# 언어정보

제9호 · 2008

고려대학교 언어정보연구소

# 차 례

# 세종 구문분석 말뭉치를 기반으로 한 확률 문맥자유문법 규칙

최재웅-송상헌-전지은*

## 1. 서론

1998년부터 2007년까지 10년에 걸친 '세종계획'이 종료되었고 그 결과물이 정리되어 공개되었다 (국립국어원 2007, http://sejong.or.kr 참고). 주로 다양한 말뭉치 및 현대국어 전자사전 등으로 이루어진 결과물은 그 자체로서도 의미가 적지 않지만, 궁극적으로는 그것이 얼마만큼 활용되느냐에 따라 그것의 가치 및 의의가 올바로 평가될 수 있을 것이다.

그동안 한국어 구문분석 공개 말뭉치로는 UPenn에서 개발된 Penn Korean Treebank (V1.1 5만여 어절, V2.0 13만 어절 규모, 이하 'PKT')가 있는 바,[1] 이를 바탕으로 하여 분석적 연구가 일부 이루어 졌다 (Han et al. 2002, Lee et al. 2004, Han 2006). 세종 구문분석 말뭉치(Sejong Korean Treebank, 이하 'SKT')의[2] 경우 중

* 고려대학교 언어학과, {jchoe,yooseon21,jejeon}@korea.ac.kr. 본 논문의 작성 과정에 도움을 주신 박진호 선생님, 노용균 선생님, 이종민 선생님, 구민모 선생님과 세 분의 심사위원께 감사를 표한다.

1) Rim (2001)에 따르면 STEP2000 Tree Tagged Corpus (31,086문장 규모)와 의존문법에 기반한 Sogang Tree Tagged Corpus (12,784 문장 규모) 등도 있다.

2) SKT에 대한 보다 기본적인 소개 및 관련 정보는 김홍규 외 (2003), 강범모・김의수 (2004) 참조.

간 단계 결과물을 기반으로 한 한국어 문형에 대한 신서인 (2006)의 연구가 아직까지는 유일한 것으로 보인다. 세종 말뭉치가 규모나 그 개발에 관여한 인적 물적 자원의 규모 및 국가적인 중점 사업의 일환으로 구축되었다는 점을 감안해 볼 때 적어도 당분간은 한국어의 대표적인 말뭉치로 자리매김할 수 있을 것이다. 그러한 점에서, SKT를 포함한 세종 말뭉치를 활용한 다양하고 체계적인 연구는 꼭 필요하다 할 것이다. 세종계획 형태소 분석 말뭉치의 경우에는 가장 기초적인 통계가 김흥규・강범모 (2000, 2004)에 정리되어 제시된 바 있으나, 구문분석 말뭉치는 그러한 가장 기본적인 통계사항에 대한 정리도 아직 충분치 못하다.

구문분석 말뭉치의 주요 활용방안의 하나는 문맥자유문법 규칙(이하 'CFG 규칙')을 도출하는 것이다 (Abeillé 2003). 해당 말뭉치가 어느 정도 대표성이 있다면 그 언어의 CFG 규칙이 대부분 망라되어 있을 것이라는 기대를 할 수 있다. 거기에 더해 말뭉치에 활용된 규칙별로 빈도수를 파악할 수 있기 때문에 규칙별 확률을 도출할 수도 있다. 이러한 정보는 그 자체로도 언어의 특성을 기술하는데 주요한 정보가 될 것이고, 또한 CFG 규칙을 바탕으로 한 관련 범주간의 상호 제약관계나 결합 가능성 등을 찾아내는데 요긴하게 쓰일 수 있다. 전산언어학적인 관점에서는, Charniak (1996)에 의해 주장되고 입증된 바처럼, 말뭉치를 기반으로 추출한 CFG 확률 규칙(Probabilistic CFG rules)은 구문분석기 개발에 유용하게 쓰일 수 있다.

따라서 본 연구에서는 우선 세종계획을 통해 구축된 약 80만 어절 규모[3] 구문분석 말뭉치로부터 CFG 규칙을 추출하여 그것을 바탕으

3) "80만 어절"이란 규모는 국립국어원 (2007)에 언급된 것을 바탕으로 한 수치로 실제 본 연구에서 파악한 수치와는 차이가 있으나(2절 논의 참고), 편의상 "80만 어절"이라고 칭하기로 한다.

로 한국어의 주요 구문적 특성에 대한 일반화를 시도하는 것을 주요 목표로 한다. 그런데 말뭉치로부터 추출한 자료 해당 언어의 주요 관련 규칙성을 반영하기 위해서는 우선 말뭉치 자체가 해당 언어의 특성을 반영할만한 규모를 갖추었는지에 대한 검토가 필요하다. 또한 말뭉치로부터 자료를 추출하는 과정 역시 명료하며 타당성을 갖추어야 할 것이다. 이를 위해 본 연구에서는 SKT에 대한 최소한의 검토와 아울러 본 연구에서 활용한 CFG 규칙 추출의 핵심 알고리즘에 대한 소개 및 검증도 병행하도록 한다.[4)]

각 절의 구성은 다음과 같다. 2절에서는 우선 CFG 규칙 추출의 전제가 되는 말뭉치 및 규칙 추출 방법에 대하여 소개하고 검토한 뒤에, 그 둘에 대한 검증을 통해 타당성을 정립한다. 3절에서는 SKT의 구성상의 특징을 기존 언어학적 분석 방법과 비교한 뒤에, SKT로부터 추출한 CFG 규칙의 종류 및 상대적 빈도 등을 정리하고, 그러한 빈도 및 분포를 근거로 규칙내 범주간 상관관계를 도출한다. 4절은 본 논문의 결론이다.

## 2. 세종 구문분석 말뭉치 및 문맥자유문법 규칙 추출

이 절에서는 SKT의 주요 통계에 대한 검토를 하고(2.1, 2.3절), 또한 SKT로부터 그러한 통계값 및 CFG 규칙 추출에 대한 방법을 소개한다(2.2절). 이 절에서는 주로 통계적 차원에서의 SKT의 특성에

4) 본 연구는 SKT를 여러 관점에서 검토하고 검증하여 SKT의 특성을 정리한다는 의미도 지니게 된다. 그런데 하나의 말뭉치가 모든 언어학적 문제점을 해결할 수는 없다는 점에서, SKT의 특성을 여러 시각에서 구체적으로 검토하고 정리하는 것은 그것의 특성과 아울러 제한점도 살펴보게 되는 셈이다. 따라서 본 연구는 SKT의 주요 특성과 함께 제한점도 살핀다는 의미가 있다고 본다.

대한 검토 및 논의를 진행하고, SKT의 구성 원칙이나 방식 등 내용상의 특성에 대한 검토는 다음 절로 넘긴다.

## 2.1 세종 구문분석 말뭉치 기초 통계

SKT의 기본적인 통계치는 아래와 같다.[5][6]

[표 13] SKT 기초 통계 (1)

| 종류 | 규모 |
|---|---|
| 파일 | 31 |
| 문장 | 77,121 |
| 태그 | 2,487,979 |
| 종단 절점 (terminal nodes) | 855,350 |
| 비종단 절점 (non-terminal nodes) | 1,633,492 |
| 비종단 절점 유형 | 402 |
| 형태소 | 1,874,623 |

위 표를 근거로 문장당 평균 어절수 및 문장당 평균 형태소 수를 계산하면 다음과 같다.

---

5) 제시된 통계치는 Unicode로 작성된 SKT를 UTF-8으로 일괄 변환한 뒤에 Perl로 작성된 프로그램(SKT_tagCount.pl)을 이용하여 추출한 수치다.

6) SKT의 종단 절점, 즉 어절의 정확한 숫자를 단순한 통계로만 파악하는 데는 어려움이 있다. SKT는 [세종 형태분석 말뭉치]의 일부를 모태로 하고 있다고 알려져 있는 바, SKT의 원본이라 할 수 있는 원시 말뭉치나 형태분석 말뭉치의 분석 기준과 SKT에서의 어절 분석 기준이 다를 수 있다. 예를 들어 원본 자체에 띄어쓰기가 되어 있지 않은 표현 “책내(BGAA0001.txt, 8368-76줄)”를 생각해 보자. 원본을 가급적 그대로 살리는 형태분석 말뭉치에서는 그것을 하나의 어절로 간주하고 내부 형태분석을 제시해 주면 충분할 것이다. 반면 구문분석 말뭉치에서는 “책”과 “내” 사이에 구절경계 절점이 들어가므로 그 둘을 구분해 주어야 하고 따라서 별도의 어절로 계산된다. 특히 부호가 들어간 표현의 경우에 큰 차이를 유발한다. “<명징성(clarity,”(BGHO0127.txt, 15627-71줄)의 경우 원시 말뭉치 방식으로 보면 어절 하나겠지만, 부호마다 독자적인 범주가 부여되는 SKT에서는 다섯 개의 어절로 계산될 수 있다. 본 연구에서는 후자의 방식을 택하여 계산하였다. 이러한 점을 알게 해 준 김주영에게 감사를 표한다.

[표 14] SKT 기초 통계 (2)

| | |
|---|---|
| 문장당 평균 어절수 | 11.1 |
| 문장당 평균 형태소수 | 24.3 |

## 2.2 문맥자유문법 규칙 추출 방법

앞에서 언급한 바와 같이 SKT로부터 CFG 규칙을 추출하기위해서는 적절한 도구의 활용이 필수적이다. 세종계획의 일환으로 개발된 도구로는 2003년도 배포된 [구문분석 말뭉치 종합관리도구]와 2007년도에 배포된 [한마루]를 들 수 있다.[7)]

그러나 본 연구의 목적에 맞게 활용하기에는 그 둘 다 적합하지 않다. [구문분석 말뭉치 종합관리도구]는 의존관계를 추출해 주기 때문에 본 연구에서 원하는 CFG 규칙 도출에 직접 활용할 수 없고, [한마루]에서는 구문 검색만 가능할 뿐 사용된 CFG 규칙을 종합적으로 추출해 주는 기능이 마련되어 있지 않다. 따라서 본 연구에서는 별도로 개발한 프로그램 Xavier를[8)] 활용하였다. 다음 소절에서는 Xavier의 주요 특성을 간략히 살펴보고, 이어지는 소절에서는 세종 구문분석 말뭉치로부터 Xavier를 통해 추출된 결과가 검색 대상을 모두 포함하며, 또한 검색 대상만을 포함하는지를 검증하도록 한다.

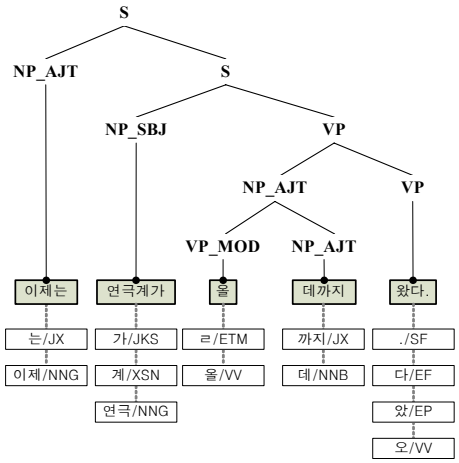
7) [구문분석 말뭉치 종합관리도구]는 말뭉치 변환도구, 말뭉치 검색도구, 말뭉치 수정도구를 묶어 하나의 통합환경을 제공하는 프로그램이다. [글잡이] 후속으로 개발된 [한마루]는 세종 현대국어 기초말뭉치에 대한 종합적인 검색 도구로, 그 중에서 구문분석 말뭉치를 다루는 부분만 본다면 구문에 대한 다양한 검색을 주로 수행하되 몇 가지 기초적인 통계도 추출할 수 있도록 구현되었다.

8) 가칭 'Xavier'는 송상헌에 의해 2007년도에 개발되었으며, 이에 대한 자세한 내용은 추후 별도로 발표할 예정이다.

### 2.2.1 Xavier의 핵심 알고리즘

두 항목 간의 관계를 표상하는 방법은 중위 표기법, 전위 표기법, 후위 표기법으로 구분할 수 있다.[9] 예컨대, 수학식에서 흔히 사용되는 '1+2', 'a ◦ b' 등은 중위 표기법이다. 즉, 두 피연산자(operand)에 대한 연산자(operator)의 위치에 따라 결정되는 것이다. 통상, 구문분석 말뭉치는 이들 가운데 전위 표기법으로 표상된다. 구조를 밝히는 데 가장 효과적이기 때문이다. (1)은 실제 SKT에서 추출한 예이며, 동일 정보를 언어학적 연구에서 일반적으로 쓰이는 수형도를 병기하였다.

(1) ; 이제는 연극계가 올 데까지 왔다.
(S (NP_AJT 이제/NNG + 는/JX)
(S (NP_SBJ 연극/NNG + 계/XSN + 가/JKS)
(VP (NP_AJT (VP_MOD 오/VV + ㄹ/ETM)
(NP_AJT 데/NNB + 까지/JX))
(VP 오/VV + 았/EP + 다/EF + ./SF))))




9) 이들 각각은 어느 하나가 보다 우월한 방법이라 할 수는 없으며, 나름의 장단점을 가지고 있을 뿐이다.

