論文 88-25-12-18

구문구조 Matrix에 의한 한국어의 수식구조와 개념구조의 해석

(An Analysis of the Korean Modificatory and Conceptual Structure by a Syntactic Matrix)

韓 光 錄*, 崔 璋 善**, 李 柱 根*

(Kwang Rok Han, Jang Seon Choi and Joo Keun Lee)

要 約

본 논문은 자연어 이해 시스템의 설계를 위하여 한국어의 구문 구조를 해석하는 방법에 대하여 논한다.

한국어가 갖는 구조적 특징을 이용하여 구문구조의 매트릭스를 유도한다. 이 매트릭스로부터 수식구조와 개념구조를 추출하고, 이들 해석과정에서 구 구조, 절 및 개념구조를 추출함으로서 일반문으로부터 술어논리의 표현을 제시해 주며, 이것은 일반문에 대한 지식 베이스의 구축과 추론의 가능성을 제시해 준다.

Abstract

This paper deals with an analyzing method of the Korean syntax to implement a natural language understanding system.

A matrix of the syntactic structure is derived by the structural features of the Korean language. The modificatoty and conceptual structures are extracted from the matrix and the predicate logic form is expressed by extracting the phrase, clause and conceptual structure in the analyzing process. This logic form constructs an knowledge base of the sentence and proposes the possibility of the inference.

*正會員, 仁荷大學校 電子工學科
(Dept. of Elec. Eng., Inha Univ.)

**正會員, 金星社 情報機器研究所
(Information System R & D Lab., GoldStar Co., Ltd.)

接受日字 : 1988年 7月 5日

I. 서 론

최근 지식정보에 의한 자연어의 이해시스템은 급속한 발전을 하고 있다. 지식은 자연어에 의하여 표출되므로 컴퓨터에 의한 자연어의 처리는 기계번역 시스템, 각종 전문가 시스템의 핵을 이룬다.[1~7]

그러나 자연어는 지극히 다양하고 복잡하며 광역에 펼쳐 있으므로 컴퓨터에 의하여 구문, 의미, 문

맥 등을 해석하여 지식베이스를 구축하는 문제는 용이한 것이 아니다.

자연언어의 구문해석은 1970년대에 많이 연구되고, 상당히 대규모의 문법도 만들어졌다. 그 가운데에서 Chomsky의 표준형으로 구성된 문장을 3각행렬에 의하여 해석하는 Kasami-Younger의 방법이 있으나, 이 방법은 한 행에 2개 이상의 문이 들어오면 본질적으로 그 수만큼 애매한 문이 존재할 수 있다.[13]

그리고 W. Woods의 ATN은 이론적으로 뛰어난 특징을 갖고 널리 사용되어 왔다. 이것은 입력문의 단어열에 맞지 않는 곳에서는 다시 backtracking을 하는 방법을 취하고 있다.[14] 또한 일본어에서는 격 frame을 이용하는 경우가 많고,[16] LFG 문법 이론을 일본어의 기계번역에 적용한 예도 있다.[15]

그러나 이와 같은 방법들은 한국어와 일본어와 같이 용언을 기준으로 전치하는 체언의 구성성분의 어순이 자유롭고 파생어가 발달한 언어에서는 쉽게 적용되지 않는다는 결점이 있다.[13] 특히 한국어의 구문해석을 위해서는 한국어 자체의 언어사실로 부터 독자적인 기본틀을 추출해내는 것이 문제가 된다고 생각한다. 그중에서도 한국어의 수식관계는 구문 해석에서 애매성이 생기는 요인의 하나가 되며 대단히 중요하다. 종래에는 주로 통사론에 따른 프로그램에 의하여 수식구조를 해석하는 것이 일반적이었다.

본 연구에서는 한국어의 수식구조를 보다 조직적으로 해석하기 위하여 구 구조 매트릭스의 해석 모델을 제안하고 실험에 의하여 확인하였다.

Ⅱ. 한국어의 구문구조의 해석

자연언어의 처리대상인 문장의 내부 구조를 구문론적인 관점에서 분석할 경우 최소 자립 형식인 단어로 나눌 수 있으며 이 단어는 최소의 의미 단위인 형태소로 분리된다.[10~12] 즉 구문구조 분석에 있어서 최소의 구문요소는 형태소가 되며 자연어 처리 시스템에서는 중요한 대상이 되고 있다.

한국어에 있어서 형태소 분류는 의미형태소(CWD)와 기능형태소(FWD)로 분류되며, 이들은 다시 여러 갈래로 하위 분류된다. 본 논문에서는 품사의 분류는 문헌[10~12]을 재 정비하여, 본 연구의 목적에 부합하도록 저자들이 이미 발표한 일반어의 형태소 해석의 결과를 이용하고,[4] 그 다음 단계의 구문구조의 해석을 새로운 각도에서 해석하려고 한다.

본 논문에서 사용된 기호들은 다음과 같다.

V－용언, V_b－동사, Va－형용사, Nc－서술격 조사, N－체언, Nc－자립명사, Ni－의존명사, Ad－부사, Det－관형사, P－조사, Pc－격조사, Pcs－주격조사, Pco－목적격, Pcc－보격, Pcd－관형격, Pca－부사격조사, Ps－보조사, E－어미, Ee－종결어미, Ec－연결어미, En－명사형 어미 Ed－관형형 어미, Ea－부사형 어미이다.

