이진 결합 중심의 한국어 Chart parser

박성숙, 심영섭, 한성국*
최운천, 지민제, 이용주**
* 원광대학교 컴퓨터 공학과
** 한국전자 통신 연구소 자동통역연구실

A Chart Parser for Korean by Binary Association

Sung-Suk Park, Young-Seop Shim, Sung-Kook Han*
Un-Cheon Choi, Min-Je Zhi, Young-Ju Lee**
* Dept. of Computer Eng., Won Kwang Univ.
** Automatic Translation Group, ETRI

요 약

한국어는 구문요소의 문법기능이 표면구조상에 명시되는 구문특성을 갖고 있다. 이러한 특성은 한국어의 문법체계가 feature중심으로 전개되고 있음을 의미한다. 한국어에서의 feature 특성과 이진 결합 관계를 중심으로 하는 chart parsing 알고리즘을 제시하고 한국어 chart parser을 구현하였다.

1. 서론

자연언어 처리에서 처음 직면하게 되는 문제가 문장의 구문/의미 구조의 추출이다. 구문/의미구조의 추출은 대부분의 자연언어 처리 시스템이 통상적으로 거치게 되는 과정이지만, 아직도 효과적인 구문/의미 구조의 분석 체계를 확립하지 못한 채 많은 논란과 여러 새로운 방식이 제시되고 있는 실정이다[1][2].

구문구조 분석을 위해서는 먼저 형식화된 문법체계가 확립되어야 한다. 변형 생성 문법이래 다양한 문법체계들이 제시되어 왔다. 이러한 문법체계들은 국지적인 언어구조는 어느 정도 잘 표현하고 있지만 체계로써의 완결성과 범용성을 확보하지

못하고 계속 수정 보완되고 있다[2]. 특히, 한국어 처리의 경우에는 이들 문법체계가 갖고 있는 이러한 문제외에도 한국어의 구조에 적합한지 그 유효성을 검토해야 하는 부담을 갖고 있다. 구문구조 분석의 또 다른 문제로 parsing 알고리즘의 구축이 있다. 각 문법체계는 고유한 기술 양식을 갖고 있기때문에 이를 고려하여 주어진 문법에 대하여 실용적인 시간/공간 복잡도(time/space complexity)을 갖는 parsing 알고리즘의 발견은 용이한 일이 아니다. 구문구조 분석의 이러한 문제점을 고려하여 GPSG, HPSG 등의 자연언어의 context-freeness와 관련한 문법체계와 chart 또는 left-corner parsing방식이 연구되고 있다[4][5].

본 논문에서는 한국어의 언어구조 특성을 분석하여 한국어의 구조 형식화를 위한 문법 기술 체계를 제시하고, 이를 바탕으로 하는 chart parser를 구현한다.

2. 한국어 문법구조 특성

한국어에서는 구성적(configurational) 특성이 보이지 않으며, 일치(agreement)와 같은 구문 현상이 없다. 이로 인하여 동일한 문법 체계 하에서도 한국어 문법체계는 커다란 차이를 보이게 된다. 한국어의 잘 알려진 구문구조 특성을 재고찰하고 한국어 문법기술에 필요한 요소를 추출해 낸다.

2.1 한국어의 문법 요소

한국어는 상대적으로 자유스러운 어순 특성을 갖고 있으며, 이는 한국어의 구문구조 형성을 위해 시사하고 있는 바가 크다. 상대적 자유 어순을 갖기 위해서는 구문요소의 문법기능(grammatical function : GF)이 문장의 구조와 관계없이 항상 유지할 수 있어야 한다. 한국어의 경우, 표면구조상에서 GF를 결정하기 위한 조사와 어미들의 기능형태소의 발달은 어순 자유성과 연관하여 대표적 가시 특성이라 할 것이다. GF와 기능형태소 그리고 어순 자유성에 내재한 한국어 전개의 기본 원리를 GF의 markedness로 요약한 바 있다[8].