전산학 일반에서는 위와 같은 그림을 흔히 파스 트리(Parse Tree)라 한다. 파스 트리 자료 구조에서는 현재 절점(Current Node)과 그 LDN, RDN이[10] 중요한 연결 고리를 형성한다. 즉, 위 그림에서 보는 바와 같이, 'S', 'NP_AJT', 'VP' 등과 같은 각각의 절점은 자신의 LDN과 RDN의 MN이 되며, 다시 이 절점은 상위 절점의 딸 절점이 되는 형태를 반복하여 전체 문장 구조가 구성된다. 달리 말하자면, 최상위 절점에 모든 절점이 계층적으로 물려 있는 것이다.[11] 구문분석 말뭉치의 각 절점이 기술된 순서대로 차례차례 입력된다고 가정하였을 때, 파스 트리를 구성하는 알고리즘의 얼개는 다음과 같다.[12]

---

10) 서술 편의상 앞으로 빈번하게 사용될 일부 용어에 대한 몇 가지 약어를 추가로 도입하기로 한다. 이분지 가설에 근간한 SKT는 세 개의 절점으로 구성된 '삼각구조'의 형태로서 표상되는 바, 그 중에서 엄마 절점(Mother Node)는 MN으로, 좌측 딸 절점(Left Daughter Node)은 LDN으로, 우측 딸 절점(Right Daughter Node)은 RDN으로 약어를 사용한다. 참고로, [한마루]에서 구문 검색시에 이용되는 대표적인 검색 방법이 삼각구조를 기준으로 한 '삼각검색'이다.

11) 파스 트리는 통상 스택(stack)에 기반하여 구축된다. 전산적 알고리즘에서 가장 빈번히 사용되는 형태인 스택은 자료에 대한 접근이 목록의 끝에서만 이루어지도록 고안된 선형 구조이다. 이 스택은 목록의 제일 끝에 자료를 추가하는 푸시(push) 동작과 다시 그 맨 끝의 자료를 꺼내는 팝(pop) 동작으로 작동하는데, 이를 다른 말로 LIFO(Last In, First Out)라 한다.

12) 위와 같은 알고리즘을 실제 구현하는 데는 참조 연산에 우수한 성능을 보이는 C 계열 프로그래밍 언어가 보다 적합하다. 또한 자료 처리의 성격으로 미루어 보건대, 처리의 양이 비교적 많을 뿐만 아니라, 복잡한 수학식 연산을 수행하여야 할 필요가 제기된다. 따라서 Xavier 모듈은 빠른 속도로 실행이 가능하면서도 표준적인 개발 방법론에 기반한 ANSI C++로 구현되었다. Xavier 모듈을 이용하여 85만 어절 규모 전체 SKT를 파스 트리 형태로 구성하는 데 소요되는 시간은 약 13초로 측정되었다. 이때 사용된 시스템은 CPU 1828 Mhz, RAM 512MB로서 2008년 1월 현재 사무용 PC에 해당한다. 일반적 환경에서 초당 6000개 내외의 문장을 처리할 수 있다는 점에서, 속도 측면에서는 안정성을 지닌다고 하겠다.

```
1   make_parse_tree(node):
2               node→left = NIL
3               node→right = NIL
4           if node is not a terminal node:
5                       node→right = pop()
6                       node→left = pop()
7                       if node→left is NIL:
8                               node→left = node→right
9                               node→right = NIL
10          push(node)
```

위 알고리즘의 구동 방식을 간단히 짚어보도록 하자. 우선 하나의 절점(node)이 파스 트리에 새로 입력이 되면 (1행), 처음 단계로 그 절점의 LDN과 RDN을 비운 상태로 만든다 (2~3행). 다음으로 그 절점이 형태소 결합 단계가 아닌 비종단 절점이라면 (4행), 그 절점의 RDN과 LDN에 스택에서 꺼내온 절점을 차례로 대입하도록 한다 (5~6행). 경우에 따라 딸 절점이 하나 뿐인 절점이 있을 수 있으므로 (7행), 이 경우에는 LDN과 RDN을 치환한 후 RDN을 비워두도록 한다 (8~9행). 끝으로 각 절점은 다른 절점의 LDN 또는 RDN으로 처리되어야 하므로 현재까지 구축된 절점은 스택에 추가 된다 (10행). 이와 같은 처리를 반복하면 최종적으로 최상위 절점의 정보에 다른 절점의 정보가 차례차례 포함되는 파스 트리를 완성할 수 있다.

### 2.2.2 Xavier 검증

Xavier는 위와 같은 파스 트리 알고리즘을 응용하여 SKT 또는 PKT와 같은 구문분석 말뭉치를 처리하기 위한 모듈이다. 그런데 이렇게 구축된 모듈이 정상적으로 작동하여 SKT의 정보를 빠짐없이, 그리고 정확하게 처리하는지에 대한 검증 작업이 필요하다.[13] 검증

13) 이러한 검증 방법을 흔히 Precision과 Recall이라고 칭한다. Precision은 찾는 자료가 "정확하게" 찾아졌는지를 평가하는 지표이다. 한편, Recall은 찾아낸 자료가

의 절차는 두 가지 차원에서 진행하였다.

첫 번째 검증 방법은, "빠짐없이"에 대한 검증 과정의 하나로, Xavier를 통해 도출된 전체 범주의 종류와 빈도의 합이 별도 방법으로 추출한 결과와 일치하는지 여부를 확인하는 것이다. 우선 말뭉치의 일부를 대상으로 실험해 본 결과 Xavier로 추출된 규칙과 수작업으로 추출한 규칙이 정확하게 일치하였다. 보다 큰 규모로도 문제없이 추출하는지 검증하는 한 가지 간접적인 방편으로 Xavier로 추출한 규칙을 근간으로 계산한 어절수와 형태소 규모를 위 2.1에 제시된 별도 방식으로 추출한 통계와 비교하였다. 어절의 경우 99.50%, 형태소의 경우 99.99% 일치하였다.[14]

---

누락된 정보가 없이 "빠짐없이" 구성되었는가를 평가하는 지표이다. 보다 자세한 논의는 Manning and Schütze (1999:268) 참조.

14) 현재로서는 오차의 원인은 몇 가지로 추정된다. 첫째는 한글 코드 변환 과정의 정보 손실 가능성이다. Xavier는 ANSI C++로 구현이 되어 있고 본래 Unicode로 작성된 SKT를 EUC-KR 코드로 변환하여 작업하였다. 반면 태그 계산만을 위해 작성한 프로그램 SKT_tagCount.pl은 SKT를 UTF-8으로 변환하여 사용하였고 프로그램자체는 Perl로 구현하였다(각주 5). 대부분의 경우는 문제가 없을 것이나 SKT에 사용된 일부 특수 한글부호 코드의 경우 변환과정에 손실이 생겼을 가능성이 있다고 본다. 오차 발생의 두 번째 가능성으로는 SKT에 내재하는 오류가 원인일 가능성으로, 특히 SKT의 형식상 오류(syntax errors)가 적지 않게 발견되었다. UTF-8으로 변환한 뒤 계산한 것에 따르면, 닫는 괄호 앞에 빈칸이 들어간 경우가 1,346회 (9유형), 태그 앞에 빈칸이 들어간 경우가 45회 (1유형), 그리고 기타 8가지가 발견되었다. 또한 구 범주와 형태소분석 부분 사이에 들어간 + 의 경우도 일부는 원칙상의 문제이겠으나 (각주 6 참조), 단순한 실수로 추정되는 경우도 적지 않게 발견되었다. 내용상의 오류로는 예를 들어 구 범주 이름이 잘못된 경우가 최대 추정치 403회 (142유형)가 발견되었다. SKT가 Penn Treebank처럼 하나의 주요 표준이 되려면 이러한 문제가 별도로 더 철저히 검증되고 수정되어야 할 것으로 판단된다. 오차 발생의 세 번째 가능성으로는 Xavier나 SKT_tagCount.pl 자체의 오류를 들 수 있다. Xavier의 경우는 앞에서 언급한 SKT의 형식상의 오류에 대하여 수작업을 포함하는 일부 사전처리를(preprocess) 거쳤기 때문에 그 과정에서 오차가 발생했을 가능성이 있다. 물론 Xavier나 SKT_tagCount.pl 프로그램 자체의 오류 가능성도 배제할 수는 없으나, 객관성을 높이기 위해서 별도 프로그래머가 각자 작성한 뒤 그 결과를 비교하는 절차를 밟았고, 프로그램 자체의 알고리즘도 가능한 모든 경우가 각각 별도의 파일로 생성

Xavier에 대한 또 다른 검증은, “정확하게”에 대한 검증으로, 임의로 선정한 특정 예를 기준으로 Xavier의 추출 결과와 [한마루]의 검색 결과를 비교해 본 후 그 결과가 상호 일치하는지를 확인해 보았다. 우선 Xavier에서 ‘S_? → AP_? S_?’의 규칙들을 모두 추출하면 아래와 같이 17회 적용된 것으로 나타났으며 관련 문장을 모두 추출할 수 있었다.

[표 3] ‘S_? → AP_? S_?’규칙 빈도

| MN | LDN | RDN | 빈도 |
|---|---|---|---|
| S | AP_AJT | S | 14 |
| S_CMP | AP_AJT | S_CMP | 1 |
| S_MOD | AP_AJT | S_MOD | 2 |

이번에는 [한마루]에서 위 구문에 해당하는 검색식 ‘문장 → 부사구 문장’을 입력하여 검색한 결과 마찬가지로 17개의 동일한 문장이 추출되었다. 아래는 그 중 일부를 보인 것이다.

(2) a. 벌써부터 異變, 대역전극이 속출하고 있지 않은가.
b. 아직까지는 선수로서보다 진행이 더 큰 역할.
c. [S_CMP] 벌써부터 브라질이 그리워진다고
d. [S_MOD] 멀리서 다시 바라볼 수 있는
e. ....

그러나 동일 검색식으로 Xavier에서 추출한 결과와 [한마루]에서 추출한 결과가 숫자상으로 약간씩 어긋나는 경우도 일부 발견되었다.

---

되도록 작성한 뒤에 그 파일을 점검하는 절차를 거쳤다.
이와 같은 여러 원인으로 발생할 수 있는 오차를 완전히 해소한다는 것은 본 연구의 범위를 벗어나는 일이고, 또한 오차의 범위가 본 연구의 목표에 비추어 별로 심각한 수준은 아니라고 판단하여 현 상태에서 나머지 연구를 진행하였다.

모두 Xavier검색 결과가 [한마루] 검색 결과보다 약간 많은 경우들로, 그 반대의 경우, 즉 [한마루]에서는 검색이 되나 Xavier에서는 검색되지 않는 예는 발견되지 않았다. 그러한 차이를 확인하기 위해 검색식 5개를 선정한 후 각 결과를 하나씩 점검해 본 결과 이는 [한마루]쪽의 문제로 밝혀졌다.[15)16)]

이상 두 가지 방식의 검증을 통해 Xavier 모듈이 사실에 부합되는 안정적인 결과를 산출해 주는 것으로 판단하였다.

## 2.3 세종 구문분석 말뭉치의 대표성

SKT를 활용하기 위해서는 SKT가 말뭉치로서 대표성을 갖추고 있는가라는 질문에 대한 검토도 필요할 것이다. 이러한 차원의 보다 본격적인 논의는 별도의 연구를 통해 앞으로도 반복해서 이루어질 것으로 기대되나, 추출하게 될 CFG 규칙이 말뭉치의 특성을 직접 반영하게 된다는 점에서, 적어도 본 연구에서도 최소한도의 검토는 필요하다.

SKT가 한국어 문장을 포괄적으로 반영하는가라는 대표성 검증의

15) 물론 이는 일부에 대한 검증에 국한하여 내린 결론이다. [한마루]의 검색 결과 숫자의 정확성에 대한 종합적인 검증은 별도로 이루어져야할 것이다.

16) 그 중 두 가지를 예로 들어 보면, 우선 'S → AP S_PRN'이나 'S → AP AP'의 경우 [한마루]에서는 검색 결과가 없다고 나오나 Xavier는 각기 1회 출현한 것으로 나왔고 실제 그 두 가지 예가 SKT에 들어 있다는 점을 확인하였다. 또 다른 예로, 'S_?→AP_? VP_?'의 형태는 [한마루]에서는 13회 출현하였으나, Xavier 추출 결과는 총 27개로 밝혀졌다.

| MN | LDN | RDN | 빈도 |
|---|---|---|---|
| S | AP | VP | 19 |
| S | AP_SBJ | VP_MOD | 3 |
| S_MOD | AP_SBJ | VP_MOD | 2 |
| S_MOD | AP | VP_MOD | 2 |
| S_CMP | AP | VP_CMP | 1 |

차원에서 SKT와 세종 형태분석 말뭉치를 비교해 보도록 하자. 2007년 12월 배포 CD에 담긴 세종 형태분석 말뭉치는 천만 어절이 넘는 자료를 수록하고 있으며 균형 말뭉치를 표방하고 있다. 천만 어절 이상이면 통상 해당 언어에서 대표성을 지니고 있다고 판단하는 점을 고려해 볼 때, 이와 비교해 보면 SKT가 어느 정도 대표성을 지니는가를 가늠할 수 있을 것이다. 아래의 표는 두 말뭉치에서 상위 빈도어 목록을 추출한 것이다. 옆 수치는 누적비율을 말한다.