1. 단위구 구조

일반적으로 문장은 각종 단어의 열로 이루어지고, 이 단어들의 연결은 일정한 규격을 가지고 문장을 구성한다. 종래의 언어학에서 한국어의 구 구조는 동심구성(ENC : endocentric construction)과 이심구성(EXC : exocentric construction)의 두가지 구성형에 기초를 둔 견해가 있다.[10~12] ENC는 그림 1과 같이 구문적 기능으로 보면 두 구성요소가 하나의 요소로 대치되고, EXC는 두 구성요소가 하나의 구성요소로 대치되지 않는 경우이다.

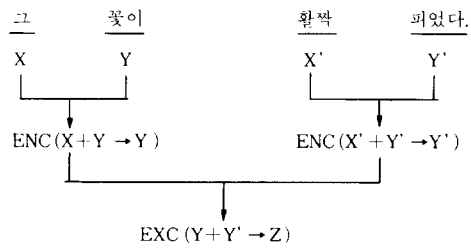


그림 1. 동심구성(ENC)과 이심구성(EXC)의 예
Fig. 1. Example of endocentric and exocentric constructions.

그림 1에서 두 성분 X와 Y가 Y로 대치되면 ENC가 되고, X와 Y가 다른요소 Z로 되면 EXC가 된다. 이와 같은 관점에서 한국어의 문장을 성분구조에 따라 축약하면 단일문이나 절은 '체언+용언'의 EXC와 '용언의 관형형+체언'의 ENC로 집약되며, 구는 여러가지 형태의 ENC로 이루어져 있음을 볼 수 있다.

구의 형성의 관점에서 ENC의 형태를 분류하면 크게 종속성과 등위성으로 나눌 수 있다. 종속성에는 '관형사+체언', '부사+용언', '부사+관형사'의 세 가지로 하위분류할 수 있으며, 등위성은 '체언+체언', '용언+용언', '부사+부사'의 형태로 하위분류를 할 수 있다.

$$\text{ENC}\begin{cases}\text{종속성} : \text{Det}+\text{N},\ \text{Ad}+\text{V},\ \text{Ad}+\text{V},\ \text{Ad}+\text{Det}\\ \text{등위성} : \text{N}+\text{N},\ \text{V}+\text{V},\ \ \text{Ad}+\text{Ad}\end{cases}$$

이들 종속성과 등위성의 관계를 살펴보면 종속성의 'Ad+Det'와 등위성의 'Ad+Ad'는 완전한 구를

이루지 못하고, 이들이 다시 체언(Det+N) 또는 용언(Ad+V)과의 종속성으로써 하나의 완전한 체언구 용언구를 형성하게 된다. 이들의 관계는 표 1에 보인다.

표 1. 종속성과 등위성의 예
Table 1. Examples of endocentric construction.

	동 심 구 성		예
종속 구성	Det+N		새 옷/그 사람
	Ad+V		아주 붉다/매우 빠르다
	Ad+Det		아주 새(책)/바로 그
등위 구성	N+N	NPcd+N	정원의 꽃/과학의 발달
		N+N	학교 설립자/삼국통일
		NPi+N	종이와 먹/책과 연필
	V+V	VEa+V	먹게 되다/읽어 주다
		VEc+V	노래하고 춤추다
	Ad+Ad		좀 더(빠르다)

2. 구 구조 규칙

앞에서 언급한 ENC는 수식어와 피수식어간의 수식구조에 의한 구성이라 할 수 있다. 한국어의 수식구조의 특징은 서구어와는 달리 수식어가 반드시 피수식어의 좌측에 위치하는 전형적인 left-branch 전개를 하고 비교차성이 유지되며, 피수식어가 그 수식구조의 핵심어가 된다.

이러한 수식구조의 특징을 이용하여 문장을 동심구성에 의해 형성되는 단위구로 분리하는 구 구조규칙을 일반화하여 수식적인 함수로 정의한다.

(정 의)

구 구조규칙 함수 fpsr는 다음과 같이 정의한다.

fpsr(Morpl, Morp2) : → Morp3

여기서, Morp1= {cwd|cwd∈N, V, Ad, Det, NP, VE}

Morp2= {fwd|fwd∈P, E, N, V, Ad, Det}

Morp3= {phr|phr∈NP, VE, N, V, Ad, Det, NIL}

이상의 고찰을 기초로 하여 단위구 분리 알고리듬 PSG를 정한다.

〔알고리듬 : PSG〕

단위구 분리 및 ISA-property Xi의 생성

데이타 구조 : PMS(전위 형태소 stack)
PWS(전위 단어 stack)
CMR(형태소 register)
CWR(단어 register)
PINFO(결합 정보 register)
NSFLAG(자립명사 flag)
i(ISA-property count)
j(단위구 count)

입력 : MOSTK(형태소 stack)

출력 : PHSTK*(단위구 stack)
PHSTK(결합 정보 stack)
ISASTK(ISA-property stack)

PROC1 : (초기상태)
1) 모든 stack과 register의 값을 NIL로 세트한다.
2) i와 j의 값을 0으로 세트하고 PROC2 로 간다.