GF의 markedness는 문법체계가 성분구조를 중심으로 전개됨을 의미한다. 각 구문요소는 이미 자신이 표현하고자 하는 GF가 결정되어 있기 때문에 구문의 전개는 구문요소의 내적 구성 방식보다는 표출된 GF에 의하게 된다. 즉 구문요소의 내재 구성 방식은 중요하지 않게 된다. 이와 같은 상황은 영어의 경우와 비교해 보면 한국어의 구조적 특성을 명확히 대비할 수 있다.

(2-1) a. A man loves a woman.

b. A woman loves a man.

c. 철수가 영희를 사랑한다.

d. 영희를 철수가 사랑한다.

(2-1 a,b)의 영어 문장과 (2-1 c,d)의 한국어 문장은 각각 그림 2-1의 구조로 표현될 수 있다.

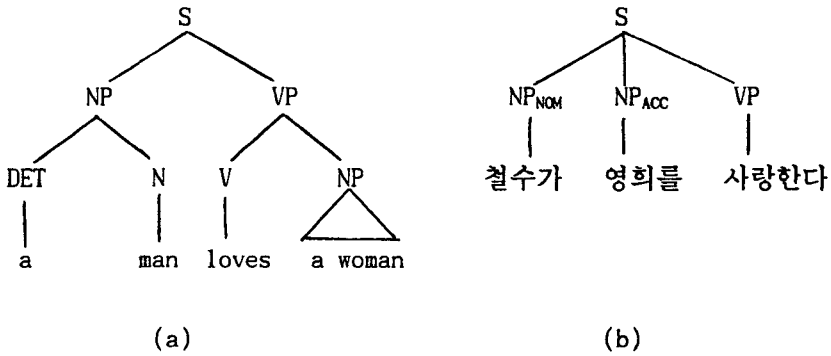


그림 2-1 구문요소의 문법기능

영어에서 구문요소는 문장의 구조적 위치에 따라 성분의 GF가 결정된다. 즉 GF와 문법구조는 상호 의존 형태로 독립적으로 취급될 수가 없다. 반면에 한국어에서는 GF가 문법 구조와는 무관하게 독립적으로 유지되는 비구성적(nonconfigurational) 특성을 갖고 있다. 전술한 바와 같이 성분의 내적 구성 보다는 marked된 GF을 구성단위로 하여 문장이 연결된다.

그런데 언어가 marked GF에 의한 비구성적 방식을 취하게 되면 언어전개는 feature 중심으로 전개되어 진다. 이는 구성으로 문법성을 확보할 수 없기 때문에, 표출된 형태에 의해서 성분의 GF가 결정되는데 기인한다. 관형화문을 통해서 feature전개 방식을 구체적으로 고찰해 보기로 한다. (2-2 a)의 구문 구조는 그림 2-2로 표현된다.

(2-2) a. the man who Bill saw

b. 철수가 만난 사람

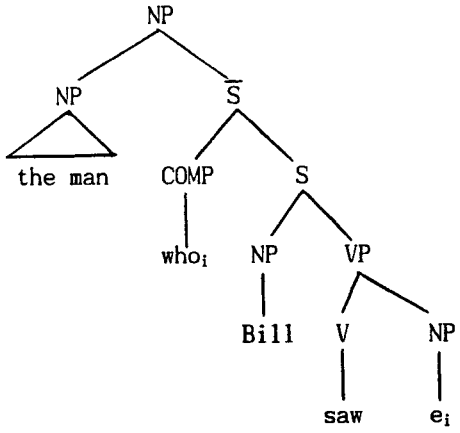


그림 2-2 영어의 관형화 구조

COMP who는 co-indexing에 의해 GF를 부여받고 있으며, 내포문은 구성적(configurational)구조에 의해서 GF를 확보하게 된다. 한국어의 관형화 구조는 그림 2-3으로 표현될 수 있다. 그림 2-3의 추정한 한국어 관형화 구조를 고찰하면, 구문구조의 형태는 관형화의 문법 특성과 거의 무관함을 할 수 있다. "-는, -을" 등의 관형화 어미가 COMP의 역할을 하는 것으로 간주되기도 하나, 영어의 COMP와 비교할 때 커다란 언어학적 차이가 엿보인다. 따라서 그림 2-3보다는 GPSG에서의 foot feature principle 등과 같은 feature중심의 구조를 설정하는 것이 보다 합리적일 것이다. 이것은 한국어 성분의 GF는 문법범주와 feature에 의존적임을 의미한다. 그림(2-3)에 보이는 유추는 구성성분의 GF 표현에 있어 문법범주와 feature의 역할을 가능케 해 준다.