[표 4] 세종 형태분석 말뭉치와 SKT 상위 빈도어 비교

| 순위 | 명사 | | | | 동사 | | | | 형용사 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 형태 | | 구문 | | 형태 | | 구문 | | 형태 | | 구문 | |
| 1 | 말 | 1.07% | 사람 | 1.06% | 하 | 7.75% | 하 | 7.07% | 없 | 17.91% | 없 | 16.92% |
| 2 | 사람 | 2.02% | 말 | 2.11% | 있 | 14.90% | 있 | 13.83% | 같 | 27.24% | 같 | 26.55% |
| 3 | 때 | 2.74% | 때 | 2.87% | 되 | 19.28% | 되 | 18.47% | 그렇 | 33.02% | 그렇 | 32.39% |
| 4 | 일 | 3.34% | 일 | 3.44% | 보 | 21.52% | 보 | 20.79% | 크 | 37.65% | 많 | 37.40% |
| 5 | 생각 | 3.84% | 생각 | 3.97% | 가 | 23.58% | 대하 | 22.74% | 많 | 42.03% | 크 | 41.73% |
| 6 | 사회 | 4.20% | 세계 | 4.36% | 대하 | 25.57% | 가 | 24.19% | 좋 | 45.88% | 좋 | 44.83% |
| 7 | 문제 | 4.53% | 사회 | 4.73% | 위하 | 27.10% | 위하 | 25.58% | 어떻 | 48.39% | 새롭 | 47.74% |
| 8 | 속 | 4.86% | 자신 | 5.09% | 받 | 28.41% | 받 | 26.95% | 이렇 | 50.72% | 어떻 | 50.22% |
| 9 | 집 | 5.15% | 문제 | 5.43% | 알 | 29.65% | 알 | 28.17% | 새롭 | 52.67% | 이렇 | 52.52% |
| 10 | 자신 | 5.44% | 속 | 5.76% | 보이 | 30.79% | 들 | 29.30% | 높 | 54.39% | 다르 | 54.34% |
| 11 | 경우 | 5.72% | 문화 | 6.08% | 들 | 31.84% | 보이 | 30.40% | 어렵 | 56.09% | 높 | 55.69% |
| 12 | 시작 | 5.96% | 집 | 6.37% | 오 | 32.85% | 오 | 31.45% | 다르 | 57.73% | 어렵 | 57.03% |
| 13 | 세계 | 6.20% | 민족 | 6.66% | 따르 | 33.77% | 만들 | 32.41% | 작 | 59.05% | 쉽 | 58.22% |
| 14 | 앞 | 6.43% | 눈 | 6.94% | 살 | 34.68% | 살 | 33.34% | 쉽 | 60.29% | 작 | 59.41% |
| 15 | 아이 | 6.66% | 시작 | 7.21% | 모르 | 35.55% | 나오 | 34.24% | 길 | 61.33% | 젊 | 60.37% |
| 16 | 정부 | 6.88% | 아이 | 7.49% | 나오 | 36.43% | 모르 | 35.13% | 아름답 | 62.36% | 아름답 | 61.30% |
| 17 | 사실 | 7.10% | 영화 | 7.74% | 쓰 | 37.29% | 쓰 | 35.98% | 깊 | 63.22% | 길 | 62.20% |
| 18 | 눈 | 7.32% | 나라 | 8.00% | 만들 | 38.13% | 그러 | 36.81% | 어리 | 63.96% | 깊 | 63.05% |
| 19 | 뒤 | 7.53% | 인간 | 8.25% | 그러 | 38.84% | 따르 | 37.50% | 젊 | 64.70% | 어리 | 63.86% |
| 20 | 인간 | 7.75% | 경우 | 8.50% | 지나 | 39.53% | 나 | 38.13% | 멀 | 65.41% | 강하 | 64.57% |
| 21 | 이상 | 7.96% | 시간 | 8.74% | 가지 | 40.20% | 듣 | 38.75% | 가깝 | 66.09% | 멀 | 65.25% |
| 22 | 시간 | 8.17% | 소리 | 8.99% | 통하 | 40.86% | 의하 | 39.36% | 힘들 | 66.76% | 가깝 | 65.92% |
| 23 | 전 | 8.37% | 돈 | 9.23% | 먹 | 41.48% | 가지 | 39.97% | 적 | 67.43% | 힘들 | 66.50% |
| 24 | 소리 | 8.57% | 교육 | 9.46% | 듣 | 42.03% | 통하 | 40.54% | 낮 | 68.01% | 적 | 67.05% |
| 25 | 여자 | 8.78% | 사실 | 9.69% | 나 | 42.57% | 주 | 41.09% | 아프 | 68.54% | 나쁘 | 67.57% |
| 26 | 필요 | 8.97% | 선생 | 9.91% | 주 | 43.08% | 지나 | 41.62% | 심하 | 69.06% | 짧 | 68.08% |
| 27 | 정도 | 9.16% | 운동 | 10.14% | 의하 | 43.58% | 먹 | 42.14% | 넓 | 69.53% | 있 | 68.58% |
| 28 | 날 | 9.34% | 아버지 | 10.36% | 만나 | 44.03% | 찾 | 42.66% | 빠르 | 70.01% | 아프 | 69.06% |
| 29 | 손 | 9.52% | 전 | 10.57% | 느끼 | 44.48% | 갖 | 43.14% | 나쁘 | 70.46% | 커다랗 | 69.53% |
| 30 | 마음 | 9.70% | 역사 | 10.79% | 찾 | 44.92% | 만나 | 43.60% | 늦 | 70.91% | 싫 | 69.98% |

우선 명사의 경우에는 약간의 차이를 보이는 것을 알 수 있다. 예컨대, '문화', '민족', '역사' 등의 명사가 SKT에서는 30위권 안에 들지만 형태분석 말뭉치에서는 그렇지 않다. 이들은 대체로 신문 사설 또는 논설문 등의 텍스트에서 주로 출현하는 단어라고 생각한다면, SKT가 형태분석 말뭉치에 비해 평균적으로 건조체 양식의 글을 더 다루고 있다고 추정할 수 있다.[17] 즉, 장르에 따른 영향이 명사 출현 빈도에 작용하고 있는 것이다. 그러나 동사와 형용사의 경우에는 양자의 분포가 거의 유사하다. 문장 구조의 특성은 주로 용언에 의해 결정된다는 점을 고려하면, 문체의 성격에 따라 일부 차이는 보이고 있으나 SKT는 한국어의 특성을 충분히 반영하고 있다고 볼 수 있다. 즉 80만 어절의 SKT가 천만어절 이상의 세종 형태분석 말뭉치와 비슷한 수준의 대표성을 지니고 있다고 잠정적인 결론을 내릴 수 있다.

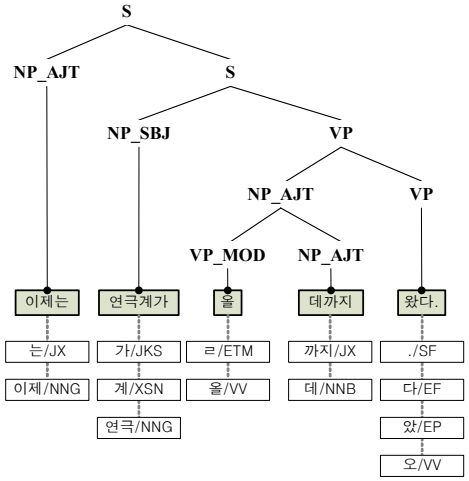
이번에는 SKT의 전체 형태소 및 문장 수 대비 비종단/종단 규칙별 비율을 제시한다. Xavier를 이용하여 추출한 각 규칙 항목별 비율은 아래와 같다. 이는 SKT의 특성을 또 다른 차원에서 나타내준다.

[표 5] SKT 기초 통계 (3)

| 구분 | 수치 | 결과치 |
|---|---|---|
| 비종단규칙토큰 / 형태소토큰 | 774,045 / 1,868,152 | 0.41 |
| 비종단규칙토큰 / 문장수 | 774,045 / 77,121 | 10.04 |
| 종단규칙토큰 / 형태소토큰 | 851,126 / 1,868,152 | 0.46 |
| 종단규칙토큰 / 문장수 | 851,126 / 77,121 | 11.04 |
| 비종단규칙타입 / 형태소타입 | 3,626 / 50,844 | 0.07 |
| 비종단규칙타입 / 문장수 | 3,626 / 77,121 | 0.05 |
| 종단규칙타입 / 형태소타입 | 215,235 / 50,844 | 4.23 |
| 종단규칙타입 / 문장수 | 215,235 / 77,121 | 2.79 |

17) 이는 2007년도 12월에 배포된 결과물 CD-ROM에서 제시된 코퍼스 맵에서도 확인할 수 있는 사항이다.

위 표에서 비종단 규칙과 종단 규칙의 차이를 앞의 예 (1)을 참조하여 설명하기로 하자. 편의상 (1)의 수형도를 아래에 다시 제시하였다.




위 수형도에서 상위 S 절점 2개와 상위 VP 절점 그리고 두번째 NP_AJT 절점만 비종단 절점으로 처리되며, 그 외 절점(NP_AJT, NP_SBJ, VP_MOD, NP_AJT, VP)은 모두 종단 절점으로 처리된다.[18]

## 3. 문맥자유문법(CFG) 규칙

세종 구문분석 말뭉치에서 추출한 CFG 규칙은 필연적으로 세종 구문분석 말뭉치의 구성 방식을 직접 반영한다. 그런 점에서 세종 구문분석 말뭉치 구성상 주요 특성을 우선 살펴볼 필요가 있다(3.1절). 이어지는 소절에서는 S 규칙(3.2절), VP 규칙(3.3절), NP 규칙(3.4절), 기타 규칙(3.5절)의 순서로 논의를 전개한다.

---

18) 여기에서 제시된 수치는 모두 Xavier를 이용하여 추출한 것으로, 다른 방식으로 추출한 2.1절에서의 수치와는 약간의 차이가 있다. 각주 14 및 관련 본문 참조

### 3.1 세종 구문분석 말뭉치 작성 원칙 및 특성

김홍규 외 (2003) 및 강범모・김의수 (2004)에 따르면 세종 구문분석 말뭉치는 아래와 같은 기본 원칙하에 구축되었다.

(4) a. 자연언어처리에서 일반적으로 고려되는 일관성 유지와 효율성 제고에 초점을 두되, 일반언어학적 관점에서도 크게 벗어나지 않도록 한다.
b. 표층 구조를 중시하여 분석한다.
c. 이분지 가설을 취하며 다분지를 허용하지 않는다.
d. 공범주를 인정하지 않는다.
e. 어절을 분석의 기본 단위로 한다.
f. 보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.
g. 원칙적으로 접속과 내포를 구별하지 않으며 접속절은 모두 부사절로 분석한다. (다만 명사구 접속만은 인정한다.)
h. 하나의 주어가 모문과 내포문 모두에 관련되어 있어 구문 분석의 중 의성이 발생할 경우, 모문의 주어로 우선 분석한다.

위의 원칙에 따라 작성된 세종 구문분석 말뭉치의 주요 특징을 살펴보는 방편으로 간단한 문장을 예로 하여 기존 주요 언어학 이론에 따른 수형도와 비교해 보자. 아래 예는 실제 SKT에 나오는 문장이다. 이어지는 그림은 [한마루]에서 제시하는 해당 문장의 수형도이다.[19]

(5) ; “내 권리는 누구도 못 빼앗는다!”
(S (L ″/SS)
(S (S (NP_OBJ (NP_MOD 나/NP + 의/JKG)
(NP_OBJ 권리/NNG + 는/JX))

19) 세종 구문분석 말뭉치에 사용된 범주는 아래와 같다 (김홍규 외 2003). 아래 표에서 X는 ‘의사구’를 뜻하는 바 명확히 규정하기 어려운 경우들로, 주로 조사나 어미가 단독으로 어절을 문장 부호등을 의미한다.

(S (NP_SBJ 누구/NP + 도/JX)
(VP (AP 못/MAG)
(VP 빼앗/VV + 는다/EF + !/SF ))))

(5)의 문장은 주어와 목적어의 순서가 도치된 주제화 구문에 해당한다. 흔히 언어학 일반에서는 위 문장의 목적어에 해당하는 '나의 권리는'이 흔적(trace)을 남기고 문장의 다른 곳으로 이동한 것으로 파악한다. 즉, 위 문장은 기저 구조와 표층 구조가 다르다. 그러나 전술한 바와 같이 SKT는 철저하게 문장의 표면형만을 대상으로 기술한다.[20]

위는 다음 두개의 수형도와 여러 가지 점에서 비교가 된다.

우선 지배-결속(Government and Binding, GB)이론 틀 내에서는 위와 같은 주제화 구문을 아래와 같이 파악한다 (이윤표 1989:92).

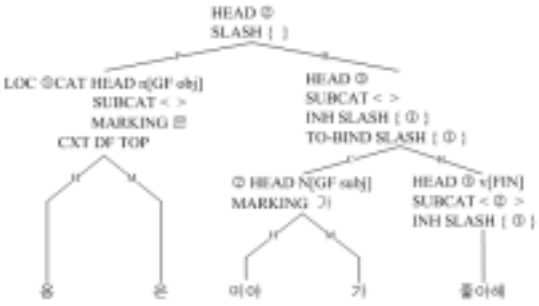
| | 범주 | 사례 |
|---|---|---|
| S | 문장 | |
| Q | 인용절 | 인용부호("") 안에 들어 있는 두 개 이상의 문장 |
| NP | 체언구 | 체언(명사, 대명사, 수사) |
| VP | 용언구 | 용언(동사, 형용사, 보조용언) |
| VNP | 긍정지정사구 | 긍정지정사 '이다' |
| AP | 부사구 | 부사 |
| DP | 관형사구 | 관형사 |
| IP | 감탄사구 | 감탄사 |

20) 이와는 대조적으로 PKT에서는 기저 구조를 반영하여 분석 정보에 흔적을 명시하고 있다.

