Corpus-Based Modality Generation for Korean Predicates

DONG UN AN and GIL CHANG KIM
Korea Advanced Institute of Science and Technology, Taejon, Korea

JONG HYEOK LEE
Pohang University of Science and Technology, Pohang, Korea

Abstract

This paper describes a corpus-based modality generation of a Korean synthesizer. Modalities may be expressed by modality morphemes such as auxiliary verbs and verb endings. To form a complete predicate, they are concatenated together with a main-verb stem, being arranged in the Korean-specific modality order, which is neither a linear order nor a partial order mathematically. To lexicalize a modality, the synthesizer must choose the best one among several different morpheme candidates whose meanings are very similar to one another, since each of them shows a subtle difference from the others as far as stylistic naturalness is concerned. To cope with these difficulties, a corpus-based modality generation is suggested, where a large corpus is analysed to acquire reliable linguistic knowledge on modalities. Through the corpus analysis, we derive a linear modality order covering as much actual ordering information as possible, and also select a representative morpheme for each modality. Finally, by performance evaluation, we show that the corpus-based approach may be a great help to the improvement of the conventional rule-based Korean synthesizer.

1. Introduction

A Korean synthesizer has been developed in a joint research project between KAIST and NEC, based on the Japanese synthesizer of PIVOT E/J which is an English-to-Japanese machine translation (MT) system of NEC (Murakai, 1986; Ichiyama, 1989). The PIVOT E/J has been built under the interlingua strategy, so the Korean synthesizer could be easily integrated into the PIVOT E/J/K, covering English, Japanese, and Korean (Lee *et al.*, 1991).

A *modality* may be interpreted as a psychological attitude on the part of the speaker toward an event, an action, or a state. In this paper, its scope is extended so that it can cover any extra meanings attached to a main verb such as modal, passive, causative, tense, aspect, negation, honorific, politeness, and mood. In most sentences, since one or more modalities are morphosyntactically reflected in a predicate, their correct analysis and generation may have a great influence on the high quality of translation. In the PIVOT E/J/K, the modality information analysed from a source sentence is represented by modality features called 'auxiliary concept CPs (Conceptual Primitives)' in an interlingua (Murakai *et al.*, 1989). One of the major issues of language synthesis is how to naturally lexicalize the modality features into surface morphemes, words, or clauses in a target sentence.

It is well-known that Korean is very similar to Japanese from typological and grammatical viewpoints. Therefore, the Korean synthesizer can use almost the same knowledge bases and sentence-generation methodologies as the Japanese synthesizer in the PIVOT E/J/K. In spite of the similarities, however, Korean has its own characteristics of modalities which have to be reflected in the Korean synthesizer. Modalities are too diverse and even too complex to be handled in the Korean synthesizer which is a conventional rule-based system.

In this paper, major characteristics of Korean modalities are described together with some problematic points of modality generation in Section 2. After analysing a large corpus of Korean sentences to acquire linguistic knowledge of Korean modalities in Section 3, we suggest a corpus-based modality generation in the Korean synthesis system in Section 4. In Section 5, we summarize experimental results of its performance evaluation which was carried out using a small set of test sentences (or their conceptual structures) embodying various modal constructions, and then we draw a conclusion. To represent Korean alphabets, the Yale Romanization is used in this paper.

2. Characteristics of Korean Modalities

From the viewpoint of morphological typology, Korean is an agglutinative language in that the boundaries between morphemes in a word may be clear-cut. From the viewpoint of word-order typology, it is a flexible word-order SOV language, which has the following universal tendencies: plenty of inflectional suffixes, auxiliary verbs after main verb, postpositions instead of prepositions, and so on. In general, these typological and grammatical characteristics may be shared by other languages of the same class as Korean. However, Korean has its own characteristics, especially on modalities, even compared with Japanese which is known to be the most grammatically similar language to Korean. This is the reason why the Korean Synthesizer can not adopt the same method of modality generation as in the Japanese version. In this Section, after describing major characteristics of Korean modalities along with a contrastive analysis of Korean and Japanese, we raise some problematic points of Korean modality generation.

Correspondence: Dong Un An, Department of Computer Science, Korea Advanced Institute of Science and Technology, 373–1, Kusong-dong Yusong-gu, Taejon, 305-701, Korea. E-mail: duan@csone.kaist.ac.kr

Table 1 A structure of a Korean predicate

Predicate	Stem	(Extended) Auxiliary Verbs				Verb-endings			
Modality		PASSIVE	ASPECT	MODAL	NEGATION	HONORIFIC	TENSE	POLITENESS	MOOD
Appearance		optional	optional	optional	optional	optional	obligatory	optional	obligatory
Example	*cap*	*hi*	*e iss*	*ko siph*	*ci anh*	*u-si*	*ess*	*sup*	*ni-ta*
	catch	passive	perfect	hope	negative	honorific	past	politeness	declarative
	cap-hi-e iss-ko siph-ci anh-u-si-ess-sup-ni-ta								

Characteristic 1: A predicate can be formed by concatenating a main-verb stem and a series of 'modality morphemes'.