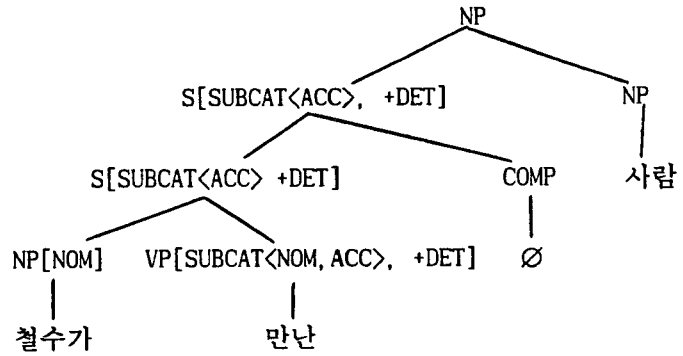


그림 2-3 한국어의 관형화

(2-3) a. 그 사람 D[DET]

b. 철수의 책 NP[DET]

c. 철수가 말했다. NP[NOM]

d. 철수가 만난 사람 S[DET]

GF표현에 있어 문법범주와 feature의 역할 분담은 한국어에 있어 (2-4 a)와 같은 규칙의 설정은 비합리적임을 의미한다.

(2-4) a. DET → D | NP ·{의} | VP ·{은, 는}

b. NP[] → X[DET] · NP[]

반면에 (2-4 b)와 같은 GPSG형식의 규칙 기술체계가 한국어에 보다 적합함을 알 수 있다. (2-4 b)에서는 feature에 문법 범주에 의해 규칙체계가 개념적으로 기술되어 진다. 특히, 한국어의 부분 어순 자유 특성을 기술하기 위해서는 체언과 용언의 결합을 규정하는 용언의 하위 범주 feature 특성에는 어순에 대한 제약성을 갖지 않는다. feature는 성분에 대한 구문특성을 표시하는 정보일 뿐 구조체는 아닌 것이다.

2-2 한국어 문법 구조의 형식화

한국어는 비구성적 구조를 갖기 때문에 문법규칙의 기술에 어려움이 있다. 어순의 강한 제약과 구성적 구조하에서 문법규칙은 명확한 형식을 갖을 수 있지만, 문법구조가 자유로운 구성을 할 경우에는 규칙을 유도하고 기술하는 것은 아주 어렵게 된다.

한국어의 문법 기술은 전술한 바와 같이 feature와 위주로 기술하는 것이 바람직하다. 본 연구에서는 feature을 중심으로 한 GPSG을 기준으로 하여, 규칙기술의 명료성과 parser의 구현의 효율성을 고려한 문맥 자유(context-free)형식의 기술체계를 이용한다. 문법 기술 체계의 일반 형식은 (2-5)과 같다.

(2-5) $A\{\sigma, \varepsilon, \tau\} \rightarrow B[\alpha] \cdot C[\beta]$

여기서 A, B, C는 구문요소, α와β는 feature이며 $\sigma, \varepsilon, \tau$는 구문처리 action이다.