PROC2 : (형태소 입력)
1) NSFLAG를 리세트(reset) 한다.
2) MOSTK로부터 pop하여 첫번째 요소를 CWR에, 두번째 요소를 CMR에 넣는다.
3) CWR과 PWS의 값이 모두 NIL이면 RETURN.
4) CMR의 값이 Nc 이면 NSFLAG를 세트하고 PROC3으로 간다.

PROC3 : (구 구조규칙 처리)
1) fpsr(PMS, CMR)의 값을 PINFO에 넣는다.
2) NSFLAG가 세트상태이면 PROC4로 간다.
3) PINFO의 값이 NIL이면 PROC5로, 아니면 PROC6으로 간다.

PROC4 : (ISA-property Xi 생성)
1) i의 값을 하나 증가시킨다.
2) ISA-property Xi를 형성하여 ISASTK에 push한다.
3) CWR의 값을 'Xi'로 치환한다.
4) PINFO의 값이 NIL이면 PROC5로, 아니면 PROC6으로 간다.

PROC5 : (단위구 분리)
1) PWS의 값과 PMS의 값을 list로 하여 PHSTK*에 push하고, PWS과 PMS를 NIL로 세트한다.
2) PINFO의 값을 PHSTK에 push 한다.
3) j의 값을 하나 증가시키고 PROC6으로 간다.

PROC6 : (단위구 형성)
1) CWR의 값을 PWS에 push 한다.
2) CMR의 값을 PMS에 push하고 PROC2로 간다.

여기서 Xi는 체언개념의 ISA-property 변수로서, 제Ⅳ장에서 자세히 논한다.

그림 2는 알고리듬 PSG에 의하여 입력문장이 단위구로 분리된 결과의 한 예이며, 구 구조 결합규칙이 재귀적으로 적용된다.

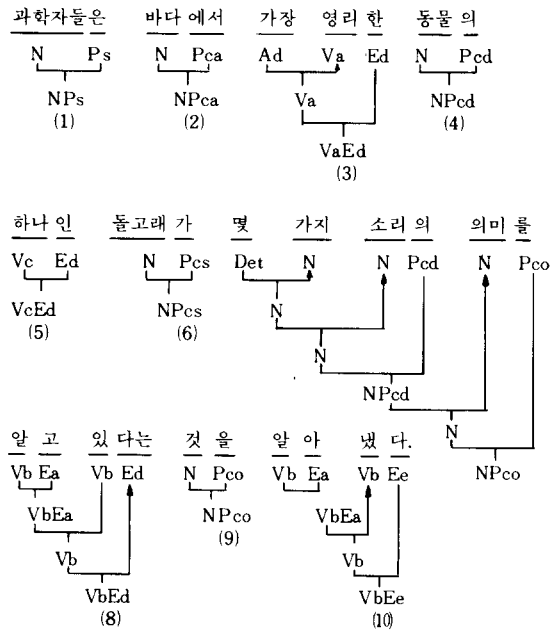


그림 2. 단위구로 분리한 입력문장의 예
Fig. 2. Example of separated phrase units.

3. 문법적 수식 구조

한국어의 문장에 있어서 문법적 수식 구조는 용언이 의미중심이 되어 그의 지배를 받는 체언과 결합하여 형성하는 구조적 결합관계와 용언이 관형어로서 활용되는 체언의 의미적 범위를 수식하는 한정적 결합관계로 이루어진다.[15] 따라서 본 논문에서는 문법적 수식 구조를 구문적 결합관계와 한정적 결합관계로 나누어 해석한다.

1) 구조적 결합관계

구조적 결합관계는 용언구와 그의 지배를 받는 체언구와 결합하여 형성되는 관계로, 용언구를 중심으로 left-branch 전개로 이루어져 있다. 이 구조적 결합관계를 이루는 단위구(용언구, 체언구)는 사고력 내용의 단위인 동시에, 문법형식의 구성요소이다. 따라서 단위구는 문장의 의미를 표현하는 구문 정보를 지닌 것으로, 구문구조 분석의 중요한 대상이 되는 것이다.

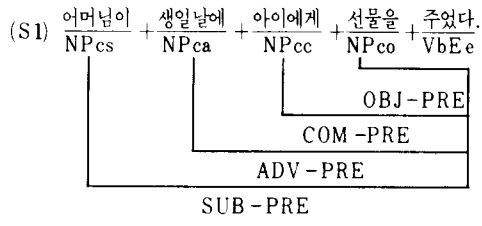
이러한 단위구를 설정하는 방법은 그 대상을 단일 문으로 한정하고, 단어의 구문상의 기능을 형식화하여 몇개의 비슷한 개념으로 묶는 일이다. 그러므로 구조적 결합관계는 이러한 단위구의 설정 방법에 따라 5가지의 관계로 분류한다.[10]

(1) 주술 관계 (주어+서술어)
(2) 객술 관계 (목적어+서술어)
(3) 보술 관계 (보어+서술어)
(4) 부체 관계 (부체어+체언)
(5) 부용 관계 (부용어+용언)

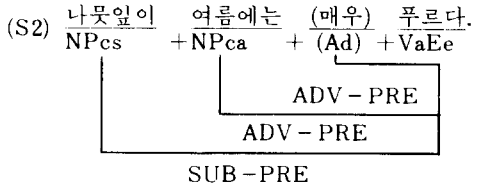
이때, 괄호안의 요소들은 각 단위구들의 문장 성분에 대한 기능을 나타내며, 앞의 성분은 체언구이고, 뒤의 성분은 용언구가 된다.

