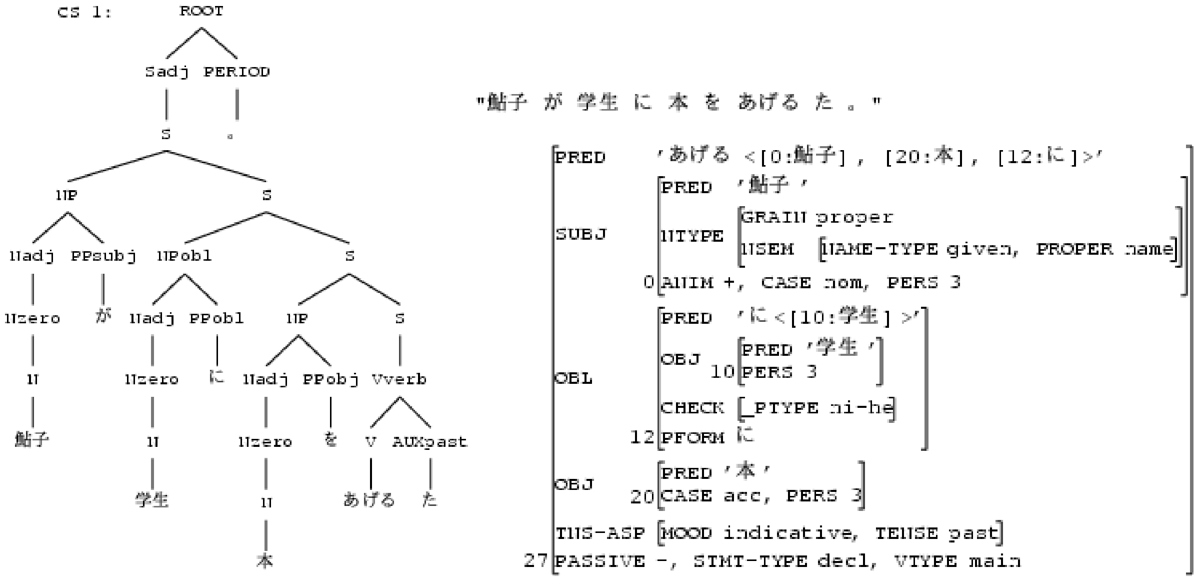

CS 1:
ROOT
Sadj PERIOD
S 。
NP S
Nadj PPsubj NPobl S
Nzero が Nadj PPobl NP S
N Nzero に Nadj PPobj Vverb
鮎子 N Nzero を V AUXpast
学生 N あげる た
本
"鮎子 が 学生 に 本 を あげる た 。"
PRED 'あげる<[0:鮎子], [20:本], [12:に]>'
SUBJ 0 PRED '鮎子' NTYPE GRAIN proper NSEM NAME-TYPE given, PROPER name ANIM +, CASE nom, PERS 3
OBL 12 PRED 'に<[10:学生]>' OBJ 10 PRED '学生' PERS 3 CHECK PTYPE ni-he PFORM に
OBJ 20 PRED '本' CASE acc, PERS 3
TNS-ASP MOOD indicative, TENSE past
27 PASSIVE -, STMT-TYPE decl, VTYPE main

Figure 2: Japanese c-structure and f-structure for (1a)

wo, *ga*, or *no* cannot be topicalized by *ha*. Instead, the postposition is dropped and only *ha* appears, as in (3c). In addition, these phrases are marked in the f-structure as to their topic status; this f-structure information controls their syntactic distribution in the sentence. The corresponding topicalizing postposition in Korean is *un/nun*, with the allomorph *un* following consonant-final nominals and *nun* following vowel-final nominals, as in (3b). Just as with the Japanese *ha*, the Korean topicalizer also cannot cooccur with postpositions marking the basic grammatical functions, as in (3d).

(3)	a.	kinoo	made	ha	b.	uhjeh	kaji	nun
		yesterday	until	TOPIC		yesterday	until	TOPIC
		'until yesterday (topic)' (Japanese)				'until yesterday (topic)' (Korean)		
	c.	*kinoo ga ha ⟹ kinoo ha			d.	*uhjeh ga nun ⟹ uhjeh nun		

Nominal internal structure was also ported directly from the Japanese grammar to the Korean. This includes the analysis of adjectival, nominal, and postpositional modifiers of the head noun. For example, the rules used to produce the analysis for the Japanese complex nominal in (4a) were ported directly to produce the analysis of the Korean nominal in (4b).