(2-5)의 기술형식의 특성은 다음으로 요약된다. 규칙기술은 이진 결합(binary association)형식을 갖는다. parsing tree에서 각 노드는 문법범주와 feature을 갖고 있을 뿐만 아니라 한국어에서는 문법구조의 계층성은 큰 의미를 갖지 못한다. 두 문법요소간의 결합관계만으로 문장의 구문구조 파악이 가능해진다. 규칙 기술에는 구문처리를 위한 고유 action이 기술된다. feature중심의 문법체계는 GPSG의 HFC 등에서 보이는 바와 같이 feature에 대한 기본 원리나 feature의 unification으로 전개해 나가는 것이 일반적이다. 이러한 경우 단순한 처리를 위해서도 원리들을 모두 적용해야 하며, 예외적인 feature연산을 위한 meta-rule등 보조 장치의 도입으로 규칙체계가 복잡하게 되어진다. 실제의 경우 대상이 되는 feature간의 처리는 단순하고 제한되어 있기 때문에 해당 규칙에 처리 action을 기술해 두는 것이 편리하다. σ와 ϵ는 B와 C의 feature처리를, τ는 규칙 자체의 feature처리를 의미한다. (2-5)의 규칙기술은 이와 같은 처리를 생략하고 feature의 원리나 unification을 적용하도록 할 수도 있다.

(2-5)으로 기술한 한국어의 문법체계는 (2-6)과 같다.

(2-6)
```
Y{,@,} --> X[DET]  · Y[N]
Y{,@,} --> X[ADV]  · Y[N]
Y{,@,} --> X[VEC]  · Y[V]
Y{,@,} --> X[PK]   · X[]
Y{,@,} --> X[PC]   · Y[V PC]
Y{,@,} --> X[PS]   · Y[V PC]
```

(2-6)은 문법요소의 추상화된 결합관계를 잘 나타내고 있다. feature는 문법범주보다 광법위하고 강력한 구문능력을 갖고 있어, 문장성분을 전개하는데 편리하다. 한국어를 위한 구문규칙 (2-6)에서 주목해야 할 것은 feature에 의해 구문이 전개된다는 사실이다. 영어등에서 feature는 구문전개가 아닌 구문검사의 도구로 사용되고 있음에 유의해야 한다. 대표적인 예로써 (2-7) 의 feature 부가 규칙이 있다.

(2-7)
```
S --> NP VP
<NP arg> = <VP arg>
```

(2-7)은 문장의 전개는 구문범주 NP 와 VP 에 의하며, 다만 이들의 결합 가능 조

건을 명시하기 위해 feature 중심의 수식이 사용되고 있음을 알 수 있다. 그러나 한국어에서 feature 는 구문구조의 전개를 위한 문법요소의 기능을 담당하고 있다.

3. 이진 결합 중심의 한국어 Chart parser

한국어의 문법체계는, 형태소 또는 어휘수준에서는 구문 범주로, 성분수준에서는 feature구조로 이원화 된다. 어휘 본래의 문법요소의 가시적 특성은 문장 성분이 되면서 추상화된 feature 중심의 문법성만 남게 된다. feature중심 문법체계에서 parsing은 DAG를 이용한 unification으로 구현하는 것이 일반적이다. 한국어에 있어 feature는 문법 적용의 제약조건(constraint)보다는 근본적인 문법 규칙의 전개 기능을 갖고 있어 CFG중심의 parsing방법을 적용하는 것이 효과적이다. 본 논문에서는 (2-6)의 문법규칙에 대한 chart parsing 방식을 보이고자 한다.

형태소 분석을 통해서 얻어진 구성 성분의 어휘수준 정보는 그림3-1의 cell에 저장된다. chart parsing은 구성성분간의 node와 arc로 구성되는 tuple로 구현하지만 본 논문에서는 feature cell은 형태소 분석에 의해 우선적으로 채워진다[8].

(3-1) 철수가 영희에게 그 책을 주었다.