Agglutinative languages, like Korean and Japanese, are defined as languages whose (inflectional) morphology is wholly concatenative, and where fairly long words can be formed by a series of morphemes concatenated together. So a predicate can be formed by the concatenating together a main-verb stem and several '*modality morphemes*'. In this paper, the term '*modality morpheme*' stands for any morpheme which adds an extra meaning to its main verb; for example, an (extended) auxiliary verb, a pre-final verb ending, and a final verb ending. Table 1 shows that a word formation of complex predicate is *purely concatenative*; in addition, all modality morphemes including auxiliary verbs follow their main-verb stem.

Characteristic 2: Most modalities can be expressed by (extended) auxiliary verbs, pre-final verb endings, and final verb endings.

In Korean, some modalities such as *passive, causative, aspect, modal*, and *negation* are syntactically expressed by (extended) auxiliary verbs. And the other modalities are morphologically expressed by verb endings: *mood* by a final verb ending, and the others such as *honorific, tense*, and *politeness* by pre-final verb endings. On the other hand, in Japanese, verb endings have no special modal meanings, so some modal particles are used instead.

Characteristic 3: For a predicate formation, modalities should be arranged in the Korean-specific modality order.

When a main-verb stem and its modality morphemes are combined to form a complete predicate, the modalities should be arranged in the Korean-specific modality order. If the modality order is violated, the resulting predicate will be ungrammatical and unintelligible. Judging from the examples of usage, the modality order is neither a total (i.e. linear) order nor a partial order. And there has been no available data on the modality order which is reliable enough to be used in a conventional rule-based Korean synthesizer because there has been no comprehensive survey of Korean modalities so far. To cope with this difficulty, we attempt to derive a linear modality order through the analysis of a large corpus.

Characteristic 4: A modality may be lexicalized into several different modality morphemes, whose meanings are very similar to one another.

Despite similarities, each modality morpheme differs slightly from the others, according to stylistic naturalness in a complete predicate. Therefore, the Korean synthesizer has to make a proper choice among several candidates with similar meanings. To solve this problem, we can use the representative modality morphemes which we selected in advance through the analysis of a large corpus.

Characteristic 5: An auxiliary verb requires that its preceding verb should take a specific type of conjunctive verb ending.

When an auxiliary verb is attached to a main verb (or another auxiliary verb), it requires that its preceding verb should take a specific type of conjunctive verb ending. Furthermore, the conjunctive verb ending between two verbs can be absolutely determined by the following auxiliary verb, not by the preceding one. A verb ending can be clearly separated from its verb stem in Korean, but not in Japanese. This is because the Japanese symbols '*Kana*' are syllabic so that symbols can not be separated into consonants and vowels. Due to the clear-cut separation between verb stems and endings, a Korean auxiliary verb can be defined together with its conjunctive verb ending. That is, an (extended) auxiliary verb can be of the form [*conjunctive-ending* + ' ' + *auxiliary-verb-stem*], where a space ' ' is needed for a word boundary in Korean. By virtue of the new definition of an auxiliary verb, a major operation of predicate formation becomes simple concatenation. Table 1 shows that, in a predicate formation, several auxiliary verbs can be easily combined by simple concatenations, notwithstanding the type of their preceding verb endings.

In summary, Korean has its own characteristics of modalities, even compared with Japanese, which have to be fully reflected in a Korean synthesizer. Modalities may be expressed by modality morphemes such as auxiliary verbs and verb endings. To form a complete predicate, the modality morphemes are concatenated together with a main-verb stem, arranged in the Korean-specific modality order. To lexicalize a modality, there may be several different modality morphemes whose meanings are very similar to one another. Therefore, the Korean synthesizer has to select the best one among them. To cope with these difficulties, a corpus-based modality generation is suggested, where a large corpus is analysed for deriving reliable linguistic knowledge on modalities.

3. Analysis of Corpus

In general, the corpus-based approaches in machine translation research may be classified into two groups (Hutchins, 1993):

(i) the direct use of information derived from corpora for the analysis, transfer, and generation of machine translation.

(ii) the indirect use of corpora as sources of information for deriving or compiling linguistic knowledge.

Our corpus-based approach of modality generation can be regarded as the second group. That is, since the Korean synthesizer have been basically developed as a conventional rule-based system, an analysis of a large corpus is needed for the Korean synthesizer to provide reliable data on modalities. The corpus analysis has two major tasks: one is to select a representative morpheme for each modality, and the other is to obtain a linear ordering among the modalities.

To acquire more reliable linguistic knowledge on modalities, the corpus should reflect both the standard Korean usage and the diversity of sentence styles. We choose the following three kinds of textbooks: all the elementary school textbooks (73 volumes), the high-school textbooks on the Korean language (2 volumes), and the high-school textbooks on literature (2 volumes).