(6) 영희는 철수가 사랑한다.


또한 한국어 핵어 구구조문법(Head-driven Phrase Structure Grammar, HPSG)에서 판단하는 주제화 구문의 구조는 아래와 같다 (장석진 1995:122).

(7) 용은 미아가 좋아해.




이와 같은 GB 및 HPSG 수형도에 비추어 SKT의 특성을 살피면 다음과 같다. 우선, 어절이 분석의 기준이므로 GB 또는 HPSG에서처럼 문법형태소 중심의 구성 정보가 표면에 직접 드러나지 않는다. 다만, 각 종단 절점에서 세부 정보를 확인할 수 있을 뿐이다. 또한 표면형 기준이므로 GB에서처럼 이동흔적이나 기타 공범주를 상정한

이론의 관점에서는 어느 정도 '왜곡된' CFG 규칙이 SKT에서는 사용되는 셈이다.[21] 한편, 표면형을 중심으로 하는 HPSG와는 공범주를 가정하지 않는 점, 표면형을 분석의 기본 골자로 하는 점 등에서 유사성을 가진다. 이는 SKT와 HPSG 모두 전산적 구현을 염두에 둔 것이라는 점에서 표층을 중심으로 하는 것이 보다 유용하기 때문이다. 단, HPSG 수형도에서는 충족되어야 할 논항 정보를 위해 SLASH 자질 구조가 이용되는 반면, SKT에서는 그 자체만으로 더 충족되어야 할 정보가 무엇인지 알기가 어렵다.[22]

세종 구문분석 말뭉치의 또 하나의 주요 특징은 각 절점별로 범주뿐만 아니라 그 범주가 해당 수형도에서 어떤 기능을 하고 있는지도 병기하고 있다는 점이다. 예를 들어, 체언구의 경우 NP라는 범주를 준 경우도 있지만 많은 경우는 그 체언구가 어떤 기능으로 쓰였느냐에 따라 SBJ, OBJ, MOD 등의 기능표지를 덧붙여 놓았다. 문법기능은 원칙적으로 수형도의 범주관계로부터 도출될 수 있는 개념이라는 점을 고려하여, 본 연구에서의 CFG 규칙에 대한 논의에서는 이러한 기능표지는 일단 배제하였다.[23] 다만 세부 항목별 논의에서 필요할 경우 그 기능표지의 분포도 논의에 포함시키기로 한다.[24]

지금까지는 세종 구문분석 말뭉치의 구문 구성상 특징에 대하여

---

21) 심사위원 중의 한 분도 이러한 점에 대한 우려를 강하게 제기하였다. "접속문을 인정하지 않고 모두 부사절로 처리한다면, 등위 접속문에 대해 정당한 처리라고 하기 어렵다"는 점이나 "생략된 성분 등으로 인해 생기는 통계적 허구성" 등의 문제점이 제기 되었는 바, 필자들도 일단 그러한 우려에는 동의를 한다. 그러나 그러한 "왜곡"이나 "허구"가 어느 정도인가에 대한 논의는 별도로 검증되어야 할 문제라고 본다.

22) 예를 들어, SKT로부터 그 안에 사용된 용언별 논항 구조를 추출하고자 할 때 이러한 점은 직접 도출을 어렵게 하는 제약이 된다고 본다.

23) 세종 구문분석 말뭉치를 기반으로 하여 그러한 도출 가능성을 검증해 보는 것도 한 가지 흥미로운 연구가 될 것이다. 이어지는 소절에서 각 규칙별 분포적 특성을 논하는 과정에 관련 논의가 일부 포함된다.

24) SKT에서 사용된 문법기능표지는 아래와 같다 (김홍규 외 2003).

간략히 살펴보았다. 이제 이러한 구성상의 특성을 바탕으로 각 규칙별로 분포적 특성을 검토해 보기로 한다. 적지 않은 숫자의 규칙이 논의의 대상이 되므로, 서술의 편의상 MN의 범주를 중심으로 논의를 전개하기로 한다. 우선 각 MN을 기준으로 CFG 규칙의 출현빈도를 살펴보면 아래와 같은 분포를 보인다.[25]

[표 6] MN 기준 CFG 규칙 분포

| MN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 282,292 | 36.47% | 36.47% |
| VP | 277,624 | 35.87% | 72.34% |
| S | 162,373 | 20.98% | 93.32% |
| 기타 | 51,709 | 6.68% | 100.00% |
| 소계 | 773,998 | | |

가장 많은 출현빈도를 보이는 것은 NP 절점이며, VP 절점 역시 그와 거의 유사한 빈도를 보이고 있다. 그 다음은 S 절점이며 이상의 세 절점의 빈도가 전체의 93%를 초과하고 있다. 이어지는 소절에서 이 S 규칙, VP 규칙, NP 규칙 순서로 세밀히 살펴보고, 기타 MN 항목 중에서 대다수를 차지하는 VNP와 AP는 묶어서 '기타 규

| | 범주 | 사례 |
|---|---|---|
| SBJ | 주어 | 주격 체언구, 명사 전성 용언구, 명사절 (NP_SBJ, VP_SBJ, S_SBJ) |
| OBJ | 목적어 | 목적격 체언구, 명사 전성 용언구, 명사절 (NP_OBJ, VP_OBJ, S_OBJ) |
| CMP | 보어 | 보격 체언구, 명사 전성 용언구, 인용절 (NP_CMP, VP_CMP, S_CMP) |
| MOD | 체언수식어 | 관형격 체언구, 관형형 용언구, 관형절 (NP_MOD, VP_MOD, S_MOD) |
| AJT | 용언수식어 | 부사격 체언구, 문말어미+부사격조사 (NP_AJT VP_AJT, S_AJT) |
| CNJ | 접속어 | 접속격 체언(NP_CNJ) |
| INT | 독립어 | 체언 (NP_INT) |

25) 아래 표에서 '기타'에 속하는 규칙들의 분포는 3.5절에 제시한다.

칙’ 소절에서 서술한다.

### 3.2 S 규칙

세종 구문분석 말뭉치에 나오는 S 규칙은 모두 162,373회 적용된 것으로 파악된다. 그 규칙의 유형은 63개로 토큰/타입 비율은 2,577이나, 상위 6개로 한정할 경우엔 25,000 가까이 되고, 나머지 57개는 223개 수준이 되는 것으로 보아 상위 소수의 규칙이 월등히 많은 빈도로 적용된다는 점을 알 수 있다. 즉, 빈도수 기준 상위 6개의 S 규칙은 아래에 표에서 보듯 합계 92% 이상으로 출현하고 있다. 나머지 57개의 규칙은 빈도로는 2,348개~1개, 비율로 보아 1.45%~0.00% 정도로만 분포되어 있다.[26]

[표 7] 상위 6개 S 규칙 분포

| MN | LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|---|
| S | NP | VP | 72,566 | 44.69% | 44.69% |
| S | NP | S | 28,551 | 17.58% | 62.27% |
| S | NP | VNP | 14,291 | 8.80% | 71.07% |
| S | VP | S | 11,948 | 7.36% | 78.43% |
| S | AP | S | 11,502 | 7.08% | 85.51% |
| S | S | S | 10,766 | 6.63% | 92.14% |
| 소계 | | | 149,624 | | |

한편, S 규칙의 범주 S가 쓰인 기능별 분포는 다음과 같다.

26) 표에 제시된 모든 수치는 소수점 이하 세 번째 자리에서 반올림한 것을 이용하였다. 이에 따라 소수점 이하의 단위에서 약간의 변동 폭이 있을 수 있다.

[표 8] 범주 S의 기능별 분포

| 기능 | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| S | 124,105 | 76.43% | 76.43% |
| S_MOD | 29,296 | 18.04% | 94.47% |
| S_CMP | 5,335 | 3.29% | 97.76% |
| S_OBJ | 1,623 | 1.00% | 98.76% |
| S_AJT | 1,142 | 0.70% | 99.46% |
| S_SBJ | 529 | 0.33% | 99.79% |
| S_PRN | 257 | 0.16% | 99.95% |
| S_CNJ | 75 | 0.05% | 100.00% |
| S_INT | 11 | 0.01% | 100.00% |
| 소계 | 162,373 | | |

위 표를 보면 기능이 명시되지 않은 경우가 대부분으로, 일반적으로 문장이 수형도상 최상위 절점이 된다는 점에서 자연스러운 결과이다. S가 관계절을 포함하는 체언수식절(S_MOD)로는 18% 이상 활용된 반면, 같은 수식기능을 하되 용언수식절(S_AJT)로 쓰인 경우는 0.7%로 상대적으로 미미한 수치였다. 절이 논항으로 쓰인 경우(S_CMP, S_OBJ, S_SBJ)는 모두 합하면 5% 가까이 된다.

다음 이어지는 두 소절에서는 S 삼각구조, 즉 S 다시쓰기 규칙(S rewrite rules)의 LDN과 RDN을 기준으로 관련 CFG 규칙들을 차례대로 살펴보기로 한다.

### 3.2.1 S 규칙 좌측 절점 (LDN) 기준

LDN 범주를 기준으로 S 규칙의 적용 빈도를 보면 다음과 같다.

[표 9] LDN 기준 S 규칙 분포

| LDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 118,231 | 72.81% | 72.81% |
| S | 14,545 | 8.96% | 81.77% |
| VP | 13,078 | 8.05% | 89.82% |
| AP | 11,535 | 7.10% | 96.92% |
| L | 2,411 | 1.48% | 98.40% |
| VNP | 1,760 | 1.08% | 99.48% |
| IP,X,DP,Q,LP,R | 713 | 0.51% | 100.00% |
| 소계 | 162,373 | | |

위 표를 보면, S 규칙의 LDN에 나오는 범주 중 NP가 72.81%로 단연 다수를 점하고 있고, S, VP, AP 등이 각각 7~9%사이의 빈도로 발생하고 있다. 그리고 그 넷의 합이 97%정도를 차지하면서 나머지 범주는 거의 무시할 수준으로만 나타나고 있다.

이번에는 S 규칙의 LDN 범주를 기준으로 각 범주별로 그것의 자매 절점(Sister Node, 이하 'SN')이 어떤 것들이 어떤 비율로 나오는지 살펴보기로 하자. 첫 번째로 LDN에 가장 많은 빈도로 출현하는 NP 범주의 SN에 나오는 범주의 분포부터 살펴본다.

[표 10] LDN이 NP일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | VP | 72,566 | 61.38% | 61.38% |
| NP | S | 28,551 | 24.15% | 85.53% |
| NP | VNP | 14,291 | 12.09% | 97.62% |
| NP | NP | 1,960 | 1.66% | 99.28% |
| NP | AP | 760 | 0.64% | 99.92% |
| NP | Q,R,X,DP,IP | 103 | 0.09% | 100.00% |
| NP 소계 | | 118,231 | | |

위의 분포를 보면 S 규칙에서 LDN이 NP일 때 RDN이 VP인 경우가 61.38%로 다수를 차지하고 있고, 같은 술어성 범주인 VNP까

지 합치면 73% 이상이 된다. 이 경우 LDN NP에는 주어기능표지(SBJ)가 부착되어 있을 거라는 예측이 가능한 바, VP에서는 99.87%가, VNP에서는 99.90%의 LDN NP에 SBJ표지가 붙어 있는 것으로 확인되었다.[27] 나머지 중에서는 RDN이 S가 되는 경우가 대다수로, 이 경우 LDN NP는 용언수식어(AJT, 63%), 주어(SBJ, 23%),[28] 목적어(OBJ, 10%)등의[29] 기능 표지를 달고 있는 경우가 대부분이다. 'S → NP NP'규칙에서는 'LDN-NP'의 98%가 주어 기능 표지를 달고 있다. 'RDN-VNP'는 긍정지정사 '이다'가 붙은 명사구이고, 'RDN-NP'는 체언구가 서술어로 쓰인 경우로 대개 긍정지정사가 표면에 드러나지 않은 경우로 보인다.

다음에는 LDN이 S인 S 규칙을 살펴보자.

[표 11] LDN이 S일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| S | S | 10,766 | 74.02% | 74.02% |
| S | R | 2,278 | 15.66% | 89.68% |
| S | X | 829 | 5.70% | 95.38% |
| S | VP | 322 | 2.21% | 97.59% |
| S | NP | 136 | 0.94% | 98.53% |
| S | VNP | 129 | 0.89% | 99.42% |
| S | AP | 85 | 0.58% | 100.00% |
| S | IP,Q,L | 32 | 0.22% | 100.00% |
| S 소계 | | 14,545 | | |

27) 지면의 제약 상 관련 수치를 모두 표로 제시할 수는 없으므로 주요 수치 중심으로 표로 제시하되, 논의 상 필요한 추가 수치는 본문 서술부분에서도 언급을 한다. 즉 어떤 수치는 본문의 표에 직접 제시되었거나 표에 제시된 수치를 근거로 한 것이고, 또 다른 경우는 본문 서술에만 제시하도록 한다. 경우에 따라서는 소수점 이하는 반올림한 수치를 제시하기도 한다.

28) 'S → NP S' 규칙에서, NP가 주어(SBJ)의 기능을 하는데도 SN으로 S를 취한다는 것은 그 S에 아직 주어기능을 하는 것이 들어 있다는 의미로, 예를 들어 다중주어 구문이 이에 해당한다.