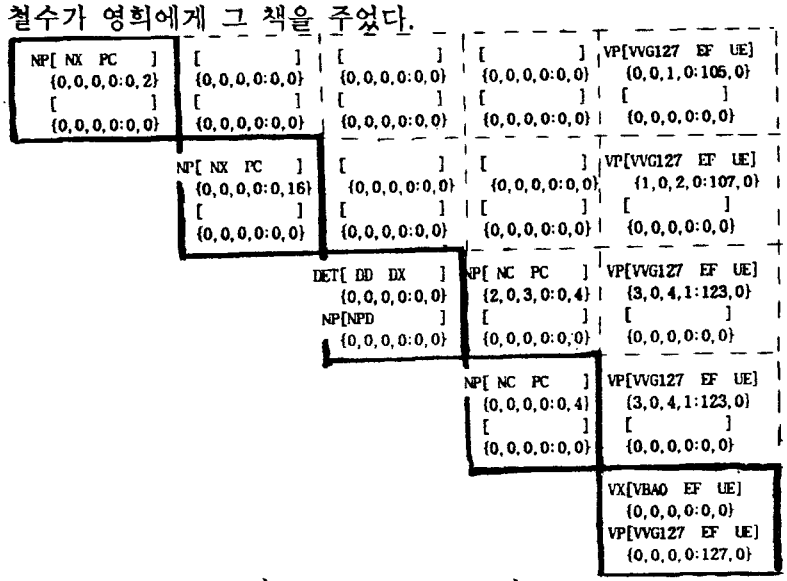
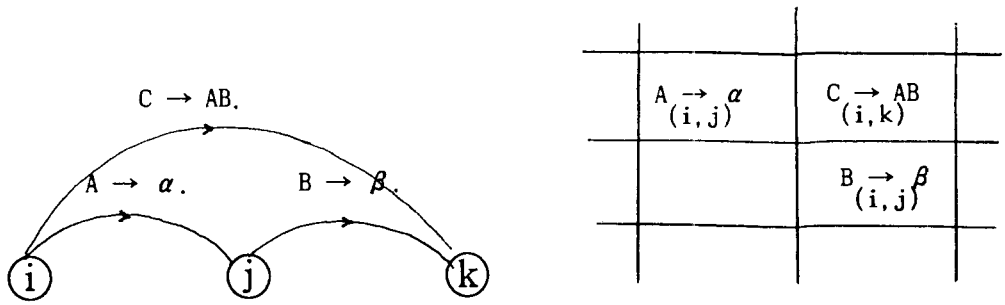


그림 3-1 chart parser의 cell

그림 3-1의 cell 구조와 chart parsing의 기본 구조와의 관계를 비교해 보면 그림 3-2와 같다.



(a) (b)

그림 3-2 Chart parser의 기본처리

tuple구조는 다루기 어렵고 적용해야 할 규칙을 찾아내기가 힘들다. cell구조에서는 규칙의 적용이나 feature정보관리가 보다 편리해진다. 이진 결합관계로 구성되는 구문 규칙하에서의 chart parsing 알고리즘은 다음과 같다.

```
[SK Chart Paser]
 • x, y, ninfo /* 현재 cell의 (x,y)좌표와 cell 내의 정보수 */
 • px, py, pninfo /* 전상태 cell의 (x,y)좌표와 cell내의 정보수 */
 • c, pc /* 현재의 cell 위치와, 전상태의 cell 위치 */

/* 전체 성분에 대하여 반복 */
for (x=1; x<nToken ; x++) {
      /* cell 지정의 y축 이동 */
   for (y=x; y>0; y--) {
      • 현재 cell의 위치를 가져온다.
      /* 현재 cell 내의 정보수 만큼 반복한다. */
      for (ninfo =0; ninfo < (cell 내의 정보수) ; ninfo++) {
      /* 전상태의 cell 좌표의 y 축 이동 */
         for ( py= px=y-1; py >= 0.3 ; py--) {
         • 규칙의 LHS가 저장될 위치를 구한다.
          /* 전상태 cell 내의 정보수 만큼 반복 */
           for (pninfo=0; pninfo < (cell내의 정보수) ; pninfo++) {
           • 규칙의 RHS 패턴 매칭
           • 일치한 규칙의 LHS을 (x, py)에 저장
            }
         }
       }
     }
   }
```