As mentioned previously, most modalities can be represented by modal verb endings and auxiliary verbs. There are only a few modal verb endings, and the modality order among them is also well-described in school grammar. On the other hand, there are plenty of auxiliary verbs, whose representatives and modality order have not been fully surveyed yet. Thus, we focus on auxiliaries in the corpus analysis.

3.1 *Representative auxiliary verbs for modalities*

In the corpus, there are 23,981 sentences that contain at least one auxiliary verb. Among them, about 11% sentences have two auxiliaries or more. The total frequency of auxiliary verbs in the corpus in 26,775 as shown in Table 2, and the distinct auxiliaries total 143. Since some of auxiliaries differ only in phonologically-conditioned allomorphs, if we regard them as the same, there are 71 kinds of auxiliary verbs in the corpus.

Although there may be several auxiliary verbs whose modal meanings are very similar to each other, in most cases, each of them shows some difference from the others, especially according to stylistic naturalness in a complete predicate. This means that a completely-formed predicate may become even unintelligible depending on which auxiliary verb is selected for it. To generate stylistically-natural modalities, a representative from each group of similar auxiliary verbs should be determined in advance. The following steps illustrate representative auxiliary verbs were selected:

(a) *Grouping*. First, the 71 kinds of auxiliary verbs were classified into 34 groups according to their modal meanings and grammatical functions. The grain size of the classification was fine enough to map each modality feature into one of the groups without any ambiguities.

(b) *Selection*. After grouping, some groups may contain two or more auxiliary verbs whose meanings are too similar to distinguish from each other. A representative of each group should be able to semantically cover the other auxiliary verbs. In addition, it should stylistically reflect more naturalness than the others. The auxiliary verb, which appears more frequently than the others in the corpus, may be considered to satisfy the above conditions. Therefore, the most-frequently-used auxiliary verb was selected as the representative one.

Table 3 shows the 34 modality groups and their representative auxiliary verbs, which are arranged in the order of frequency. It should be noticed that, in most cases, there are remarkable differences in frequency between representatives and the others. As an exception, the representative '*e peli*' in the '*Completion*'

Table 2 Statistics of auxiliary verbs in the corpus

	Number of sentences										Total frequency
		Number of auxiliary verbs in a sentence									of auxiliary
Textbooks	Total	1		2		3		4			verbs
		freq.	%	freq.	%	freq.	%	freq.	%		
Type I	20,345	18,564	91.2	1,659	8.2	116	0.6	6	0.0		22,254
Type II	1,087	836	76.9	222	20.4	26	2.4	3	0.3		1,370
Type III	2,549	2,009	78.8	482	18.9	54	2.1	4	0.2		3,151
Total	23,981	21,409	89.2	2,363	9.9	196	0.8	13	0.1		26,775

Type I: the elementary school textbooks
Type II: the high-school textbooks on the Korean language
Type III: the high-school textbooks on literature

Table 3 Korean modality groups and their representative auxiliary verbs (in frequency order)

Group	Freq.	Representative auxiliary verb	The other auxiliary verbs	English equivalents
Attempt	8,492	e po(8,492)		try
Progressive	2,725	ko iss(2,196)	e o(313),e ka(174),e naka(43)	be ~ing
Passive	1,694	toy(757),eci(462),i(293),li(84), hi(66),ki(16),pat(11),tangha(6)		~ed
Possibility	1,615	l swu iss(1,615)		can, be able to
Negation	1,592	ci anh(1,156),ci mosha(323), ci mal(113)		not
Change	1,343	key toy(1,343)		came to, become, get
Conclusion	1,237	n kesi(1,221)	n pa iss(12),n seymi(4)	
Service	1,199	e cwu(1,199)		do something for
Need	1,126	eya ha(1,107)	eya toy(16)	must, have to
Perfect	911	e iss(911)		have ~ed
Approval	740	kito ha(740)		really
Expectation	713	l kesi(616)	l thei(96)	will, may, might
Causative	693	key ha(378),sikhi(130),i(98), ki(31),li(23),wu(17),hi(16)		make, have, get
Impossibility	531	l swu eps(531)		cannot
Completion	463	e peli(150)	e nay(160),ko mal(153)	finish
Cause	422	ki ttaymwuni(422)		because
Guess	284	n kes kath(205)	n tus ha(66),nka po(3), l moyangi(8)	seem as if
Keep	241	e noh(148)	e twu(93)	keep
Hope	241	ko siph(225)	ki pala(11),myen ha(5)	wish, want, hope
Intend	199	tyeko ha(156)	koca ha(44)	intend, plan
Concentration	85	l ppwuni(80)	l ttalum(5)	only
Permission	49	myen toy(41)	eto toy(5),eto coh(3)	had better
Inevitableness	44	l swupakkey eps(44)		cannot help ~ing
Disapproval	39	myen an toy(31)	senun an toy(8)	must not
Habit	20	kon ha(20)		used to
Tendency	16	n pyeni(16)		tend to, be apt to
Past possibility	14	l ppenha(14)		come near
Beginning	11	e tul(11)		begin, start
Worthy	11	l manha(8)	m cikha(3)	be worthy of
Pretense	8	n chekha(4)	n cheyha(4)	pretend
Limit	6	l cikyengi(6)		
Situation	5	n thei(5)		
Accent	4	e tay(4)		heavily, hard
Certainty	3	ki malyeni(3)		certain