이와 같이 구문의 구조적 결합관계를 다섯 가지로 등식화 할 수 있으며, 각각의 서술부의 형식에 따라 구조적 결합관계가 성립한다.

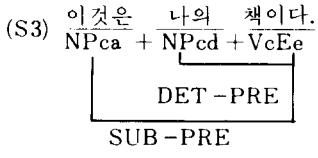
(1) 동사문



(2) 형용사문



(3) 서술격문



여기서 'SUB-PRE'는 주술관계, 'OBJ-PRE'는 객술관계, 'COM-PRE'는 보술관계, 'DET-PRE'는 부체관계, 'ADV-PRE'는 부용관계를 나타낸다. 예문에서 괄호로 묶는 부분은 단위구의 분리과정에서 용언과 결합하여 하나의 단위구로 형성되는 요소이다.

이 밖에 보조사가 첨용된 체언구와 용언구의 구조적 결합관계의 성립 여부는 다음의 예문 (S4)에서와 같이 용언의 활용형태에 따른 제약을 받는다.

(S4) 영수는 나무를 잘 타는 원숭이를 보았다.
NPs NPco Ad VbEd NPco VbEe

SUB-PRE

문장 (S4)에서 보조사 '는'의 체언구 NcPs는 용언의 관형형 VbEd와는 주술관계가 성립되지 않고, '보았다'와는 주술 관계가 성립된다. 이것은 용언이 전성어미로 활용되었을 때는 보조사가 첨용된 체언구와 구조적 결합관계가 성립하지 않는 것을 보이는 것이다.

표 2. 구조적 결합관계
Table 2. Relations of structural cohension.

용언구 \ 체언구		NPcs	NPco	NPcc	NPs	NPca	NPcd	Det	Ad
Va_	Ee	SUB	-	COM	SUB	ADV	-	-	ADV
	Ec	SUB	-	COM	SUB	ADV	-	-	ADV
	En	SUB	-	COM	-	ADV	-	-	ADV
	Ed	SUB	-	COM	-	ADV	-	-	ADV
	Ea	SUB	-	COM	-	ADV	-	-	ADV
Vb_	Ee	SUB	OBJ	COM	SUB	ADV	-	-	ADV
	Ec	SUB	OBJ	COM	SUB	ADV	-	-	ADV
	En	SUB	OBJ	COM	-	ADV	-	-	ADV
	Ed	SUB	OBJ	COM	-	ADV	-	-	ADV
	Ea	SUB	OBJ	COM	-	ADV	-	-	ADV
Vc_	Ee	SUB	-	-	SUB	-	DET	DET	ADV
	Ec	SUB	-	-	SUB	-	DET	DET	ADV
	En	SUB	-	-	-	-	DET	DET	ADV
	Ed	SUB	-	-	-	-	DET	DET	ADV
	Ea	SUB	-	-	-	-	DET	DET	ADV

표 2는 체언구의 첨용형태와 용언구의 활용형태에 따른 구조적 결합관계의 성립 유무를 table로써 나타낸 것으로, "-"는 구조적 결합관계가 성립하지 않음을 뜻한다.

구조적 결합관계를 일반화하여 수식적인 함수로 정의하면 다음과 같다.

(정 의)

구조적 결합관계 함수 frs는 다음과 같이 정의한다.

frs(Phrase1, Phrase2) : → Crs

여기서, Crs = {rel| rel∈SUB, OBJ, COM, ADV, DET, NIL}

Phrase1 = {np|np∈NP, Det, Ad}

Phrase2 = {vp|vp∈VE}

2) 한정적 결합관계

한정적 결합관계는 용언이 관형형으로 활용되었을 때에만 성립되며, 용언구를 중심에 우측에 나타나는 하나의 체언과만 결합관계가 이루어 진다. 그러므로 한정적 결합관계는 ENC로서의 '관형어-체언'의 등식으로 성립된다.

(S5) 아름다운 꽃과 좋은 열매가 많이 달렸다.
VaEd N VaEd N

PRE-MOD PRE-MOD

(S6) 그는 씩씩한 군인이다.
VaEd Vc

PRE-MOD

위의 예문들은 한정적 결합관계가 이루어 지는 경우를 보인 것으로, 'PRE-MOD'는 한정적 결합관계가 성립함을 나타낸다.

한정적 결합관계를 일반화하여 수식적인 함수로정의하면 다음과 같다.

(정 의)

한정적 결합관계 함수 frd는 다음과 같이 정의한다.

frd(Phrase3, Phrase4) : → Crd

여기서, Crd = {rel|rel∈PRE-MOD, NIL}

Phrase3 = {vp|vp∈VEd}

Phrase4 = {np|np∈NP, Vc}

Ⅲ. 구문구조의 매트릭스

1. 구문구조의 매트릭스 표현

문장으로부터 추출한 결합관계는 용언구의 갯수를 m, 단위구의 갯수를 n이라 할때, 그림 3과 같이 m×n 매트릭스의 형태로 표현할 수 있으며, 이것은 구문 구조의 수식 관계를 도출할 수 있다.