제시된 chart parsing 알고리즘은 최악의 경우 대략 $O(cn^2)$의 기억공간의 요구와 $O(cn^3)$의 시간 복잡성을 갖는다. 평균적인 경우 훨씬 양호한 결과를 보이고 있다. (3-1) 단문처리의 결과를 그림 3-1에 보였다. 모호성(ambiguity)을 갖는 문장에 대해 알고리즘은 가능한 모든 parsing tree을 생성한다.

```
# Morphoeme Output Table # ------------------------------ #
00 :: 00 : 00 :: 착하(VVG good)_(EXD VERB DET)
01 :: 01 : 00 :: 철수(NX CHULSU)와(PK PPS)
02 :: 01 : 01 :: 철수(NX CHULSU)와(PC PCS)
03 :: 02 : 00 :: 영희(NX younghee)가(PC SUB)
04 :: 03 : 00 :: 학교(NC school university)에(PC PCX)
05 :: 04 : 00 :: 가(VVG127 go)_(EP A1 PRESENT)다(EF EFX).(UE PER)
06 :: 04 : 01 :: 가(VBA0 PRG)_(EP A1 PRESENT)다(EF EFX).(UE PER)
# Parse Table # ---------------------------------------- #
 VP[VVG  DX    ] NP[ NX  PK    ] NP[ NX  PC    ]  [           ] VP[VVG127  EF  UE]
   {0,0,0,0:255,0}   {0,0,1,0:0,0}   {1,0,2,0:0,2}   {0,0,0,0:0,0}   {2,0,3,0:93,0}
   [           ] NP[ NX  PC    ] NP[ NX  PC    ]  [           ] VP[VVG127  EF  UE]
   {0,0,0,0:0,0}   {0,0,1,1:0,128}   {0,0,1,0:0,2}   {0,0,0,0:0,0}   {2,1,3,0:93,0}

                 NP[ NX  PK    ] NP[ NX  PC    ]  [           ] VP[VVG127  EF  UE]
                   {0,0,0,0:0,0}   {1,0,2,0:0,2}   {0,0,0,0:0,0}   {2,0,3,0:93,0}
                 NP[ NX  PC    ]  [           ]  [           ]  [           ]
                   {0,0,0,0:0,128}   {0,0,0,0:0,0}   {0,0,0,0:0,0}   {0,0,0,0:93,0}

                             NP[ NX  PC    ]  [           ] VP[VVG127  EF  UE]
                               {0,0,0,0:0,2}   {0,0,0,0:0,0}   {2,0,3,0:93,0}
                               [           ]  [           ]  [           ]
                               {0,0,0,0:0,0}   {0,0,0,0:0,0}   {0,0,0,0:0,0}

                                         NP[ NC  PC    ] VP[VVG127  EF  UE]
                                           {0,0,0,0:0,32}   {3,0,4,0:95,0}
                                           [           ]  [           ]
                                           {0,0,0,0:0,0}   {0,0,0,0:0,0}

                                                        VP[VVG127  EF  UE]
                                                          {0,0,0,0:127,0}
                                                        VX[VBA0  EF  UE]
                                                          {0,0,0,0:0,0}
```

```
# Parse Tree Graph # ------------------------------ .

#0 : (S(VP[VVG127 EF UE](NP[NX PC](NP[NX PK](VP[VVG DX](착하(VVG good)_(EXD VERB DET))NP[NX PK](철수(NX CHULSU)와(PK PPS)))NP[NX PC](영희(NX young
SUB)))VP[VVG127 EF UE](NP[NC PC](학교(NC school university)에(PC PCX))VP[VVG127 EF UE](가(VVG127 go)_(EP A1 PRESENT)다(EF EFX).(UE PER))))))

S
|
VP[VVG127 EF UE]
|-------------------------------------------------------------------------|
NP[NX PC]                                                                 VP[VVG127 EF UE]
|-----------------------------------------------|                         |-----------------------------------|
NP[NX PK]                                       NP[NX PC]                 NP[NC PC]                         VP[VVG127 EF UE]
|-----------------------------|                 |-----------|             |-------------------|             |---------|-------------|------|
VP[VVG DX]                    NP[NX PK]         영희        가            학교                에            가        _             다     .
|---------------|             |---------|       |           |             |                   |             |         |             |      |
착하            _             철수      와      NX younghee PC SUB       NC school university PC PCX       VVG127 go EP A1 PRESENT EF EFX UE PER
|               |             |         |       |           |             |                   |             |         |             |      |
VVG good        EXD VERB DET  NX CHULSU PK PPS

#1 : (S(VP[VVG127 EF UE](NP[NX PC](VP[VVG DX](착하(VVG good)_(EXD VERB DET))NP[NX PC](NP[NX PK](철수(NX CHULSU)와(PK PPS))NP[NX PC](영희(NX young
SUB)))VP[VVG127 EF UE](NP[NC PC](학교(NC school university)에(PC PCX))VP[VVG127 EF UE](가(VVG127 go)_(EP A1 PRESENT)다(EF EFX).(UE PER))))))

S
|
VP[VVG127 EF UE]
|-------------------------------------------------------------------------|
NP[NX PC]                                                                 VP[VVG127 EF UE]
|-----------------------------|                                           |-----------------------------------|
VP[VVG DX]                    NP[NX PC]                                   NP[NC PC]                         VP[VVG127 EF UE]
|---------------|             |-----------------------------|             |-------------------|             |---------|-------------|------|
착하            _             NP[NX PK]                   NP[NX PC]     학교                에            가        _             다     .
|               |             |---------|                 |-----------|                   |             |         |             |      |
VVG good        EXD VERB DET  철수      와                영희        가                    NC school university PC PCX       VVG127 go EP A1 PRESENT EF EFX UE PER
                              |         |                 |           |
                              NX CHULSU PK PPS            NX younghee PC SUB
```