group was preferred to the others because it has wider semantic coverage in spite of its relatively-low frequency. It is also noticeable that some modality groups such as '*Passive*', '*Causative*', and '*Negation*' do not have their representatives. This is because all the auxiliary verbs of the groups have their own distributive characteristics of usages. That is, the selection of auxiliary verbs for such groups depends on the class of the verb stem or even meaning of the main verb. Therefore, all main verbs should have dictionary information about types of auxiliary verbs for modality groups in advance.

3.2 *Linear modality order*

As shown in Table 2, a sentence may rarely contain three or more auxiliary verbs (less than 1% in frequency), so that such unusual cases may not be useful when deciding a modality order of auxiliary verbs. Instead, the cases with two auxiliary verbs, which are 9.9% in frequency, are far more useful. Since all the auxiliary verbs are classified into 34 modality groups, the ordered pairs among the modality groups total 1156 (34 × 34), among which only 195 ordered pairs appear in the corpus as shown in Table 4.

To derive a corpus-based linear modality order, we define an ordering relation '$\prec$' over the set of all modality groups as follows:

Definition: $\alpha \prec \beta$ if, in the corpus, there exists a sentence in which the modality group α precedes the other β.

Since the ordering relation '$\prec$' is given over some pairs of modalities but not among all of them, it is not a linear ordering. It does not obey the *transitive*, *irreflexive*, and *asymmetric* properties either, so that it can not be even a partial ordering mathematically. Figure 1 shows a typical part of the directed graph representing

Table 4 Frequencies of ordered pairs over modality groups in the corpus

The former	The latter																																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	a	b	c	d	e	f	g	h
A. Passive	7	2			13	129	175				1	1		34		10	77	1	5	33			37	43	1					60	13	14	2	1
B. Causative				3	1	44		41		35	2	7		17			16	1	3	4	2		16	34						20	1	7		
C. Beginning						1																												
D. Keep		1			1	5					1						4		2		2		1	9						4		2		
E. Completion					7	13				1	2	1		15		3	5		8				7	3						18		11		
F. Progressive					2	54								7		9	14						4	7			1			54	25	6	5	
G. Perfect																6	8	1	1	4			4	4						10	10	5	1	
H. Service						51				1	15	3		14		2	17		5	2			36	29			1			26	2	20	3	
I. Accent																																		
J. Attempt						1					13	3		28	1		19	1		2			3	17						3		2		
K. Hope						1				1						1	11						1								8		1	
L. Intend						2											13													7	2	1		
M. Pretense					1									1																				
L. Possibility		13														8	6			78				13						35	22	54	1	
O. Worthy																	1																	
P. Guess																	6						1								2			
Q. Negation		21			3	23		1			1	3		3		4			32	15	2	12		22			1		2	16	14	37	1	
R. Inevitableness																															1			
S. Impossibility																2	2			15						1				13	11	10		
T. Change					1	1	13							2			2						2							45	4	42		
U. Permission																	1															1		
V. Disapproval																														1				
W. Approval																								1						2				
X. Need																	5													9	5	52	1	
Y. Past possibility																																		
Z. Limit																																		
a. Habit																																		
b. Situation																																		
c. Tendency																																		
d. Conclusion																							1								4	1	1	
e. Cause																							1									3		
f. Expectation																															1			
g. Concentration																																		
h. Certainty																																		

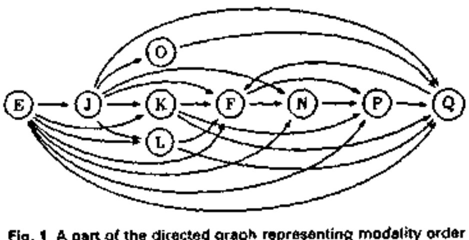

Fig. 1 A part of the directed graph representing modality order (see Table 4 for node symbols)

the ordering relation '≺'. However, we hope to obtain a linear modality order that may cover as much actual ordering information of the corpus as possible. For this, first we attempt to convert the ordering relation '≺' into a partial ordering by removing some ordered pairs, which will have a negligible effect on the modality order, from Table 4. Second, through the well-known topological sorting, we can easily obtain a linear modality order from the partial one.

The corpus-based ordering relation '≺' can be approximately transformed into a partial ordering by the following steps:

(a) *Transitivity*. Since most linguists often assume that the modality order is *transitive*, we also follow the same assumption.