(4)	a.	Ayuko-no	ookii	e	hon	b.	Myungwoniui	kun	kurim	chaek
		Ayuko-GEN	big	picture	book		Myungwoni-GEN	big	picture	book
		‘Ayuko’s big picture book’ (Japanese)					‘Myungwoni’s big picture book’ (Korean)			

Similarly, the rules building oblique noun phrases, i.e., noun phrases with postpositions that serve as oblique arguments of verbs, were ported directly. An example in Japanese is shown in (5a) with the corresponding Korean phrase in (5b).

(5)	a.	ooki	ie	ni	b.	kun	jib	eh
		big	house	at		big	house	at
		‘in the big house’ (Japanese)				‘in the big house’ (Korean)		

A rule fragment for these oblique noun phrases from the Japanese grammar is shown in (6).

```
(6)  NPobl ---> { Nadj: (ˆ OBJ)=!
                  PPobl: ˆ=!
                 |AN: (ˆ OBJ)=!
                  PPobl: ˆ=!
                        (! CHECK POST-TYPE)=c 'to-ni'
                 |...}.
```

In the first disjunct in Rule (6) (from { to |), the NPobl consists of an Nadj which is the OBJ of the corresponding f-structure (Nadj: (ˆOBJ)=!) followed by a PPobl postposition which is the head of the corresponding f-structure (PPobl: ˆ=!).[1] The second disjunct (from | to |) is similar except that the PPobl is restricted to postpositions of the type to-ni and the OBJ of this postposition is an AN instead of the usual Nadj. Note that the to-ni value is particular to Japanese; however, the corresponding Korean form can be provided with this value to satisfy the constraints.[2] Other disjuncts are found in this rule, indicated here by |...}.

We were also able to port the implementation of pro-drop for subjects and objects. Examples of sentences with a pro-dropped subjects for Japanese and Korean are seen in (7). The analysis corresponding to the Korean sentence is shown in Figure 3.

(7)	a.	jitensha	de	ie	ni	kaeru.	
		bicycle	by	home	to	return	
		‘(I/You/He/She/We/They) return home by bicycle.’ (Japanese)					
	b.	jajungu	ro	jib	eh	dorakanda.	
		bicycle	by	home	to	return	
		‘(I/You/He/She/We/They) return home by bicycle.’ (Korean)					

Pro-drop is analyzed by optionally providing a null pronominal subject and/or object for each verb frame that subcategorizes for these functions. If an overt subject or object is found in the clause, then the pro-drop option is not chosen because the PRED of the overt subject would fail to unify with the PRED of the optional pronominal subject. However, if there is no overt subject, then the pro-drop option must be chosen because otherwise the subcategorization requirements of the verb would not be met. The null-anaphor template NA that provides the pronominal arguments is shown in (8a), where GF is a grammatical function value that is passed in by the verbal template. (8b) shows the expansion of the template for pro-dropped subjects.

[1] In the XLE grammar development platform, the ˆ corresponds to the traditional LFG ↑ and the ! corresponds to the traditional LFG ↓.

[2] In a number of places in the grammar, surface form features are checked; the values of these features must be changed from Japanese to their Korean equivalents. However, once this translational port is made, the rules can be used directly.