29) NP가 목적어(OBJ) 기능을 하면서 SN으로 S를 취하는 경우는 목적어가 주어 앞으로 이동한 경우뿐만 아니라 "[S[NP_OBJ현실을] [S[NP_SBJ조망할 수] 있을까]]"같은 구문도 포함한다.

S 규칙에서 LDN이 S일 경우 RDN의 74%는 S가 된다. 이때 LDN S는 '-어', '-고', '-면', '-면서', '-는데' 등으로 끝나는 것들로 전통문법에서는 부사절로도 분류될 수 있는 것들이다. 범주 R이나 L은 문장 부호로, 'LDN-S'의 SN인 'RDN-R'은 문장을 닫는 괄호나 인용부호 등이 적지 않게 쓰이고 있음을 보여준다.

이어서 LDN이 VP인 S 규칙을 살펴보자.

[표 12] LDN이 VP일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| VP | S | 11,948 | 91.36% | 91.36% |
| VP | VP | 872 | 6.67% | 98.03% |
| VP | VNP | 168 | 1.28% | 99.31% |
| VP | NP | 51 | 0.39% | 99.70% |
| VP | AP | 26 | 0.20% | 99.90% |
| VP | R,X,L | 13 | 0.10% | 100.00% |
| VP 소계 | | 13,078 | | |

위 표를 보면 S 규칙에서 LDN이 VP인 경우 RDN으로 S가 나오는 경우가 압도적으로 많다는 점을 알 수 있다. 사실 이 경우 VP는 실제로는 문장임에도 표면에 주어가 나타나지 않기 때문에 VP라는 범주가 부여된 경우다. 따라서 내용적으로는 앞에서의 'S → S S' 규칙과 마찬가지 경우라 할 수 있다. 'S → VP VP' 규칙이 적용되는 경우는 주로 [S[VP객관화되기가] [VP어렵다]]처럼 'LDN-VP'가 주어로 쓰인 경우가 93%에 해당되며, 이러한 점은 S → VNP VP의 경우에도 마찬가지로 VNP가 주어로 쓰인 경우가 97%에 이른다.

끝으로 S 규칙에서 LDN이 AP인 경우를 보자.

[표 13] LDN이 AP일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| AP | S | 11,502 | 99.71% | 99.71% |
| AP | VP | 24 | 0.21% | 99.92% |
| AP | NP,VNP,AP,L | 9 | 0.08% | 100.00% |
| AP 소계 | | 11,535 | 100.00% | |

위 표에 따르면 세종 구문분석 말뭉치에서 MN이 S고 LDN이 AP면 RDN이 S가 될 확률은 거의 100%에 이른다.

### 3.2.2 S 규칙 우측 절점 (RDN) 기준

S 규칙의 RDN 범주를 기준으로 S 규칙의 적용 빈도를 보면 VP, S 두 범주가 87%정도를 점하고 있고, 이어서 VNP가 9%를 차지하여 상위 세 범주가 96%라는 절대 다수의 분포적 특성을 보인다.

[표 14] RDN 기준 S 규칙 분포

| RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| VP | 74,049 | 45.60% | 45.60% |
| S | 67,381 | 41.50% | 87.10% |
| VNP | 14,653 | 9.02% | 96.12% |
| R | 2,316 | 1.43% | 97.55% |
| NP | 2,187 | 1.35% | 98.90% |
| X | 853 | 0.53% | 99.43% |
| AP | 833 | 0.51% | 99.94% |
| Q,IP,DP,L | 99 | 0.06% | 100.00% |
| 소계 | 162,373 | | |

즉, S 규칙의 RDN 범주 중 거의 대부분이 VP, S, VNP라는 점은 문장의 RDN이 주로 서술어이거나 '주어+서술어'라는 말로 이는 일반적인 기대에 부합한다.

그러면 위 표에 나오는 RDN 범주를 기준으로 각 범주별로 그것

의 SN, 즉 관련 S 규칙의 LDN이 어떤 분포를 보이는지 살펴보자. RDN이 VP일 때의 분포는 다음과 같다.

[표 15] RDN이 VP일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | VP | 72,566 | 98.00% | 98.00% |
| VP | VP | 872 | 1.18% | 99.18% |
| S | VP | 322 | 0.43% | 99.61% |
| VNP | VP | 217 | 0.29% | 99.90% |
| L,AP,IP,X,DP,LP,Q | VP | 72 | 0.10% | 100.00% |
| 소계 | | 74,049 | | |

위 표는 S 규칙에서 RDN이 VP일 때 LDN이 NP일 가능성이 98%에 이른다는 점을 보이고 있다. 즉, MN이 S고 RDN이 VP라면 LDN은 거의 확실히 NP라고 볼 수 있다. 또한 RDN이 VP면서 LDN이 NP인 경우 99.87%가 주어표지가 부착되어 있다는 앞 절에서의 일반화를 결합하면, LDN의 98%이상이 NP 주어가 되는 셈이다.

이어서 RDN이 S인 S 규칙을 살펴보자.

[표 16] RDN이 S일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | S | 28,551 | 42.37% | 42.37% |
| VP | S | 11,948 | 17.73% | 60.10% |
| AP | S | 11,502 | 17.07% | 77.17% |
| S | S | 10,766 | 15.98% | 93.15% |
| L | S | 2,348 | 3.48% | 96.63% |
| VNP | S | 1,478 | 2.19% | 98.82% |
| IP | S | 588 | 0.87% | 99.69% |
| X,DP,Q,R,LP | S | 200 | 0.30% | 100.00% |
| 소계 | | 67,381 | | |

3.1.1절에서 보았듯 S 규칙에서 LDN이 NP인 경우 RDN이 S일 가능성은 24%정도인 반면, RDN이 S일 때 LDN이 NP가 될 가능성

은 42%가 넘는다. 이러한 불균형은 LDN이 VP, AP인 경우에 더 극단적으로 나타나는 바, 위 표에서 보듯 RDN이 S일 때 LDN이 VP나 AP가 될 가능성은 17%대에 머물러 있는 반면, LDN이 VP나 AP일 때 RDN이 S가 될 확률은, 각각 91%, 99% 이상이 된다.

끝으로 S 규칙에서 RDN이 VNP인 경우를 보자.

[표 17] RDN이 VNP일 경우 S 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | VNP | 14,291 | 97.53% | 97.53% |
| VP | VNP | 168 | 1.15% | 98.68% |
| S | VNP | 129 | 0.88% | 99.56% |
| VNP | VNP | 46 | 0.31% | 99.87% |
| L,X,AP,IP,LP | VNP | 19 | 0.13% | 100.00% |
| 소계 | | 14,653 | | |

위 NP의 99.9%가 주어 표지가 부착된 범주였다. 즉 앞의 VP와 관련한 일반화를 확대하여 말하자면 S 삼각구조에서 RDN이 VP나 VNP라면 LDN이 NP_SBJ일 확률은 97%를 넘는다고 볼 수 있다. 이는 물론 일반적으로 기대되는 바라 할 수 있겠으나, 대규모 말뭉치를 통해 구체적인 수치로 입증되고 있다는 점은 의의가 있다.

### 3.3 VP 규칙

세종 구문분석 말뭉치에 나오는 VP 규칙의 빈도는 277,624로 나타났다. 그 규칙의 유형은 59개로 토큰/타입 비율은 4,705이나, 대부분이 상위 소수의 규칙에 높은 빈도로 나타난다. 즉, 빈도수 기준 상위 4개의 VP규칙은 아래에 표에서 보듯 합계 95% 이상으로 활용되고 있다. 나머지 55개의 규칙은 빈도로는 4,069개~1개, 비율로 보아 1.47%~0.00%정도로만 분포되어 있다.

[표 18] 상위 4개 VP 규칙 분포

| MN | LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|---|
| VP | NP | VP | 141,806 | 51.08% | 51.08% |
| VP | VP | VP | 80,431 | 28.97% | 80.05% |
| VP | AP | VP | 33,074 | 11.91% | 91.96% |
| VP | S | VP | 10,072 | 3.63% | 95.59% |
| VP | L | VP | 2,071 | 0.75% | 96.34% |
| VP | VP | R | 2,062 | 0.74% | 97.08% |
| VP | NP | NP | 1,222 | 0.44% | 97.52% |
| 소계 | | | 270,738 | | |

고빈도 상위 4개의 VP 규칙을 보면, 세종 구문분석의 이분지 구조의 기본 원칙이 잘 드러난다. 대체로 RDN의 구문 표지는 MN의 구문 표지가 된다. 즉, VP를 머리어로 갖는 구 역시 VP로 분석됨을 확인할 수 있다. 빈도를 보면, 일반적으로 용언구가 NP논항을 취하는 규칙이 51.08%로 가장 많이 나타났고, 용언구와 용언구가 합쳐져서 상위절의 용언구로 확장되는 규칙이 28.97%로 나타났다. 이러한 구조로는 본용언 + 본용언 구성, 본용언 + 보조용언 구성 등이 해당한다. 그 다음으로는 부사구가 용언구를 수식해주는 규칙이 11.91%, 하나의 문장이 용언구의 논항으로 나타난 규칙이 3.63%로 나타났다.

한편 VP 규칙의 범주 VP가 쓰인 기능별 분포는 아래와 같다.

[표 19] 범주 VP의 기능별 분포

| 기능 | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| VP | 174,561 | 62.88% | 62.88% |
| VP_MOD | 89,761 | 32.33% | 95.21% |
| VP_CMP | 5,983 | 2.16% | 97.36% |
| VP_OBJ | 3,147 | 1.13% | 98.50% |
| VP_AJT | 2,697 | 0.97% | 99.47% |
| VP_SBJ | 1,242 | 0.45% | 99.92% |
| VP_PRN | 165 | 0.06% | 99.98% |
| VP_CNJ | 65 | 0.02% | 100.00% |
| VP_INT | 3 | 0.00% | 100.00% |
| 소계 | 277,624 | | |

위 표를 보면 VP기능에 따른 빈도의 순서가 S 규칙의 기능 분포와 동일하다. 기능이 명시되지 않은 경우가 가장 많은데, 세종 구문분석 말뭉치에서는 전통적인 통사구조와는 달리 최상위 절점이 S뿐 아니라, VP도 가능하다. 즉, 주어가 없으면서 머리어가 용언인 모든 절은 VP로 분석된다. 또한 용언구 + 용언구 구성의 경우 역시 기능이 부착되지 않는다. 그 다음으로 VP가 관계절을 포함하는 체언수식절(VP_MOD)로는 32%이상 활용된 반면, 같은 수식기능을 하되 용언수식절(VP_AJT)로 쓰인 경우는 0.97%로 미미한 수치였다. 절이 논항으로 쓰인 경우(VP_CMP, VP_OBJ, VP_SBJ)는 모두 합하면 3.5% 가까이 된다. 또한 VP가 PRN, CNJ, INT 기능으로 쓰인 경우는 드물게 나타났다. PRN는 주로 명사구 삽입구문으로, CNJ는 체언 접속 구성의 표지로, INT는 역시 명사구 독립어 표지로 주로 사용된다.

### 3.3.1 VP 규칙 좌측 절점 (LDN) 기준

LDN 범주를 기준으로 VP 규칙의 적용 빈도를 보면 아래와 같다.

[표 20] LDN 기준 VP 규칙 분포

| LDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 143,341 | 51.63% | 51.63% |
| VP | 83,460 | 30.06% | 81.69% |
| AP | 33,154 | 11.94% | 93.64% |
| S | 10,192 | 3.67% | 97.31% |
| VNP | 4,090 | 1.47% | 98.78% |
| L | 2,089 | 0.75% | 99.53% |
| LP,DP,Q,X,R,INT,U | 1,298 | 0.47% | 100.00% |
| 소계 | 277,624 | 100.00% | |

위 표를 보면, 이는 S규칙과 유사하게 NP 논항이 51.63%로 절반 정도로 가장 많이 차지하고, 그 다음 순으로 자신의 머리어 VP를 취

하고 있다. AP, S 까지 합하면 97%정도를 차지하고, 나머지 범주들은 미미한 분포를 보인다. 이에 VP 규칙의 LDN 범주를 기준으로 각 범주별로 그것의 RDN이 어떤 것들이 어떤 비율로 나오는지 살펴보기로 하자. 위의 표의 빈도순으로 NP, VP, AP, S의 순으로 보면 다음과 같다.

[표 21] LDN이 NP일 경우 VP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | VP | 141,806 | 98.93% | 98.93% |
| NP | NP | 1,222 | 0.85% | 99.78% |
| NP | S,VNP,AP,R,X,L,IP | 313 | 0.22% | 100.00% |
| NP 소계 | | 143,341 | | |

[표 22] LDN이 VP일 경우 VP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| VP | VP | 80,431 | 96.37% | 96.37% |
| VP | S,NP,VNP,AP,IP,X,Q,R | 3,029 | 3.63% | 100.00% |
| VP 소계 | | 83,460 | | |

[표 23] LDN이 AP일 경우 VP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| AP | VP | 33,074 | 99.76% | 99.76% |
| AP | NP,VNP,S,IP,R | 80 | 0.24% | 100.00% |
| AP 소계 | | 33,154 | 100.00% | |

[표 24] LDN이 S일 경우 VP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| S | VP | 10,072 | 98.82% | 98.82% |
| S | S,NP,VNP,DP,R,Q,X | 120 | 1.18% | 100.00% |
| S 소계 | | 10,192 | 100.00% | |

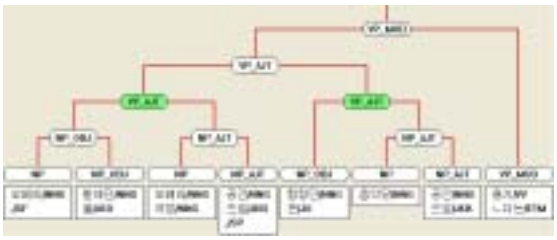
위 표에 따르면 세종 구문분석 말뭉치에서 MN이 VP이면 LDN에

관계없이 RDN은 대부분 VP가 될 확률이 거의 100%에 이른다.