(b) *Irreflexivity*. From the viewpoint of modality generation, we do not have to consider the case in which the same modality feature appears more than once in an interlingua. So, we can remove the three ordered pairs like (*Passive, Passive*), (*Completion, Completion*), and (*Progressive, Progressive*) for irreflexivity.

(c) *Asymmetry*. This property, together with Transitivity, requires that there should be no cycles in the directed graph illustrating Table 4. So, from each cycle, we basically remove a single arc (that is, an ordered pair) whose frequency is low enough to ignore so as to make the graph acyclic. However, the arc removal in some cycles may be carried out from the viewpoint of machine translation. Although a negative rhetorical question is syntactically a negative sentence, it is semantically a strong positive assertion. This means that a negative rhetorical question may be represented in an interlingua without *Negation* feature. As a result, we could remove five arcs ending in *Negation* which only appear in rhetorical questions.

Table 5 Partial ordering relation ' ≺ ' over Korean modality groups

The former	The latter																																	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	a	b	c	d	e	f	g	h
A. Passive	ir	≺			≺	≺	≺				≺	≺		≺		≺	≺	≺	≺	≺			≺	≺	≺					≺	≺	≺	≺	≺
B. Causative				≺	≺	≺		≺		≺	≺	≺		≺			≺	≺	≺	≺	≺		≺	≺						≺	≺	≺		
C. Beginning						≺																												
D. Keep		as			≺	≺					≺						≺		≺		≺		≺	≺						≺		≺		
E. Completion					ir	≺				≺	≺	≺		≺		≺	≺		≺				≺	≺						≺		≺		
F. Progressive					as	ir								as		≺	≺						≺	≺			≺			≺	≺	≺	≺	
G. Perfect																≺	≺	≺	≺	as			≺	≺						≺	≺	≺	≺	
H. Service						≺				≺	≺	≺		≺		≺	≺		≺	≺			≺	≺			≺			≺	≺	≺	≺	
I. Accent																																		
J. Attempt						≺					≺	≺		≺	≺		≺	≺		≺			≺	≺						≺		≺		
K. Hope						≺				as						≺	≺						≺								≺		≺	
L. Intend						≺											≺													≺	≺	≺		
M. Pretense					≺									≺																				
N. Possibility		as														≺	as			≺				≺						≺	≺	≺	≺	
O. Worthy																	≺																	
P. Guess																	≺						≺								≺			
Q. Negation		as			as	as		as			as	as		≺		as			≺	≺	≺	≺		≺			≺		≺	≺	≺	≺	≺	
R. Inevitableness																															≺			
S. Impossibility																as	as			≺						≺				≺	≺	≺		
T. Change					as	as	as							as			as						≺							≺	≺	≺		
U. Permission																	as															≺		
V. Disapproval																														≺				
W. Approval																								≺						≺				
X. Need																	as													≺	≺	≺	≺	
Y. Past possibility																																		
Z. Limit																																		
a. Habit																																		
b. Situation																																		
c. Tendency																																		
d. Conclusion																							as								≺	≺	≺	
e. Cause																							as									≺		
f. Expectation																															as			
g. Concentration																																		
h. Certainty																																		

Types of removal: ir(irreflexivity), as(symmetry).

Table 6 Summary of the removed arcs (i.e. ordered pairs)

Type of removal		Number	Arcs
Irreflexivity		3	(passive,passive) (completion,completion) (progressive,progressive)
Asymmetry	Rhetorical Question	5	(possibility,negation) (impossibility,negation) (change,negation) (permission,negation) (need,negation)
	Low Frequency	21	(keep,causative) (progressive,completion) (progressive,negation) (perfect,change) (hope,attempt) (possibility,causative) (negation,causative) (negation,completion) (negation,service) (negation,hope) (negation,intend) (negation,guess) (change,possibility) (conclusion,approval) (expectation,cause) (negation,progressive) (impossibility,guess) (change,completion) (change,progressive) (change,perfect) (cause,approval)
Total		29	

Table 7 A linear order over Korean modality groups

Order	Group	Auxiliary concept CP
1	Situation	
2	Pretense	
3	Accent	
4	Beginning	XBEGIN
5	Passive	XPASS
6	Certainty	
7	Past possibility	XNEAR
8	Perfect	PREF
9	Causative	XCAUS
10	Service	
11	Keep	
12	Completion	XFINISH
13	Attempt	XTRY
14	Inevitableness	XNECE
15	Worthy	XESTIM
16	Intend	XPLAN
17	Hope	XWISH
18	Progressive	PROG
19	Possibility	XPOSS
20	Guess	XSEEM
21	Negation	NEG
22	Tendency	XTEND
23	Habit	XRULE
24	Disapproval	NEG∧XMUST
25	Permission	XRECO
26	Impossibility	NEG∧XPOSS
27	Limit	
28	Change	XCHAN
29	Approval	
30	Need	XNEED
31	Conclusion	XCONCLU
32	Concentration	XONLY
33	Cause	XREA
34	Expectation	XINFE

After the arc removal steps, a partial ordering among modality groups can be obtained as shown in Table 5, where the removed arcs are marked as 'ir (irreflexivity)' or 'as (asymmetry)'. Table 6 sums up all of the removed arcs. Finally, the partial ordering can also be embedded in a linear order through the so-called topological sorting. Table 7 shows the linear order over Korean modality groups, together with the corresponding 'auxiliary concept CPs' of the PIVOT interlingua.