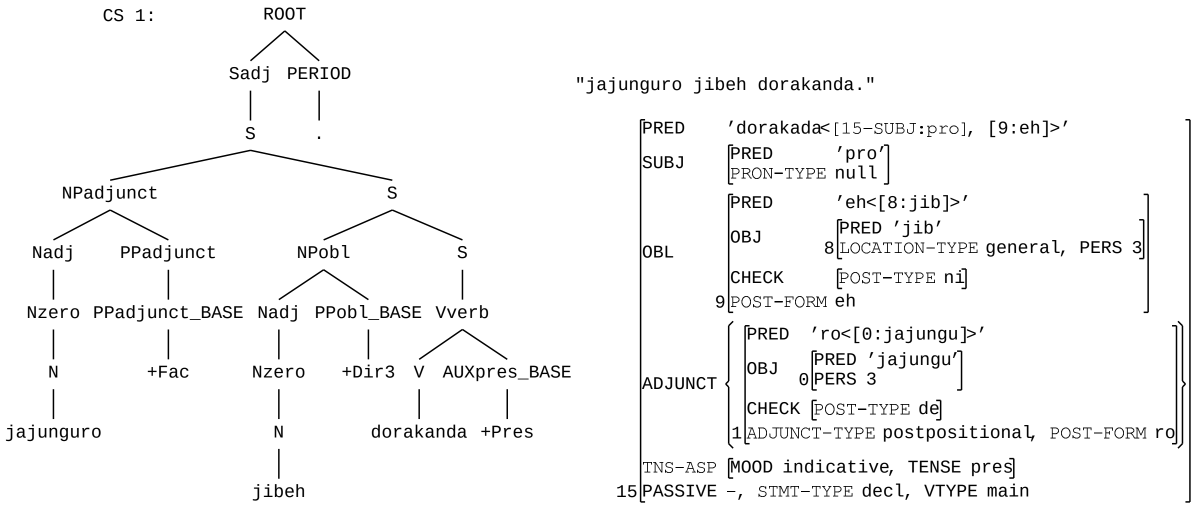

CS 1:
ROOT
Sadj
PERIOD
S
.
NPadjunct
S
Nadj
PPadjunct
NPobl
S
Nzero
PPadjunct_BASE
Nadj
PPobl_BASE
Vverb
N
+Fac
Nzero
+Dir3
V
AUXpres_BASE
jajunguro
N
dorakanda
+Pres
jibeh
"jajunguro jibeh dorakanda."
PRED 'dorakada<[15-SUBJ:pro], [9:eh]>'
SUBJ PRED 'pro' PRON-TYPE null
OBL PRED 'eh<[8:jib]>'
OBJ PRED 'jib' 8 LOCATION-TYPE general, PERS 3
CHECK POST-TYPE ni
9 POST-FORM eh
ADJUNCT PRED 'ro<[0:jajungu]>'
OBJ PRED 'jajungu' 0 PERS 3
CHECK POST-TYPE de
1 ADJUNCT-TYPE postpositional, POST-FORM ro
TNS-ASP MOOD indicative, TENSE pres
15 PASSIVE -, STMT-TYPE decl, VTYPE main

Figure 3: Pro-drop: Korean c-structure and f-structure for (7b)

(8) a. NA(GF) = @(DEFAULT (ˆ GF PRED) (ˆ GF PRON-TYPE) 'pro' null).
b. (ˆ SUBJ PRED)='pro' (ˆ SUBJ PRON-TYPE)=null

The ability to drop postpositional case and discourse function markers was also ported directly. In a standard SOV (or OSV) sentence, if the accusative case marker is dropped, but the nominative case marker is not, as in (9b), the sentence is given only one parse with the case-less noun phrase taking the object function and the nominative case-marked noun phrase being the subject of the sentence. The same holds when only the nominative case marker is dropped but the accusative is not, as in (9c). When both case markings are dropped, as in (9d), the sentence is given two parses with each noun phrase being the subject in one parse and the object in the other. The Japanese equivalents of (9) receive the same analyses.

(9) a. Minjaga Taesunul boattda.
Minja-NOM Taesun-ACC saw
'Minja saw Taesun.' (Korean)
b. Minjaga Taesun boattda.
c. Minja Taesunul boattda.
d. Minja Taesun boattda.
'Minja saw Taesun.' (preferred due to default word order)

Thus, due to the similarity of the c-structure between Japanese and Korean and to the similarity of the mapping from c-structure to f-structure, annotated phrase structure rules already existing in the Japanese grammar could be ported without change to a Korean grammar, as illustrated here by Rule (6) and Template (8).

3 Grammar Porting: Necessary Changes

The previous section described ways in which Japanese grammar rules could be used directly in the Korean grammar. This was true for many core constructions, thus eliminating the need to construct such core rules for the Korean grammar. However, there are several aspects of the Japanese grammar that could not be ported directly. In this section we discuss three places where this was the case: the tokenizer and morphology, the lexicon, and some of the rules.

3.1 The Tokenizer and Morphology

Tokenization and morphological analysis could not be carried over from Japanese to Korean. The Japanese grammar uses the independently developed ChaSen module for a single processing stage that divides the text into tokens and at the same time provides certain part-of-speech information. The output of this stage becomes the immediate input to the syntactic analysis component, bypassing the tokenizing and morphology transductions that perform these functions for the other ParGram languages.[3] However, Korean typographical conventions, both in the Romanization we used in this experiment and also in normal Hangul script, are typologically much more similar to those of the other ParGram languages in the way that spaces and punctuation are used to delimit words. We were thus able to port the English finite-state tokenizing transducer to Korean with only a few minor modifications: we removed the provisions in the English transducer for lower-casing capitalized words and we eliminated the special treatment of periods as abbreviation indicators.