일반 문장이 단위구로 분리된 뒤에 용언구를 중심으로 구조적 결합관계와 결합관계를 추출하여 구문구조 매트릭스로 표현하기 위한 알고리듬 MATGEN을 설정한다.

〔알고리듬 : MATGEN〕

구문 구조 Matrix (SSM) 생성

데이타 구조 : PTR1(용언구 pointer)

PTR2(체언구 pointer)

RLIST(결합관계 stack)

$$SSM = \begin{bmatrix} R11 & R12 & R13 & \cdots\cdots & R1m \\ R21 & R22 & R23 & \cdots\cdots & R2m \\ \cdots & \cdots & \cdots & & \\ \cdots & \cdots & \cdots & & \\ Rn1 & Rn2 & Rn3 & \cdots\cdots & Rnm \end{bmatrix}$$

m; 용언구의 갯수, n; 단위구의 갯수
Rij; i 번째 용언구와 j번째 단위구와의 결합관계

그림 3. 구문구조의 매트릭스
Fig. 3. Syntactic structure matrix.

입력 : PHSTK(결합 정보 stack)
CNT(단위구의 갯수)
출력 : SSM (구문 구조 Matrix)
PROC1 : (초기상태)
PTR1을 0으로 세트하고 PROC2로 간다.
PROC2 : (용언구 검색)
1) PTR2를 1로 세트한다.
PTR1의 값을 하나 증가시킨다.
RLIST를 NIL로 세트한다.
2) PTR1>CNT 이면 RETURN.
3) PHSTK(PTR1)이 용언구이면 PROC3로, 아니면 PROC2로 간다.
PROC3 : (구조적 결합관계 처리)
1) PTR1=PTR2 이면 RLIST에 'PRE'를 push하고, PRT2의 값을 하나 증가시킨 후 PROC4로 간다.
2) frs(PHSTK(PTR2), PHSTK(PTR1))의 값을 RLIST에 push한다.
3) PTR2의 값을 하나 증가시킨 후 PROC3로 간다.
PROC4 : (한정적 결합관계 처리)
1) PTR2>CNT 이면 RLIST의 값을 SSM에 push한 후 PROC2로 간다
2) frd(PHSTK(PTR1), PHSTK(PTR2))의 값을 RLIST에 push 한다.
3) PTR2의 값을 하나 증가시킨 후 PROC4로 간다.

그리고 이 매트릭스는 하나의 체언구가 두 개 이상의 용언구와 구조적 결합관계가 성립되고 있거나, 하나의 용언구가 두개 이상의 체언과 한정적 결합관계가 형성되는 애매성(ambiguity)을 가지고 있다. 이러한 애매성을 제거하고 완전한 구문구조 매트릭스를 얻기 위해서 각 용언구가 지배하는 체언구 또는 체언에 대한 지배 scope를 결정하는 처리가 필요하기 때문에 용언구의 지배 scope를 결정하는 규격을 정한다.

〔규칙〕 용언구의 지배 scope 결정 규칙

1) 각 체언구는 단 하나의 용언과 구조적 결합관계가 성립하며, 각 용언구는 하나 이하의 체언과 한정적 결합관계가 성립한다.

2) scope는 상위행에서 하위행의 순으로, 좌측열에서 우측열의 순으로 결정해 간다.

3) 구조적 결합관계의 우선순위는 상위행에 있으며, 한정적 결합관계의 우선순위는 좌측열에 있다.

2. 구문구조 매트릭스의 형태

문장의 종류에 따른 매트릭스의 형태는 기본문, 포유문, 복합문 그리고 포유문과 복합문의 합성에 의해 확장된 혼합문의 4가지 유형에 의한 매트릭스의 형태를 나타난다.

1) 기본문의 경우

(S7) 그는 책을 나에게 주었다.
NPs NPco NPcc VbEe
SSM=[SUB OBJ COM PRE]

위의 경우와 같이 1×n 매트릭스의 형태로 나타난다.
(단, n=단위구의 갯수)

2) 포유문의 경우

(S8) 향기가 좋은 꽃이 들에 만발하였다.
NPcs VaEd NPcs NPca VaEe

$$SSM = \begin{bmatrix} SUB & PRE-MOD & 0 & 0 \\ 0 & 0 & \downarrow SUB & ADV & PRE \end{bmatrix}$$

m×n 매트릭스의 형태로 나타나며, 한정적 결합관계 "MOD"에 의하여 수식구조로서 연결되며 이것을 ↓표로 나타낸다. 한정적 결합관계는 m−1 개가 존재한다. (단, m=용언구의 갯수)

3) 복합문의 경우

(S9) 봄이 되면 꽃이 피고 새가 운다.
NPcs VbEc NPcs VbEc NPcs VbEe

$$SSM = \begin{bmatrix} SUB & PRE : & 0 & 0 & : & 0 & 0 \\ 0 & 0 & : SUB & PRE : & & 0 & 0 \\ 0 & 0 & : & 0 & 0 & : SUB & PRE \end{bmatrix}$$

m×n 매트릭스의 형태로 나타나고, 한정적 결합관계가 존재하지 않으므로 점선 부분과 같이 m−1 개의 불연속점이 생긴다.