4. Corpus-based Modality Generation

In PIVOT E/J/K, the Korean synthesizer generates a surface sentence from a language-independent conceptual structure, which may be the analysis result of either an English or a Japanese sentence. The sentence generation is carried out by the following three phases (Kim *et al.*, 1988; Lee *et al.*, 1991). First, the phase one (sentence-structure planner) transforms the language-independent conceptual structure into a Korean-dependent semantic structure so that the target sentence to be generated will be pragmatically and stylistically much more natural. Then, for each node of the semantic structure, the phase two (syntactic generator) determines its syntactic and word-order information. A grammatical structure is produced as an output. Finally, using the word-order information, the phase three (morphological generator) arranges all nodes of the grammatical structure in a linear order. The nodes are, then, lexicalized into surface morphemes or words.

In this section, under a rule-based system architecture of the Korean synthesizer, we propose a corpus-based modality generation for Korean predicates. Figure 2 shows the overall processing flow of the Korean synthesizer, focusing on the modality generation. The synthesizer makes use of two knowledge bases obtained through the corpus analysis to access the data on modalities: one is the modality ordering table which represents a linear order over modality features (i.e., auxiliary concept CPs of PIVOT interlingua), and the other is the modality lexicalizing table through which each modality feature can be lexicalized into a representative auxiliary verb. The connectivity checking table, which represents the morphological dependency between neighbouring morphemes, is used to select a correct modality morpheme among several phonologically-conditioned allomorphs.

In the following, we concentrate our description on the stages needed for the modality generation. Among them, the first and the second stages are performed in the sentence-structure planning and the syntactic generation phases, respectively. The others are carried out in the morphological generation phase.

(a) *Activo-passivization.* If a verb does not have a morpho-syntactic passive form, but instead there is a verb that represents the same activity in an opposite direction, then the sentence-structure planner changes the related CP and deep cases of the verb for *activo-passivization* (Lee and Kim, 1988). Then, the passive voice should be altered into the active, because the target sentence will be morpho-syntactically active, although it is semantically passive.

(b) *Agreement checking.* In Korean, there are various kinds of agreements which are closely related to modalities; for example, politeness, honorific, tense, and mood agreements. In the syntactic generation phase, the related modality features may be modified or added according as the agreements are checked.

(c) *Modality ordering.* When two or more modality features are contained in a predicate, the morphological generator can arrange them in the linear order specified by the modality ordering table.

(d) *Decision of passive/causative forms.* Some Korean verbs are formed with a verb stem and a morphologically conditioned passive/causative morpheme, which has several allomorphs, one of which should be chosen correctly according to the class and even the meaning of its verb stem.

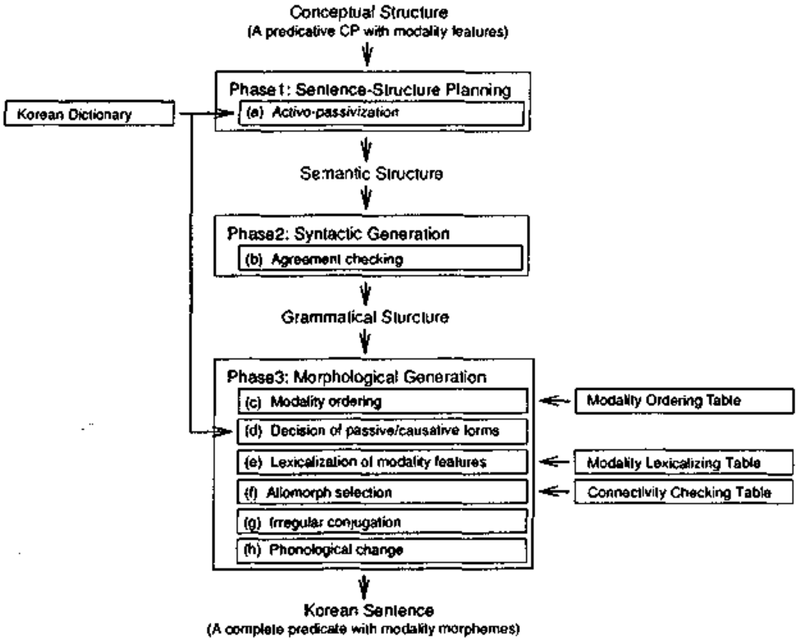
Conceptual Structure
(A predicative CP with modality features)
Phase1: Sentence-Structure Planning
(a) Activo-passivization
Korean Dictionary
Semantic Structure
Phase2: Syntactic Generation
(b) Agreement checking
Grammatical Sturcture
Phase3: Morphological Generation
(c) Modality ordering
(d) Decision of passive/causative forms
(e) Lexicalization of modality features
(f) Allomorph selection
(g) Irregular conjugation
(h) Phonological change
Modality Ordering Table
Modality Lexicalizing Table
Connectivity Checking Table
Korean Sentence
(A complete predicate with modality morphemes)