We were also able to use standard finite-state technology (see (Beesley and Karttunen, 2003) and references therein) to construct a simple morphological analyzer for Korean.[4] The morphology transducer receives the word-tokens produced by the tokenizing transduction and decomposes the Korean nouns and verbs it identifies into stems and affixes. For example, *chaekul* is analyzed by the morphology as *chaek +Noun +Acc*. The stems and affixes are then referred to the Korean lexicon to obtain their f-structure meaning and inflectional properties.

3.2 The Lexicon

Unlike the majority of annotated c-structure rules, the lexical items differ significantly between Japanese and Korean, and the lexicon needed substantial modification in the grammar porting. However, once the lexical item head-words were changed, the information in many entries remained the same. For example, the entry for the Japanese accusative postposition *wo* is identical to that of the Korean accusative postposition *ul/rul* other than the fact that the Korean postposition is mapped onto a *+Acc* morphological tag, e.g., both assign accusative case in the same environments. Thus, the Japanese entry for *wo* was ported to the Korean entry for *+Acc*. This was similar for the majority of closed class items.

At this point we are working with a toy lexicon for open class items, although all the closed class items have been translated. We anticipate no problem for open class items with predictable subcategorization frames such as nouns, adjectives, and adverbs. In fact, no direct translation porting of these items is necessary because given a large morphological analyzer for Korean, these forms will not need explicit lexical entries. Instead, lexical entries will be created “on the fly” based on the morphological tags. For example, there would be no overt lexical entry for the name *Minja*. Instead, the morphology would produce the stem and tags of the form *Minja +Noun +Proper*. Based on these tags, the noun would receive the correct part of speech and f-structure information for a proper noun. This is the system currently employed in the other ParGram grammars for items with predictable (or no) subcategorization frames. It is implemented in the Korean grammar for the small morphology that is currently being used and will need little modification to scale to a broad coverage morphology. Unfortunately, this system cannot be used for items with unpredictable subcategorization frames, such as verbs. We plan to attempt a port for these items, but we anticipate that there will be mismatches in the Japanese and Korean equivalents of certain verbs due to translational mismatches, and so other methods of creating a lexicon for these Korean verbs will be needed.

Thus, by using a Japanese-Korean dictionary to translate the head words in the lexicon and a large finite-state morphology, a detailed lexicon can be semi-automatically created for Korean.

[3] The Japanese grammar could in principle also use a cascade of transducers similar to the other ParGram grammars. However, the availability and high accuracy of the ChaSen tokenizer and part-of-speech tagger made it appealing for the Japanese grammar to use a different system design. See (Masuichi and Ohkuma, 2003) for details of the system architecture.

[4] As this project scales up, we will most likely exchange our simple transducer for a larger, commercially available transducer based on the same technology and also compatible with the XLE system.

3.3 Sublexical and Phrase Structure Rules

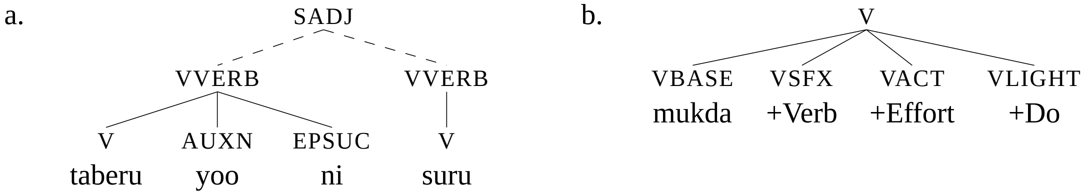
In addition to these lexicon and tokenization and morphology preprocessing steps, some minor changes to the core annotated phrase structure rules were needed. Most of these occurred in the domain of suffix syntax. That is, they are a result of *morphosyntactic* differences between Japanese and Korean that require modification to the c-structure rules. For example, the Japanese grammar allows both orders for the location suffix in conjunction with the topic suffix. However, in Korean, only the order location-suffix followed by topic-suffix is possible (e.g., *eh nun*); thus, the rule had to be further constrained for the Korean grammar. There were a number of places in the grammar where this type of minor, but significant, change was made in the ordering restrictions on inflectional morphemes (be they affixes or independent words).