첫 번째 표에서 보면, LDN과 RDN이 둘 다 NP일 때가 1,222회 나타났는데, 특수한 형태의 구문임에도 적지 않은 수가 발견된다는 점이 흥미롭다. 따라서 그 구문을 좀 더 살펴보기로 한다. [한마루] 삼각구조 검색을 통해 검색해보면 다음과 같은 예문들이 추출된다.

(8) a. [VP[NP[NP원한과 사랑의] [NP갈등구조를 중심축으로]]] 이야기를 전개하는 이 연극은 이런 불화와 경원, 애증이 허망한 것임을 보여준다.

b. 실제로 문화부쪽은 3개 단체 이전안이 장관결재를 받을 때까지 예술의 전당 관계자들과 공식협의를 거의 거치지 않았고 지난달 중순 김명곤 극장장 내정 뒤에야 [VP[NP[NP오페라. 발레단을] [NP 오페라극장 공간으로, 합창단은 음악당 공간으로 옮긴다는]]] 방침을 통보했다.



위의 예문들은 NP와 NP가 합해져서 VP를 이루는 구조로, 기존의 X′-theory와는 다르게 분석되고 있다. 우선 '오페라 발레단을'은 NP이고 목적어 기능을 한다. '오페라극장 공간으로' 역시 NP이나 서술어 기능을[30] 하는 것으로 '오페라 발레단을 오페라 극장 공간으로' 전체는 VP가 된다. 'VP → NP NP'의 구조를 지니는 예문을 살펴본 결과 'VP_AJT → NP_OBJ NP_AJT'을 지닌 예문은 464회 출현하였다. 나머지 세 개의 표는 RDN이 VP인 경우를 제외하면 그 다음으로 R이나 X와 같은 부호가 나타났다.

### 3.3.2 VP 규칙 우측 절점 (RDN) 기준

VP 규칙의 RDN 범주를 기준으로 VP 규칙의 적용 빈도를 보면 VP 하나가 98%이상이라는 절대 다수의 분포를 보이고 있다. 또한 그 다음으로 오는 범주들의 순서 역시 차이를 보인다. LDN은 NP, VP, AP, S 순이었다면, RDN은 VP, R, NP, X 등의 순으로 나타났다. VP 규칙의 RDN 범주를 기준으로 나타낸 표는 다음과 같다.

[표 25] RDN 기준 VP 규칙 분포

| RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| VP | 272,812 | 98.27% | 98.27% |
| R | 2,364 | 0.85% | 99.12% |
| NP | 1,482 | 0.53% | 99.65% |
| X | 643 | 0.23% | 99.88% |
| S | 122 | 0.04% | 99.92% |
| VNP | 108 | 0.04% | 99.96% |
| IP,AP,Q,DP,L | 93 | 0.03% | 100.00% |
| 소계 | 277,624 | 100.00% | |

30) 서술성 명사와 유사한 기능을 하는 것이다. 예를 들어, [백화점들과 농협, 일선 행정기관은 이번 주부터 상설코너를 마련, 국기를 팔고 있다.]라는 문장에서 '마련'과 같은 서술어 기능을 말한다.

즉, VP 규칙의 RDN 범주 중 거의 대부분이 VP라는 점은 앞에서 언급한 바와 마찬가지로 용언구의 머리어는 용언(구)라는 일반적인 기대에 부합한다. 또한 LDN과 달리 NP가 거의 나타나지 않는 것은 한국어 용언의 논항은 왼쪽에 오기 때문이다. 용언구 오른쪽에 NP가 오는 예문을 [한마루]에서 검색하면 다음과 같은 구문이 추출된다.[31)]

(9) a. [VP[VP잘 알지요,] [NP민섭씨.]]
b. [VP[VP로마 제국이 위태로워지자 서울을 동쪽 콘스탄티노플로 옮겼단 얘기는 했지?] [NP그리고 476년, 게르만 족에 의해 서로마 제국이 멸망했다는 것도.]]
c. 어느 한 올림픽 구기종목에서 우리 한국팀이 4 강에 오르면 3천만 원씩, [VP[VP결승에 오르면] [NP4천만 원씩]], [VP[VP우승을 하면] [NP5천만원씩] 포상을 한다는 보도에 접하고 보니 이 나귀의 당근이야기가 생각나는 것이다.

주로 (9a-b)의 예문은 어순이 뒤바뀐 경우이다. (9c)의 예문의 경우는 '4천만 원씩'은 NP이고 앞에 '3천만 원씩'과 함께 뒤에 나오는 VP의 목적어 기능을 하고, '포상한다'는 동사가 생략되어 있다고 볼 때, '결승에 오르면 4천만 원씩,'이라는 구는 VP가 된다. 이러한 구조는 공범주를 표상해 준 PKT와는 달리 SKT는 표면 구조에 충실한 특성이 잘 반영되어 있기 때문이다.

다음은 RDN 범주를 기준으로 그것의 SN, 즉 관련 VP 규칙의 LDN이 어떤 분포를 보이는지 살펴본 것이다. RDN이 VP일 때의 분포 외에 다른 범주는 빈도가 미미한 수준으로 제외하고 VP일 경우만 보면 아래와 같다.

31) 검색식 '용언구 → 용언구 체언구_목적어'로 검색한 결과 9개의 문장이 추출되었다.

[표 26] RDN이 VP일 경우 VP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | VP | 141,806 | 51.98% | 51.98% |
| VP | VP | 80,431 | 29.48% | 81.46% |
| AP | VP | 33,074 | 12.12% | 93.58% |
| S | VP | 10,072 | 3.69% | 97.28% |
| VNP | VP | 4,069 | 1.49% | 98.77% |
| L,IP,DP,Q,X,LP,R,INT,U | VP | 3,360 | 1.12% | 100.00% |
| 소계 | | 272,812 | | |

위의 표에 따르면 NP가 51.98% 절반 이상의 분포로 나타났고, 그 다음은 VP, AP, S 순으로 나타났다. 이는 3.2절에 제시하였던 VP 전체 규칙의 고빈도 상위 4개의 범주 표의 순서 및 분포와 거의 일치한다. 즉, MN과 RDN이 둘 다 VP인 경우 그 둘은 서로 밀접한 관계를 맺고 있음을 알 수 있다.

### 3.4 NP 규칙

NP 규칙은 SKT에서 총 282,292회 출현하는 데, 그 타입은 995개에 지나지 않는다. 다시 1회 출현하는 것을 제외하면[32] 토큰은 총 281,759, 타입은 총 498, 토큰/타입 비율은 565으로 산정된다. 출현빈도 1인 것 까지 포함하여 규칙의 분포를 살피면 빈도 출현 상위 9위까지의 범주 규칙이 전체의 99%에 이르는 것으로 나타났다. 또한 'NP → NP NP', 'NP → VP NP' 의 두 규칙이 전체의 75%를 상회하는 높은 비율을 보임을 알 수 있었다. 이는 한국어에서 NP의 대부분이 복합명사구 또는 '용언수식어구 + 명사구'의 형태로 사용된다는 점을 방증한다 하겠다.

32) 규칙 가운데서 단 1회 출현하는 것은 재고의 여지가 있다. 예컨대, 1회 출현하는 'NP_AJT → NP_AJT VP'와 같은 규칙은 머리어가 후치하는 한국어의 특성상 상당히 특이한 경우에 속한다.

[표 27] 상위 9개 NP 규칙 분포

| MN | LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|---|
| NP | NP | NP | 146,201 | 51.79% | 51.79% |
| NP | VP | NP | 68,903 | 24.41% | 76.20% |
| NP | DP | NP | 21,895 | 7.76% | 83.96% |
| NP | S | NP | 17,571 | 6.22% | 90.18% |
| NP | VNP | NP | 7,199 | 2.55% | 92.73% |
| NP | NP | X | 5,242 | 1.86% | 94.59% |
| NP | L | NP | 5,002 | 1.77% | 96.36% |
| NP | NP | R | 4,972 | 1.76% | 98.12% |
| NP | AP | NP | 3,341 | 1.18% | 99.30% |
| 소계 | | | 280,326 | | |

한편 NP 범주의 기능별 분포는 아래와 같다. 'NP_SBJ', 'NP_AJT', 'NP', 'NP_OBJ', 'NP_MOD', 'NP_CNJ' , 'NP_CMP' 등의 7개의 기능별 NP 범주가 전체의 98%이상을 점유하고 있으며, 그 가운데서 특히 'NP_SBJ', 'NP_AJT', 'NP_OBJ'의 세 유형이 전체의 3분의 2에 해당함을 알 수 있다. 이는 명사구의 대부분이 주어, 목적어, 용언수식어로서 기능을 하고 있다는 점을 말한다.

[표 28] 범주 NP의 기능별 분포

| 기능 | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP_SBJ | 70,212 | 24.87% | 24.87% |
| NP_AJT | 66,721 | 23.64% | 48.51% |
| NP | 54,726 | 19.39% | 67.89% |
| NP_OBJ | 51,452 | 18.23% | 86.12% |
| NP_MOD | 20,801 | 7.37% | 93.49% |
| NP_CNJ | 10,043 | 3.56% | 97.05% |
| NP_CMP | 4,922 | 1.74% | 98.79% |
| NP_PRN | 2,699 | 0.96% | 99.75% |
| NP_INT | 690 | 0.24% | 99.99% |
| 소계 | 282,266 | | |

### 3.4.1 NP 규칙 좌측 절점 (LDN) 기준

LDN을 기준으로 NP 규칙의 분포를 살펴보면 아래와 같다.

[표 29] LDN 기준 NP 규칙 분포

| LDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 157,090 | 55.65% | 55.65% |
| VP | 69,013 | 24.45% | 80.10% |
| DP | 21,904 | 7.76% | 87.85% |
| S | 17,613 | 6.24% | 94.09% |
| VNP | 7,205 | 2.55% | 96.65% |
| L | 5,028 | 1.78% | 98.43% |
| AP | 3,351 | 1.19% | 99.61% |
| X, IP, Q, R, LP | 1,088 | 0.39% | 100.00% |
| 소계 | 282,292 | | |

위 표를 살펴보면, LDN이 NP인 경우, 또는 VP인 경우가 전체의 80%를 상회하고 있음을 알 수 있다. 이 두 범주를 중심으로 하여 각각의 RDN에 어떠한 범주가 자리하는지 살펴보도록 하자. 위 VP 규칙에서 파악한 것과 같이, SKT의 특성상 RDN에는 MN과 같은 범주가 오는 것이 일반적이다. 먼저 LDN이 NP인 경우를 검토해 보도록 하자.

[표 30] LDN이 NP일 경우 NP 규칙 분포

| RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 146,201 | 93.07% | 93.07% |
| X | 5,242 | 3.34% | 96.41% |
| R | 4,972 | 3.17% | 99.57% |
| VP | 273 | 0.17% | 99.74% |
| AP | 182 | 0.12% | 99.86% |
| S | 126 | 0.08% | 99.94% |
| VNP, L, IP, Q | 94 | 0.06% | 100.00% |
| 소계 | 157,090 | | |

예상한 바와 같이 RDN에 NP가 출현하는 것이 전체의 93%를 상

회하고 있다. 문제는 상대적으로 낮지만 유의미한 수치를 보이는 VP, AP, S 등의 범주이다. 우선, 'NP → NP VP'의 범주의 예로는 [NP[NP청주공단 영태전자][VP(다시 뛴다)]], [68년 NP[NP[12 월 성경공부중] [VP영혼의 지진을 경험.]]] 등이 있다. 다음으로 'NP → NP AP'의 예로는 [NP_CNJ[NP'양'] [AP또는]], [NP_CNJ [NP_CNJ그의 지지자들], [AP그리고]] 등이 [한마루] 검색결과로 발견되었다. 즉, NP 뒤에 곧바로 '또는', '그리고', '및'과 같은 접속 부사가 자리하는 경우가 이에 해당한다. 끝으로 'NP → NP S'의 경우는 [NP[NP'明日來他盜] [S(내일이면 또 다른 도둑이 오려니)]]와 같은 특수한 예가 발견되는 바, 이는 NP 다음에 괄호로 그 NP에 대한 상세 정보를 주는 문장이 나오는 경우이다.

이상으로 보아, 적은 빈도를 보이는 사례들은 흔히 말하는 특수 구문에 해당하는 것으로 이 역시 SKT가 한국어의 특성을 세세한 부분까지 잘 반영하고 있음을 보여준다.

다음으로 MN이 NP일때, LDN이 VP인 경우를 보도록 하자.

[표 31] LDN이 VP일 경우 NP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| VP | NP | 68,903 | 99.85% | 99.85% |
| VP | R | 53 | 0.08% | 99.93% |
| VP | VP | 31 | 0.04% | 99.97% |
| VP | S | 13 | 0.02% | 99.99% |
| VP | VNP, X, AP | 7 | 0.01% | 100.00% |
| 소계 | | 69,007 | | |

기대하는 바와 같이 NP가 전체의 99.85%를 차지한다. 기타의 경우에는 모두 0.1% 미만으로 극히 드문 출현을 보인다.