Fig. 2 Overall processing flow of Korean synthesizer

That is, since the passive and causative modality groups have no representative auxiliary verbs, the morphological generator must select a proper form by using the lexical information of each verb.

(e) *Lexicalization of modality features.* Each modality feature in a predicate can be lexicalized into a (representative) modality morpheme by means of the modality lexicalizing table.

(f) *Allomorph selection.* Since a modality morpheme may have several phonologically-conditioned allomorphs, the morphological generator has to select a correct one depending on whether its preceding morpheme ends is a consonant or a vowel, and, in some cases, whether the sound quality of the last vowel of its preceding morpheme is clear or dark. The allomorph selection can be carried out by checking the morphological connectivity between two neighbouring morphemes, where the connectivity checking table is used.

(g) *Irregular conjugation.* In Korean, there are 12 types of irregular conjugations for verbs (including auxiliary verbs). When a verb stem and a verb ending are combined with each other in the way of irregular conjugation, the last and the first syllables of the stem and the ending, respectively, may be omitted, changed, or added depending on the type of irregular conjugation.

(h) *Phonological change.* The morphological generator handles phonological phenomena such as vowel contraction and vowel omission, except for vowel harmony. The vowel harmony can be handled easily by the connectivity checking of the stage of allomorph selection.

5. Evaluation

To evaluate the quality of the corpus-based modality generation for Korean predicates in PIVOT E/J/K, some experiments were carried out using a small set of test sentences (or their corresponding conceptual structures). However, careful attention was given to the selection of test sentences to reflect various modal construction. The quality of generated modalities turned out to be much better than before (see Appendix). There were no cases where the modality order was violated, so that the generated predicates became intelligible and grammatical. This means that the corpus-based linear order over modalities is quite satisfactory for the Korean synthesizer. It may be safe to say that most of representative auxiliaries have been well selected for modality features, evaluating the extent of their naturalness in the generated predicates. This improvement is due to the much more reliable data on modalities which is obtained through an analysis of a large corpus, and also due to the systematic approach of modality generation.

However, the representative auxiliary verbs for '*Completion*', '*Hope*', and '*Negation*' are sometimes unnatural in specific constructions. This is because they have been selected mainly on the basis of frequency of use, and also because there must be some kind of selection restriction for them which has not yet been studied in full. Our further research will be concentrated on elaborating selection restrictions of Korean modal auxiliaries based on semantics.

6. Conclusion

Korean modalities are too diverse and there had not been any available data on modalities which are reliable enough to use in a conventional rule-based Korean synthesis system. A large corpus was analysed to acquire much more reliable knowledge on modalities. Firstly, for each group of auxiliary verbs whose modal meanings are very similar to one another, the representative for each group was selected mainly on the basis of frequency in the corpus. Secondly, the corpus-based ordering relation among a set of modality groups was transformed into a partial ordering by removing some ordered pairs, and then further into a linear ordering through the topological sorting. On the basis of the derived linguistic knowledge of Korean modalities, we suggest a corpus-based modality generation in the Korean synthesis system. A performance evaluation was carried out using a small set of test sentences (or their conceptual structures) embodying various modal constructions. On the whole, the quality of generated modalities turned out to be much better than before, and we believe that the corpus-based approach is a great help to the improvement of the conventional rule-based Korean synthesizer. Since there still remain a few cases where the frequency-based representatives result in unnatural expressions, we will concentrate our further research on elaborating semantics-based selection of Korean modal auxiliaries.

References

Hutchins, J. (1993) Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research. *MT Summit IV*, pp. 11–34.

Ichiyama, S. (1989) Multi-lingual Machine Translation System. *Office Equipment and Products*, 18.131: 46–48.

Kim, C. H., Choe, B. H., Kim, G. C., Choi, K. S. and Ichiyama, S. (1988) Generation of Korean from Conceptual Representation. *Proc. 37th Information Processing Society of Japan*, pp. 947–948.

Lee, J. H. and Kim, G. C. (1988) Voice Generation from Conceptual Representation: Syntactic, Semantic, and Pragmatic Aspects. *Literary and Linguistic Computing*, 3: 250–254.