In the clausal domain, a significant effort will be necessary for expanding (or collapsing) the Japanese rules according to the verbal morphology of Korean. The current Japanese grammar requires a certain amount of morphological preprocessing for the input string to parse. This is done via the ChaSen tokenizer. For instance, *taberuyoonisuru* ‘make effort to eat’ must be tokenized as *taberu yoo ni suru* before it is presented to the parser. In this case, *suru* maps one-to-one to the *+Do* tag which the Korean morphology analyzer produces, allowing the grammar writer to port the lexical entry for *suru* to the (sub)lexical entry for *+Do*. However, a one-to-one correspondence is not always guaranteed, and rule changes are inevitable. Artificially creating dummy tags or collapsing tags in the morphological analyzer to ensure a one-to-one correspondence is undesirable as it would not accurately reflect the morphological information embedded in the inflected Korean verb. Consider the Japanese and Korean pair in (10) for the equivalent of the English *make effort to eat*.

(10)	a.	Japanese:	surface form:	taberuyoonisuru
			morphology breakdown:	taberu yoo ni suru
			meaning:	make effort to eat
	b.	Korean:	surface form:	mukuryo hada
			morphology breakdown:	mukda +Verb +Effort +Do
			meaning:	make effort to eat

The difference in the structure of the languages for this construction results in a need for different c-structure trees. These are seen in (11) (the dashed lines indicate that some intervening nodes are not included; these nodes are relevant for preverbal argument attachment and not for the verbal complex itself).

(11)



The phrase structure rules so far have only required modification for the differences in how certain morphological endings are encoded. So, the majority of the changes have affected sublexical rules, although more major changes to the verbal complex were also required. We suspect that as the grammar port moves beyond core syntactic structures to more “peripheral” ones, there will be further changes to the annotated phrase structure rules that involve more than morphosyntactic differences. In particular, Korean allows adverbial sentential negation as well as the suffixal negation found in Japanese and allows certain double accusative constructions which are impossible in Japanese. We hope to be able to report on these in the near future.

4 Conclusion

We are encouraged by our success in this preliminary investigation. With only two man-months of effort, we found that major parts of the Japanese LFG grammar could be carried over unchanged into a grammar of Korean. Many of the core annotated phrase structure rules remain the same, and it seems that many lexical items can be ported merely by changing the head-word of the entry to its Korean equivalent. New finite-state machines for tokenization and morphological analysis had to be created and incorporated into the system, as was to be expected. Of course, much work needs to be done to bring Korean coverage up to the level of the Japanese grammar. Apart from a substantial amount of testing that needs to be done, this work will focus on peripheral syntactic rules and expansions to both the lexicon and morphology. But more generally, we conclude from this limited experiment that porting grammars across typologically similar languages is an effective method for bootstrapping grammar development.

Acknowledgements

The Japanese grammar was written by Hiroshi Masuichi and Tomoko Ohkuma of the Fuji Xerox Corporation. We gratefully acknowledge their allowing us to use their grammar in our investigation and their assistance in helping us to understand its properties. We also thank them for commenting on earlier drafts of this paper.

References

Beesley, Kenneth and Lauri Karttunen. 2003. *Finite-State Morphology: Xerox Tools and Techniques*. CSLI Publications, Stanford, California.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar project. In *COLING 2002: Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Butt, Miriam, Tracy Holloway King, Maria-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications, Stanford, California.

Dalrymple, Mary. 2001. *Lexical-Functional Grammar*. Academic Press, New York. Syntax and Semantics, volume 34.

Frank, Anette. 1999. From parallel grammar development towards machine translation. In *Proceedings of MT Summit VII*, pages 134–142.

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

Masuichi, Hiroshi and Tomoko Ohkuma. 2003. Constructing a practical Japanese parser based on Lexical-Functional Grammar. *Journal of Natural Language Processing*, 10(2). To appear; in Japanese.

Maxwell, III, John T. and Ronald M. Kaplan. 1993. The interface between phrasal and functional constraints. *Computational Linguistics*, 19:571–589.

Riezler, Stefan, Tracy Holloway King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proceedings of the ACL*.

Sohn, Ho-min. 1999. *The Korean Language*. Cambrige University Press, Cambridge. Chapter 2.4.