### 3.4.2 NP 규칙 우측 절점 (RDN) 기준

다음으로 RDN을 기준으로 한 분포에 대해서 살펴보기로 하자.

[표 32] RDN 기준 NP 규칙 분포

| RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 271,179 | 96.06% | 96.06% |
| X | 5,264 | 1.86% | 97.93% |
| R | 5,064 | 1.79% | 99.72% |
| VP | 332 | 0.12% | 99.84% |
| AP | 183 | 0.06% | 99.90% |
| S | 157 | 0.06% | 99.96% |
| VNP, L, IP, Q | 113 | 0.04% | 100.00% |
| 소계 | 282,292 | | |

당초의 예상과 같이 RDN이 NP인 경우가 대부분을 차지하고 있다. 다시 RDN이 NP인 경우에 LDN의 분포를 살피면 다음과 같다.

[표 33] RDN이 NP일 경우 NP 규칙 분포

| LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP | NP | 146,201 | 53.91% | 53.91% |
| VP | NP | 68,903 | 25.41% | 79.32% |
| DP | NP | 21,895 | 8.07% | 87.40% |
| S | NP | 17,571 | 6.48% | 93.88% |
| VNP | NP | 7,199 | 2.65% | 96.53% |
| L | NP | 5,002 | 1.84% | 98.37% |
| AP | NP | 3,341 | 1.23% | 99.61% |
| X, IP, Q, R, LP | NP | 1,067 | 0.39% | 100.00% |
| 소계 | | 271,179 | | |

이 결과의 분포는 비교적 다양하다. 우선 LDN이 NP인 경우, 다시 말해 복합명사구의 경우가 전체의 절반가량을 차지한다. 그 밖에 용언이 명사구에 선행하여 수식하는 구조('LDN-VP')나 관형사구('LDN-DP'), 또는 문장이 명사구를 수식하는 구조('LDN-S')등이

차례로 빈도 순위를 차지한다.

### 3.5 기타 규칙

세종 구문분석 말뭉치에 나오는 규칙 중 S, VP, NP를 제외한 규칙은 51,709회 적용된 것으로 파악된다. 기타 규칙의 전체 유형은 169개로 토큰/타입 비율은 305이다. 기타 규칙의 범주별 규칙 유형과 빈도 비율은 다음과 같다.

[표 34] 기타 규칙의 범주별 규칙 분포

| MN | 빈도 | 비율 | 누적 비율 |
|---|---|---|---|
| S,VP,NP | 722,289 | 93.32% | 93.32% |
| VNP | 48,040 | 6.21% | 99.53% |
| AP | 2,681 | 0.35% | 99.87% |
| IP | 513 | 0.07% | 99.94% |
| DP | 187 | 0.02% | 99.96% |
| Q | 97 | 0.01% | 99.98% |
| X | 62 | 0.01% | 99.98% |
| LP | 61 | 0.01% | 99.99% |
| L | 34 | 0.00% | 100.00% |
| R | 32 | 0.00% | 100.00% |
| U | 2 | 0.00% | 100.00% |
| 소계 | 773,998 | | |

위의 표에 따르면 빈도, 비율은 S, VP, NP 규칙을 제외한 나머지 규칙 중 거의 대부분이 긍정지정사 VCP (‘이다’ 결합용언)가 부착된 구를 다루는 VNP 규칙으로 전체의 6.21%를 차지하고 있다. 그 다음으로는 AP가 0.35%로 나타나는데, 일반부사 MAG와 접속부사 MAJ로 부착된 어절이다. 그 나머지는 미미한 비율로 인용절이나 부호를 나타내는 구문이다. 이중 빈도수로 보아 상위 2개인 VNP, AP에 대해 더 상세히 살펴보도록 하겠다.

### 3.5.1 VNP 규칙

기타 규칙 중에서 가장 높은 빈도를 보인 VNP 규칙의 분포는 다음과 같다.

[표 35] 상위 6개 VNP 규칙 분포

| MN | LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|---|
| VNP | VP | VNP | 17,571 | 36.58% | 36.58% |
| VNP | NP | VNP | 15,778 | 32.84% | 69.42% |
| VNP | S | VNP | 5,122 | 10.66% | 80.08% |
| VNP | AP | VNP | 4,577 | 9.53% | 89.61% |
| VNP | VNP | VNP | 2,516 | 5.24% | 94.85% |
| VNP | DP | VNP | 1,082 | 2.25% | 97.10% |
| 소계 | | | 46,646 | | |

위 표를 보면 우선 MN이 VNP인 경우 RDN의 98% 가까이가 VNP라는 점이 눈에 띈다. 그리고, VP에 VNP가 결합하는 구성이 36.58%, NP에 VNP가 결합하는 구성이 32.84%로 비슷한 분포로 나타났고, 이 둘의 합이 69.42%로 절반 이상을 차지하였다. 이 두 구성의 차이를 다음의 예를 통해 살펴보도록 하겠다.[33]

(10) a. [VP[NP" 과학에 무지한 독자에게 깊은 인상을 남기고 무엇보다 겁을 주려는 것 "] [VNP이다.]]
b. [VP[NP'내가 하면 로맨스, 남이 하면 스캔들'이라는] [VNP말이]] 괜히 나왔겠는가!
c. [VP[VP다음 호부터 연재될 프로그램들은 모두 '선생님들의 작은 공간' 을 통해 제공할] [VNP 계획이다.]]

33) [한마루] 검색식 '용언구 → 체언구 긍정지정사구'로 검색한 결과 47개의 문장이 추출되고, '용언구 → 용언구 긍정지정사구'로 검색한 결과 22개의 문장이 추출되었다.

(10b)의 예는 체언구와 긍정지정사구가 결합하여 긍정지정사구를 이루는 구조로 '-것이다' '-라는 -이다' 등의 예가 나타나고, (10c)의 예는 용언구와 긍정지정사구가 결합하는 구조로 '-할 -이다'

그 다음으로는 VNP의 논항이 문장인 경우가 10.66%, 부사구인 경우 9.53%로 서로 비슷한 비율로 나타났고, 이어서 VNP, DP 순으로 나타났다.

LDN를 기준으로 VNP 규칙의 적용 빈도를 보면 아래와 같다.

[표 36] LDN 기준 VNP 규칙 분포

| LDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| VP | 17,620 | 36.68% | 36.68% |
| NP | 15,846 | 32.99% | 69.67% |
| S | 5,140 | 10.70% | 80.37% |
| AP | 4,587 | 9.55% | 89.92% |
| VNP | 3,068 | 6.39% | 96.31% |
| DP | 1,086 | 2.26% | 98.57% |
| L,IP,X,Q,LP,R | 693 | 1.34% | 100.00% |
| 소계 | 48,040 | | |

위 표를 보면, 앞에서 살펴본 표와 유사한 순서와 빈도 비율로 나타났다. 즉, MN이 VNP인 경우 RDN의 대부분 VNP가 되므로, LDN 범주 분포는 RDN의 영향을 별로 받지 않는다는 것을 알 수 있다. 이는 다음의 표를 통해서도 확인이 가능하다. 아래에 제시되는 표는 VNP규칙의 RDN 범주를 기준으로 VNP 규칙의 적용 빈도를 나타낸 것이다.

[표 37] RDN 기준 VNP 규칙 분포

| RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| VNP | 47,335 | 98.52% | 98.52% |
| R,X,NP,S,VP,IP,AP,Q | 705 | 1.49% | 100.00% |
| 소계 | 48,040 | | |

표에 따르면 VNP의 RDN 범주 중 VNP가 98.52%라는 거의 절대적인 분포를 차지하고 있다. 이는 VNP는 긍정지정사 '이다'로 구성된 구로 VP와 마찬가지로 머리구가 되는 역할을 하기 때문에 MN은 RDN의 범주와 거의 일치하고 LDN의 범주 분포도 유사하게 나타남을 의미한다.

### 3.5.2 AP 규칙

다음은 AP 규칙을 빈도순으로 보인 표이다.

[표 38] 상위 6개 AP 규칙 분포

| MN | LDN | RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|---|
| AP | NP | AP | 1,481 | 55.24% | 55.24% |
| AP | AP | AP | 771 | 28.76% | 84.00% |
| AP | VP | AP | 100 | 3.73% | 87.73% |
| AP | L | AP | 71 | 2.65% | 90.38% |
| AP | AP | R | 68 | 2.54% | 92.92% |
| AP | AP | X | 58 | 2.16% | 95.08% |
| 소계 | | | 2,549 | | |

AP 규칙은 NP, AP로 이루어진 구조가 55.24%로 절반 정도의 비율로 나타나고, 다음으로 NP, NP 구성으로 이루어진 것이 28.76%로 이 둘을 합하면 84%의 분포를 보인다.

다음은 AP 규칙의 LDN을 기준으로 나타낸 표이다.

[표 39] LDN 기준 AP 규칙 분포

| LDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP | 1,494 | 55.73% | 55.73% |
| AP | 914 | 34.09% | 89.82% |
| VP | 114 | 4.25% | 94.07% |
| L | 76 | 2.83% | 96.90% |
| IP | 25 | 0.93% | 97.84% |
| S | 25 | 0.93% | 98.77% |
| DP | 23 | 0.86% | 99.63% |
| VNP | 10 | 0.37% | 100.00% |
| 소계 | 2,681 | | |

위의 분포표를 보면 AP 규칙의 LDN에 나오는 범주 중 NP가 55.73%로 절반 정도 차지하고 있고, 다음은 AP가 34.09%로 이 둘을 합하면 90%에 가까운 비율을 보였다.

AP 규칙 우측 절점 기준으로 살펴보면, 지금까지 살펴본 다른 범주들과 마찬 가지로 AP 역시 자신의 머리어가 RDN의 거의 대부분을 차지하는 것으로 나타났다.

[표 40] RDN 기준 AP 규칙 분포

| RDN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| AP | 2,501 | 93.29% | 93.29% |
| IP,NP,R,S,VNP,VP,X | 180 | 6.71% | 100.00% |
| 소계 | 2,681 | | |

### 3.6 딸 절점 (DNs) 기준

이상에서 살핀 바는 모두 MN을 기준으로 한 것이다. 여기에서는 이 과정을 역으로 뒤집어 딸 절점들(Daughter Nodes, 이하 'DNs') 에 따라 MN가 어떠한 분포를 보이는지를 살펴보기로 한다.

먼저 DNs의 분포는 아래의 표와 같다.

[표 41] DNs 분포

| DNs | 빈도 | 비율 | 누적비율 |
|---|---|---|---|
| NP-VP | 214,661 | 27.73% | 27.73% |
| NP-NP | 149,427 | 19.31% | 47.04% |
| VP-VP | 81,337 | 10.51% | 57.55% |
| VP-NP | 69,163 | 8.94% | 66.48% |
| AP-VP | 33,110 | 4.28% | 70.76% |
| NP-VNP | 30,199 | 3.90% | 74.66% |
| NP-S | 28,725 | 3.71% | 78.37% |
| DP-NP | 21,902 | 2.83% | 81.20% |
| VP-VNP | 17,787 | 2.30% | 83.50% |
| S-NP | 17,754 | 2.29% | 85.80% |
| VP-S | 12,027 | 1.55% | 87.35% |
| AP-S | 11,522 | 1.49% | 88.84% |
| S-S | 10,789 | 1.39% | 90.23% |
| 소계 | 774,005 | | |

NP와 VP가 결합하는 규칙이 가장 많고, 그 다음은 LDN과 RDN이 모두 NP인 규칙이다. 3위는 VP끼리 결합하는 형태이며, 4위는 VP가 NP를 자신의 우측에 취하는 규칙이다. 5위는 AP가 VP를 수식하는 형태에 해당한다. 여기까지의 규칙이 전체의 70%를 차지한다. 이 가운데서 상위 5개의 MN 분포가 어떠한지를 다시 살피기로 하자.

우선 DNs가 'NP-VP'인 경우 MN의 분포는 아래 표와 같다.

[표 42] DNs가 'NP-VP'인 경우 MN의 분포

| DNs | MN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP-VP | VP | 141,806 | 66.06% | 66.06% |
| NP-VP | S | 72,566 | 33.80% | 99.87% |
| NP-VP | NP, VNP, AP, Q, L, X | 289 | 0.13% | 99.99% |
| 소계 | | 214,661 | | |

NP와 VP가 결합하여 다시 VP를 이루는 경우가 전체의 3분의 2 이상에 해당하여 가장 많다. 이는 VP가 앞 NP를 자신의 보어로 취하는 경우하고 추정할 수 있다. 다음으로 NP와 VP의 결합이 S가 되

는 규칙이 오는데 이는 주어부와 술어부의 결합으로 판단된다. 그 이하의 규칙은 1% 미만의 낮은 분포를 보인다.

다음은 DNs가 'NP-NP'인 경우이다.

[표 43] DNs가 'NP-NP'인 경우 MN의 분포

| DNs | MN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| NP-NP | NP | 146,201 | 97.84% | 97.84% |
| NP-NP | S | 1,960 | 1.31% | 99.15% |
| NP-NP | VP | 1,222 | 0.82% | 99.97% |
| NP-NP | VNP, L, AP, R, X | 44 | 0.01% | 99.98% |
| 소계 | | 149,427 | | |

NP와 NP의 결합은 NP가 될 확률이 압도적임을 알 수 있다. 그 다음 NP-NP가 결합하여 S나 VP가 되는 경우는 특수한 구문들로, 각각 3.2절과 3.3절에서 논의한 바 있다.