4) 혼합문의 경우

(S10) 영수는 붉은 장미를 좋아하고
NPs VaEd NPco VbEc

영희는 흰 장미를 좋아한다.
NPs VaEd NPce VbEe

$$SSM=\begin{bmatrix} 0 & \text{PRE-MOD} & 0 & 0 & 0 & 0 & 0 \\ \text{SUB} & 0 & \text{OBJ} & \text{PRE} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \text{PRE-MOD} & 0 \\ & & & & \vdots & & \downarrow & \\ 0 & 0 & 0 & 0 & \text{SUB} & 0 & \text{OBJ} & 0 \end{bmatrix}$$

m×n 매트릭스의 형태로 나타나며, 포유문과 복합문에서의 형태가 복합적으로 나타난다.

그리고 제Ⅱ장에서 그림 2의 문장은 (3),(5),(8),(10)의 4개의 용언구와 (1),(2),(4),(6),(7),(9)의 6개의 체언구로 분리되기 때문에 이 문장의 그림 4(a)와 같은 4×10의 구문구조 매트릭스를 얻을 수 있다.

그림 4(b)는 그림 4(a)의 불완전한 구문구조 매트릭스로부터 용언구의 지배 scope 결정규칙에 따라 얻어진 구문구조의 매트릭스이다.

$$SSM=\begin{bmatrix} 0 & \text{ADV} & \text{PRE} & \text{MOD} & 0 & \underline{\text{MOD}} & \text{MOD} & 0 & \underline{\text{MOD}} & 0 \\ 0 & 0 & 0 & \text{DET} & \text{PRE} & \text{MOD} & \underline{\text{MOD}} & 0 & \underline{\text{MOD}} & 0 \\ 0 & \underline{\text{ADV}} & 0 & 0 & 0 & \text{SUB} & \text{OBJ} & \text{PRE} & \text{MOD} & 0 \\ \text{SUB} & \underline{\text{ADV}} & 0 & 0 & 0 & \underline{\text{SUB}} & \underline{\text{OBJ}} & 0 & \text{OBJ} & \text{PRE} \end{bmatrix}$$

(a)

⇩

$$SSM=\begin{bmatrix} 0 & \text{ADV} & \text{PRE-MOD} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \text{DET} & \text{PRE-MOD} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \text{SUB} & \text{OBJ} & \text{PRE} & \text{-MOD} & 0 \\ \text{SUB} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \text{OBJ} & \text{PRE} \end{bmatrix}$$

(b)

그림 4. 그림 2의 구문구조 매트릭스
Fig. 4. Syntactic structure matrix of Fig. 2.

Ⅳ. 개념구조의 해석

일반적으로 인간의 개념세계는 두 가지 주요 영역으로 이루어져 있는데, 그 하나는 추상적인 또는 구체적인 사물을 표현하는 체언이고, 다른 하나는 사물의 모습이나 성질 등을 보이는 상태와 변화하는 과정이나 움직이는 동작을 표현하는 용언이다. 체언은 의미적인 독립성 유무에 따라 자립명사와 의존명사로 분류할 수 있고, 자립명사는 그 자신만으로 하나의 개념을 가지고 있지만, 의존명사는 다른 말의 수식을 받아야만 비로소 개념을 가지게 된다. 그러므로 자립명사만을 하나로 개념단위로 인정하여 변수 X를 도입하여 ISA-property 형식으로 표현하고 체언개념의 추출은 단위구 분리과정에서 행한다.

그림 5는 체언개념을 컴퓨터 내부에서 BNF(bacus normal form)로 형식화한 것이다.

```
⟨concept⟩::=(ISA(⟨nc⟨nc⟩ ⟨noun-concept⟩))
⟨noun-concept⟩::=(⟨noun⟩ [⟨det *⟩])
⟨det *⟩::=(⟨nc⟩ [⟨par⟩]) : (⟨det-list⟩)
⟨det-list⟩::=([⟨adv⟩] ⟨det⟩) : (⟨det⟩ [⟨det⟩])
⟨nc⟩::=X1 : X2 : X3 : ····· : Xi
  adv : 부사          noun : 자립명사
  det : 관형사        par : 조사
```

그림 5. 체언개념의 표현
Fig. 5. Expression of noun concept.

문장에서 의미적 중심이 되는 것은 용언이며, 체언은 다만 이용언의 지배에 의해 수반되는 주변적인 것이다. 따라서 문장의 의미를 이해하기 위해서는 용언과 체언 사이의 결합관계를 파악할 필요가 있다. 구문구조 매트릭스는 이들 용언과 체언의 결합관계를 형식화한 것으로서, 하나의 행(row)은 하나의 용언과 그 용언의 지배를 받는 체언들과의 결합관계를 표시한다. 그러므로 용언의 개념구조는 구문구조 매트릭스로부터 추출할 수가 있다. 용언개념은 변수 Ci를 도입하여 ISA-property 형식으로 표현하여, 한 개의 개념절을 나타낸다. 각 개념절은 한정적 결합관계의 유무에 따라 다른 절의 수식을 받는 수식구조를 갖게 된다.

그림 6은 용언개념의 컴퓨터 내부에서의 표현형식을 BNF로 형식화한 것이다.