cell을 중심으로한 chart parsing은 구문 분석 과정이 그대로 cell에 투영되며 feature을 비롯한 정보의 관리가 용이하고 unification 등의 처리를 쉽게 보강할 수 있는 장점이 있다.

4. 결론

구문구조 분석을 위해서는 언어의 문법체계가 우선 정립되어야 한다. 언어의 문법 체계는 범용성을 지향하는 기존의 문법체계를 원용하거나 새로운 체계를 설정할 수도 있을 것이다. 어느 경우에도 해당 언어 구조의 기술과 처리에 적합한 체계를 갖어야 한다. 본 논문에서는 한국어의 구조 특성을 고찰하여 문법체계 기술은 feature중심이어야 함을 논증하고 이진 결합 관계를 기초로한 체계를 제시하였다. 한국어의 문법체계에 있어 feature는 다른 언어에서와는 다른 역할을 한다. feature는 문법 제약 조건을 기술하고 unification 되는 대상이 아닌 문법 형성의 근간 규칙 요소로의 기능을 갖는다.

chart parser을 구현하는데 편리한 방식을 보였다. 본 논문의 방식은 특정 문법 체계와 관계없이 적용 가능하며 적용규칙의 탐색 과정의 합리화 등 chart parsing의 성능 개선에도 도움이 될 수 있을 것이다.

참고문헌

1. Gazdar, Gerald et al, Generalized Phrase Structure Grammar, Cambridge, Mass. : Harvard University Press, 1985
2. Pullum, Geoffrey and Gerald Gazdar, " Natural Languages and Context-free Languages", Linguistics and Philosophy, 4:471-504, 1982
3. Ades, Anthony and Mark Steedman, "On the order of words", Linguistics and Philosophy, 4:517-558, 1982
4. Karttunen, L., D-PATR : a development environment for unification-based grammar, CSLI-86-61, 1986
5. Thompson, Henry, "Chart parsing and rule schemata in PSG", ACL 19th Annual Meeting, 167-72, 1981
6. 손덕진, 최기선, 김길창, "단일화 중심 문법론에서의 단일화 방법", 인지과학 2:1 1990
7. 서영훈, 김영택, "활성 차트를 이용한 중심어 후행언어의 파싱", 한국정보과학회 논문지, Vol.17, No.1, 1990
8. 이미선 외, "한국어 형태소 분석기의 정형화", 93 한국정보과학회 논문집, 20권 1호, 1993