——, Okumura, A., Muraki, K. and Kim, G. C. (1991) An English–Korean Machine Translation System: Korean Synthesis under the Environment of Japanese Generation System, *Proc. 2nd Japan–Australia Joint Symposium on Natural Language Processing*, Iizuka, Japan, pp. 219–224.

Muraki, K. (1986) VENUS: Two-phase Machine Translation System. *Future Generations Computer Systems*, 2: 117–119.

——, Kamei, S. and Nomura, N. (1989) The Interlingua for a Machine Translation System. *SIGNLP Information Processing Society of Japan*, 89. 54: 99–106.

Appendix

Some examples of improvement

(a) *Activo-passivization*

Because some verbs do not have any morpho-syntactic passive verbs for the passive construction, the auxiliary passive transformation comes to be unacceptable in *Initial Result*. There are some pairs of verbs which express activities in opposite direction, so the opposite contrastive active verb can take the place of passive verb in *Improved Result*.

a.1

Input Sentence (E) I was told about how to solve this problem.

Initial Result (K) nanun i mwunceylul ettehkey phununkaey kwanhaye *malhaye cyesssupnita*. (*)

Improved Result (K) nanun i mwunceylul ettehkey phununkaey kwanhaye *ulesssupnita*.

a.2

Input Sentence (E) They were asked if they had gone out of the city.

Initial Result (K) kutulun kutuli tosi oypwuey kassnunka ttonun ettencilul *mwule cyesssupnita*. (*)

Improved Result (K) kutulun kutuli tosi oypwuey kassnunka ttonun ettencilul *cilmwunpatasssupnita*.

a.3

Input Sentence (E) He was taught how to borrow a book from the library.

Initial Result (K) kunun chaykul tosekwanpwuthe ettehkey pillinunkalul *kaluchye cyesssupnita*.(*)

Improved Result (K) kunun chaykul tosekwanpwuthe ettehkey pillinunkalul *paywesssupnita*.

(b) *Compound passive*

When the noun N_t of a transitive denominal verb 'N_t + *ha*' has the meaning of adversity or beneficiary, the verbalizer passive 'N_t + *toy*' comes to be unacceptable in *Initial Result*. The compound passive 'N_t + *pat* ('receive')' makes a good expression in *Improved Result*.

b.1

Input (E) He was ordered to fight.

Initial Result (K) kunun kongkyekhanun kesul *myenglyengtoyesssupnita*. (*)

Improved Result (K) kunun kongkyekhanun kesul *myenglyengpatasssupnita*.

b.2

Input Sentence (E) I was begged to sit down.

Initial Result (K) nanun ancnun kesul *pwuthaktoyesssupnita*. (*)

Improved Result (K) nanun ancnun kesul *pwuthakpatasssupnita*.

(c) *Representative auxiliary verbs for modalities*

The auxiliary verb of *Improved Result* is the representative auxiliary verb obtained through the analysis of corpus. The sentence of *Improved Result* is more stylistically natural and compact than the sentence of *Initial Result*.

c.1

Input Sentence (E) He has to go to hospital tomorrow.
#MODAL(XNEED)

Initial Result (K) kunun nayil pyengweney *kal philyoka isssupnita*.

Improved Result (K) kunun nayil pyengweney *kaya hapnita*.

c.2

Input Sentence (E) I intend to learn Korean this month.
#MODAL (XPLAN)

Initial Result (K) nanun ipen tal hankwukelul *ikhinun kesul uytohako isssupnita*.

Improved Result (K) nanun ipen tal hankwukelul *ikhilye hapnita*.

c.3

Input Sentence (E) Because I like to study.
#MODAL(XREA)

Initial Result (K) nayka kongpwuhanun kesul *cohahako issumulo*.

Improved Result (K) nayka kongpwuhanun kesul *cohahaki ttaymwunipnita*.

c.4

Input Sentence	(E) I could not help laughing when I saw that drama.
	#MODAL(XNECE)
Initial Result	(K) nanun nayka ku tulamalul pon ttayeynun *wusci anhul su epssupnita.*
Improved Result	(K) nanun nayka ku tulamalul pon ttayeynun *wusul su pakkey epssupnita.*

c.5

Input Sentence	(E) I used to see him often.
	#MODAL {XRULE)
Initial Result	(K) nanun kulul cacwu *pokilo hako issesssupnita.*
Improved Result	(K) nanun kulul cacwu *pokon hayesssupnita.*

c.6

Input Sentence	(E) I tend to go to mountain every weekend.
	#MODAL{XTEND)
Initial Result	(K) nanun cwumalmata saney *kal kyenghyangi isssupnita.*
Improved Result	(K) nanun cwumalmata saney *kanun pyenipnita.*

c.7

Input Sentence	(E) His behavior is worthy of praise.
	#MODAL{XESTIM}
Initial Result	(K) kuuy hayngtongun *chingchanhal kachiisssupnita.*
Improved Result	(K) kuuy hayngtongun *chingchanhal manhapnita.*