Ⅴ. 시스템 및 처리 결과

1. 시스템 구성

사용언어는 자연어 처리에 유용한 LISP 언어를 사용하였으며, 프로그램의 16Kbyte이고, data는 30 Kbyte가 소요되었으며 그림 7은 전체 시스템의 구성도이다.

1단계로 문장을 입력받아 형태소 해석을 행한다. 형태소해석은 한일기계번역시스템의 결과를 이용하였다.[4]

2단계로 구 구조규칙에 의해 문장을 단위구로 분리한다. 이것은 문장 내에서의 체언구와 용언구의 결합관계를 추출하기 위한 전 처리단계이다.

```
⟨concept⟩ ::= (ISA (⟨vc⟩ ⟨verbial-concept⟩))
⟨verbial-concept⟩ ::= ⟨verbial-con1⟩ : ⟨verbial-
                      con2⟩ : ⟨verbial-con3⟩
⟨verbial-con1⟩ ::= (VA ⟨verbial-part⟩ [⟨subject-
                      part⟩] [⟨complement-part⟩]
                      [⟨adverbial-part⟩])
⟨verbial-con2⟩ ::= (VB ⟨verbial-part⟩ [⟨subject-
                      part⟩] [⟨object-part⟩]
                      [⟨complement-part⟩]
                      [⟨adverbial-part⟩])
⟨verbial-con3⟩ ::= (VC ⟨verbial-part⟩ [⟨subject-
                      part⟩] [⟨complement-part⟩
                      [⟨adverbial-part⟩]]) :
                    (VC ⟨verbial-part⟩ [⟨subject-
                      part⟩] [⟨determinative-
                      part⟩])
⟨verbial-part1⟩ ::= (VP (⟨infword⟩ (⟨inform⟩
                      ⟨ending⟩) [⟨adv-list⟩ :
                      ⟨det-list⟩]))
⟨adv-list⟩ ::= (⟨adv⟩ [⟨adv⟩])
⟨det-list⟩ ::= ([⟨adv⟩] ⟨det⟩) : (⟨det⟩ [⟨det⟩])
⟨subject-part⟩ ::= (SP ⟨noun *⟩)
⟨object-part⟩ ::= (OP ⟨noun *⟩)
⟨complent-part⟩ ::= (CP ⟨noun *⟩)
⟨noun *⟩ ::= (⟨inoun⟩ ⟨par⟩ [(⟨vc⟩)]) :
             (⟨nc⟩ ⟨par⟩ [(⟨vc⟩)])
⟨adverbial-part⟩ ::= (AP [⟨noun *⟩])
⟨determinative-part⟩ ::= (DP[⟨noun *⟩])
⟨nc ::= X1 : X2 : X3 : ····· : Xi
⟨vc⟩ ::= C1 : C2 : C3 : ····· : Cj
⟨inform⟩ ::= Ee : En : Ed : Ea : Ec
inoun : 의존명사 infword : 용언 ending : 어미
adv : 부사 det : 관형사 par : 조사
```

그림 6. 용언개념의 표현
Fig. 6. Expression of verbial concept.

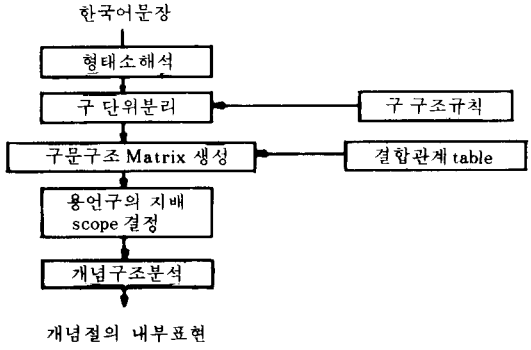


그림 7. 시스템의 블럭선도
Fig. 7. Block diagram of system.

3 단계로 체언구와 용언구의 결합관계를 추출하여 문장의 구문구조를 매트릭스의 형태로 표현하고 용언구의 지배 scope를 결정하여 완전한 구문구조 매트릭스를 얻는다.

4 단계로 개념절의 구조 및 이들 사이의 수식구조를 분석하여 컴퓨터 처리에 적합한 표현형식을 생성, 출력한다.

2. 처리 결과의 예

INPUT SENTENCE⇨저 자동차가 빨리 간다.

```
(ISA(X1 (자동차(저))))
(C1 (VB (VP (가(Ee ㄴ다) (빨리)))
        (SP (X1 가))))
T
```

INPUT SENTENCE⇨봄이 되면 꽃이 피고 새가 운다.

```
(ISA (X1 (봄)))
(ISA (X2 (꽃)))
(ISA (X3 (새)))
(Cl (VB (VP (되 (Ec 면)))
        (SP (X1 이))))
(C2 (VB (VP (피 (Ec 고)))
        (SP (X2 이))))
(C3 (VB (VP (울 (Ee ㄴ 다)))
        (SP (X3 가))))
T
```

INPUT SENTENCE⇨과학자들은 바다에서 가장 영리한 동물의 하나인 돌고래가 몇 가지 소리의 의미를 알고 있다는 것을 알아냈다.

```
(ISA (X1 (과학자들)))
(ISA (X2 (바다)))
(ISA (X3 (동물)))
(ISA (X4 (돌고래)))
(ISA (X5 (소리 (몇 가지))))
(ISA (X6 (의미 (X5 의))))
(C1 (VA (VP (영리하 (Ed ㄴ) (가장)))
        (AP (X2 에서))))
(C2 (VC (VP (하나이 (Ed ㄴ)))
        (DP (X3 의 (C1)))))
(C3 (VB (VP (알고있 (Ed 다는)))
        (SP (X4 가 (C2)))
        (OP (X6 을))))
(C4 (VB (VP (알아내 (Ee ㅆ 다)))
        (SP (X1 은))
        (OP (것 을 (C3)))))
T
```

Ⅵ. 결　　론

본 논문에서는 한국어의 용언구와 체언구의 결합관계를 이용하여 문법규칙을 형식화하고, 이를 기초로 하여 한국어 문장의 수식구조와 개념구조를 추출하는 방법에 대해서 논하였다.

본 논문에서 검토된 중요사항을 총괄하면 다음과 같다.

(1) 한국어의 수식구조는 전형적인 left-branch 전개를 이루고 있는 특성을 이용하여 문장을 단위구로 분리하는 구 구조규칙을 정의하였다.

(2) 체언구와 용언구의 결합관계를 이용하여 한국어 문장에서 구문구조 매트릭스를 유도하였다.

(3) 구문구조의 매트릭스로 부터 수식구조와 개념구조를 해석하였다.

(4) 추출된 개념단위를 술어 논리형으로 표현하였다.

이것은 일반 문장으로 부터 지식베이스(Knowledge-Base)의 구축과 추론 Machine으로 이어지며, 추후 발표할 예정이다.

參 考 文 獻

[1] H. Nomura, et. al., "Translation by understanding: a machine translation system LUTE," COLING '86, pp. 621-626, 1986.

[2] S.H. Lee, K.R. Han, J.K. Lee, "A bidirectional MT system between korean and japanese based on a pattern net," *Proc. of 10th Symposium on Information Theory and its Application*, Japan, pp. 405-410, Nov. 1987.

[3] S.K. Han, S.H. Lee, J.K. Lee, "A korean-english machine translation system based on lexical association grammar," *Proc. of TENCON 87-IEEE*, pp. 516-519, Aug. 1987.

[4] Joo Keun Lee, Kwang Rok Han, "Korean-japanese machine trasnslation system (III)," *Proc. of 1985 SISEE*, pp. 134-138 Nov. 1985.

[5] F. Hayes-Roth, "The kwoledge based expert system: a tutorial," *computer*, vol. 17, pp. 11-28, Sep. 1984.

[6] W.B. Gevarter, "Expert system: limited but powerful," *IEEE Spectrum*, pp. 39-45, Aug. 1983.

[7] J. Slocum, "Machine translation: its history, current status, and future prospects," *Proc. COLING-84*, 1984.

[8] Wilks, Y.A., Grammar, meaning and the machine analysis of language, Routledge & Kegan Paul, London, 1972.

[9] J.K. LEE, J.H. LEE, K.R. HAN, "Determination of modificatory scope and inference for korean," *Proc. of TENCON 87-IEEE*, pp. 520-523, Aug. 1987.

[10] 김민수, 국어 문법론, 일조각, 1984.

[11] 김형규, 국어학 개론, 일조각, 1985.

[12] 이길록, 국어문법 연구, 일신사, 1983.

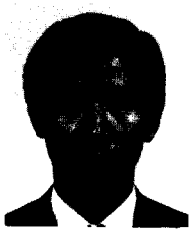
[13] 長尾眞, 言語工學, 昭晃堂(日本), 1983.

[14] Woods, W.A., "Transition network grammars for natural danguage analysis," *Communications of the ACM*, 13, pp. 591-60, 1970.

[15] I. Kudo, H. Nomura, "Lexical function transfer: a transfer framework in a machine translation system based on LFG," *Proc. COLING '86*, pp. 112-114, Aug. 1986.

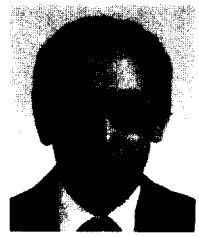
[16] M. Nagano, "The transfer phrase of the Mu machine translation system," *Proc. COLING '86*, pp. 97-103, Aug. 1986. ∗

著 者 紹 介



韓 光 錄(正會員)

1961年 5月 24日生. 1984年 2月 인하대학교 전자공학과 공학사 학위 취득. 1986年 2月 인하대학교 대학원 전자공학과 공학석사학위 취득. 1986年 3月～현재 인하대학교 대학원 전자공학과 박사학위 과정재학중. 주관심분야는 자연어 이해시스템, 기계번역 시스템 등임.



李 柱 根(正會員)

1925年 3月 5日生. 1951年 9月 서울대학교 전자공학과 공학사 취득. 1973年 3月 일본 KEIO 대학교 대학원 공학박사학위 취득. 1958年 4月～현재 인하대학교 전자공학과 교수. 주관심분야는 시각 Pattern인식, 자연어 이해시스템, 기계번역 시스템, 질문응답 시스템 등임.

崔 璋 善(正會員)

1961年 7月 30日生. 1985年 2月 인하대학교 전자공학과 졸업. 1988年 2月 인하대학교 대학원 전자공학과 석사학위 취득. 1987年 12월～현재 (주)금성사 정보기기 연구소 재직. 주관심분야는 자연어 처리 및 기계번역 분야임.