세 번째로 DNs가 'VP-VP'인 경우이다. 예상과 일치하여 VP끼리의 결합은 대부분의 경우에 다시 VP가 된다는 사실을 확인하였다.

[표 44] DNs가 'VP-VP'인 경우 MN의 분포

| DNs | MN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| VP-VP | VP | 80,431 | 98.89% | 98.89% |
| VP-VP | S | 872 | 1.07% | 99.96% |
| VP-VP | NP, AP, VNP | 34 | 0.04% | 100.00% |
| 소계 | | 81,337 | | |

네 번째로는 DNs가 'VP-NP'인 경우이다. 대부분의 경우 MN 역시 NP인 것으로 보아 이는 관계절 구문으로 예상된다.

[표 45] DNs가 'VP-NP'인 경우 MN의 분포

| DNs | MN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| VP-NP | NP | 68,903 | 99.62% | 99.62% |
| VP-NP | VP | 198 | 0.29% | 99.91% |
| VP-NP | S | 51 | 0.07% | 99.98% |
| VP-NP | L, AP, VNP, R, X | 11 | 0.01% | 99.99% |
| 소계 | | 69,163 | | |

끝으로 DNs가 'AP-VP'인 경우를 살피면 아래와 같다.

[표 46] DNs가 'AP-VP'인 경우 MN의 분포

| DNs | MN | 빈도 | 비율 | 누적비율 |
|---|---|---|---|---|
| AP-VP | VP | 33,074 | 99.89% | 99.89% |
| AP-VP | S | 24 | 0.07% | 99.96% |
| AP-VP | AP, NP, DP | 12 | 0.02% | 99.99% |
| 소계 | | 33,110 | | |

이상의 다섯 가지 경우를 종합하면 DNs가 주어지면 이에 따라 MN이 거의 확정적으로 결정된다는 점을 알 수 있다. 첫 번째의 'NP-VP' 경우에는 VP 또는 S로 분기되는 측면이 있으나, S도 넓은 범주에서 VP의 확장이라는 점을 고려하면 이 역시 일반화의 대상이 된다. 다시 말해, 한국어에서 MN은 RDN에 99% 의존하고 있음을 알 수 있다.

## 4. 결론

지금까지 본 연구에서는 SKT로부터 모든 CFG 규칙 및 규칙별 빈도수를 추출한 뒤에, 그 규칙들을 S, VP, NP, VNP, AP 등 각 범주별로 나누어 검토하였다. MN, LDN, RDN으로 구성된 삼각구조에서 각각 MN, MN-LDN, MN-RDN, LDN-RDN을 기준으로 한 범주별 분포적 특성 및 기능과의 상관관계 등을 다음의 기준을 중심으

로 살펴보았다: 1) MN 범주별 고빈도 상위 규칙 2) MN 범주가 쓰인 기능별 분포 3) MN-LDN 기준 규칙 빈도 4) MN-RDN 기준 규칙 빈도 5) LDN-RDN 기준 규칙 빈도. 이처럼 보다 체계적이고 구체적인 CFG 규칙 분석을 통하여 한국어 CFG 규칙에 대한 다음 몇 가지 특징을 포착할 수 있었다.

(11) a. S, VP, NP, VNP, AP 규칙 모두 상위 소수 규칙이 많은 빈도 비율을 차지하였다.[34]

b. MN의 기능별 분포는 S, VP, VNP가 동일하게 나타났는데, 기능이 명시되지 않은 경우가 대부분이고, 체언수식 기능을 하는 것이 그 다음을 차지하였다 ({S,VP,VNP} > _MOD > _CMP > _OBJ > _AJT > _SBJ > _PRN > _CNJ > _INT). NP 와 AP는 다른 양상을 보였는데, NP의 경우 주어기능, 수식어 기능 순으로 나타났고 (NP_SBJ > NP_AJT > NP > NP_OBJ > NP_MOD > NP_CNJ > NP_CMP), AP의 경우는 다른 범주와 달리, PRN, CNJ, INT 기능이 나타나지 않았다 (AP > AP_MOD> AP_AJT > AP_OBJ > AP_CMP > AP_SBJ).

c. 각 범주별 규칙의 LDN을 기준으로 SN의 분포를 살펴본 결과, VNP를 제외한 나머지 S, VP, NP, AP 모두 NP가 절반 이상의 비율로 가장 많이 나타났다.

d. 각 범주별 규칙의 RDN을 기준으로 SN의 분포를 살펴본 결과, S, VP, NP, VNP, AP 모두 RDN이 MN과 같은 범주가 오는 경우가 95% 이상이었다. 즉, 이를 통해 세종 구문 분석의 이분지 구조와 관련하여 RDN이 MN과 밀접한 상관관계가 있음을 확인할 수 있다.

---

34) S: 총 63개 규칙 중 빈도수 기준 상위 6개가 92% 이상 출현하였다.
VP: 총 59개 규칙 중 빈도수 기준 상위 7개가 97% 이상 출현하였다.
NP: 총 995개 규칙 중 빈도수 기준 상위 9개가 99% 이상 출현하였다.
VNP: 총 43개 규칙 중 빈도수 기준 상위 6개가 97% 이상 출현하였다.
AP: 총 31개 규칙 중 빈도수 기준 상위 6개가 95% 이상 출현하였다.

이러한 특징뿐만 아니라, 각 MN의 범주별 CFG 규칙에 대한 부분적인 일반화가 가능하다.

(12) a. MN이 S이고 LDN이 AP이면 RDN이 S가 될 확률은 거의 100%에 이른다.
b. MN이 S이고 RDN이 VP이면 LDN이 NP가 될 확률은 98%에 이른다.
c. MN이 S이고 RDN이 VP나 VNP이면 LDN이 NP_SBJ일 확률은 97%가 넘는다.
d. MN이 VP이면 LDN에 관계없이 RDN이 VP 가 될 확률은 거의 100%에 이른다.
e. MN과 RDN이 둘 다 VP인 경우 그 둘은 서로 밀접한 관계를 맺는다.
f. MN이 NP이고 LDN이 NP인 경우, 또는 VP인 경우가 전체의 80%를 상회한다.
g. MN이 NP이고 LDN이 VP이면 RDN이 NP가 될 확률은 거의 100%에 이른다.
h. MN이 VNP이면 RDN이 VNP일 확률이 거의 100%이고, 이 경우 LDN 범주 분포는 RDN의 영향을 별로 받지 않는다.
i. MN이 AP 인 경우 역시 RDN의 거의 대부분이 AP로 나타난다.
j. DNs가 NP-NP이거나 VP-NP인 경우 MN이 NP가 될 가능성은 98%이고, VP-VP이거나 AP-VP인 경우 MN이 VP일 확률이 99% 이상이며, NP-VP이거나 NP-NP의 나머지 인 경우 VP나 S가 될 확률이 99%이상이다.

위에 제시된 CFG 규칙의 분포적 특징은, 일반적인 기대나 추정과 부합하느냐 그렇지 않으냐를 떠나서, 대규모 말뭉치를 통해 종합적으로 정리되고 검증된 특징이라는 점에 가장 중요한 의의가 있다고 본다. 말뭉치가 한국어에 대한 대표성을 띠는 정도만큼 위의 일반화

는 한국어의 구문 구조상의 구체적인 특성을 올바로 반영하고 있다고 볼 수 있다.

SKT가 얼마만큼 말뭉치로서의 요건을 갖추고 있는가에 대한 검토도 본 연구에서 다루었다. 본격적인 검증은 앞으로 지속적으로 이루어져야할 문제일 것이다. 그러나 SKT의 대표성이나 무결성이 본 연구에서도 중요한 문제점인 만큼 본 연구에 필요한 수준으로 검토와 검증 작업을 거쳤다. 그 결과 SKT가 규모나 대표성의 측면에서 모두 적정 수준이라는 판단을 할 수 있었다. 다만 SKT에 아직 충분히 정제되지 못한 점들이 발견되었고 이러한 점은 정확한 자료 추출에 일부 장애요소로 작용하는 것으로 여겨진다. 이 문제는 아마도 별도의 추후 검증과정을 거쳐 필요한 후속작업으로 이루어져야 할 것으로 본다.[35]

본 연구에서 사용한 추출 도구 Xavier에 대한 소개와 검증도 또한 본 연구 결과의 타당성과 직결되는 문제라 구체적으로 다루어 졌다. Xavier의 핵심 알고리즘을 제시하였으며, 또한 그 도구의 객관성과 타당성을 입증하기 위하여 추출 규모와 결과의 정확성의 차원에서 검증 과정을 거쳤다.

서두에서 언급한 바와 같이, 본 연구는 SKT에 대한 가장 기초적이고 핵심적인 연구를 목표로 하였다. 따라서 본 연구는 SKT를 대상으로 앞으로 이루어질 연구의 초석으로서의 의의도 크다고 본다. 두 가지 예를 들어 이러한 측면을 논의해 보면서 본 논문을 마무리하기로 한다. 우선 본 연구에서 도출된 언어적 일반화 한 가지를 예로 들어 설명하자면, 'S → NP VP' 규칙에서 첫 NP는 주어기능표지(SBJ)가 부착되어 있을 거라는 예측이 98% 수준으로 검증이 되었다는 점을 밝힌 바 있다. 그렇다면 이러한 검증이 무슨 의미를 띠는

35) PKT에서처럼 (Han et al. 2002), SKT의 형식상 오류를 검증 수정할 수 있는 별도의 프로그램을 개발하여 검증을 거치는 것도 좋은 방법이라고 본다.

가? 이러한 검증을 출발점으로 연구자의 관심에 따라서는 98%에 관심을 두어 파서 개발에 활용할 수도 있고, 또는 나머지 2%에 대한 언어학적인 검토를 시도할 수도 있을 것이다.

또한, 본 연구에서는 CFG 규칙을 반영하는 삼각구조 내의 범주간 분포적 특성에 초점을 맞추어 언어적 일반화를 도출하였다. 이러한 연구를 확대하여 삼각구조와 그 주변요소, 즉 계층적으로 바로 상위 요소나 어순상 선행하는 요소, 또는 후행하는 요소와의 분포적 특성을 정리해 보는 것도 또 다른 연구가 될 수 있다고 본다.[36] 물론 각 연구자의 관심과 관점에 따라 무수한 방식의 연구가 가능할 것이다.

본 연구는 이러한 여러 추후 연구의 바탕이 되는, SKT의 가장 기본적인 특성을 정리하고 분석하여 한국어의 구문 구조 특성을 밝히는데 기여하였다는 점에서 가장 큰 의의가 있다고 본다.

---

36) 이러한 점을 제시해 준 노용균 선생님께 감사드린다.

## 참고문헌

강범모・김의수. 2004. “세종 구문분석 말뭉치를 위한 구문 분석 방법.”「코퍼스와 어휘 데이터베이스」 서울: 월인.

국립국어원. 2007.「21세기 세종계획 최종 성과 발표회 자료집」 문화관광부·국립국어원.

김홍규・강범모. 2000.「한국어 형태소 및 어휘 사용 빈도의 분석 1」 서울: 고려대학교 민족문화연구원.

김홍규・강범모. 2004.「한국어 형태소 및 어휘 사용 빈도의 분석 2」 서울: 고려대학교 민족문화연구원.

김홍규 외. 2003.「21세기 세종계획 국어 기초자료 구축 연구보고서」 문화관광부.

신서인. 2006.「구문 분석 말뭉치를 이용한 한국어 문형 연구」 서울대학교 박사학위 논문.

이윤표. 1989.「國語 空範疇의 연구」 고려대학교 박사학위 논문.

임홍빈・이홍식. 2002.「한국어 구문분석 방법론」 서울: 한국문화사.

장석진. 1995.「정보기반 한국어 문법」 서울: 한신문화사.

Abeillé, Anne. 2003. Treebanks: Building and Using Parsed Corpora. Dordrecht; Boston: Kluwer Academic Publishers.

Charniak, Eugene. 1996. Tree-bank Grammars. Paper presented at AAAI'96, Portland, Oregon.

Han, Chung-hye, Na-Rae Han, Eon-Suk Ko, and Martha Palmer. 2002. Development and Evaluation of a Korean Treebank and its Application to NLP. Paper presented at the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Spain.

Han, Na-Rae. 2006. Korean Zero Pronouns: Analysis and Resolution, The University of Pennsylvania, Ph. D.

Lee, Sun-Hee, Donna K. Byron, and Whitney Gegg-Harrison. 2004. Annotations for Zero Pronoun Resolution in Korean Using the Penn Korean Treebank. Paper presented at the Third Workshop on Treebank and Linguistic Tools, Tuebingen,

Germany.

Manning, Christopher D. and Hinrich Schütze. 1999. Foundations of statistical natural language processing. MA: MIT Press.

Rim, Hae-Chang. 2001. Language Resources In Korea. Paper presented at ALR-1 Workshop.

*Language Information* 9, 2008, pp. 87-139

# Probabilistic Context-Free Grammar Rules based on Sejong Korean Treebank

Choe, Jae-Woong, Sang-houn Song, Jieun Jeon

Keywords Sejong Korean Treebank, probabilistic context-free grammar rules, corpus, frequency, parsed nodes, trees

Abstract

The Sejong Korean Treebank (SKT) was built as part of 10 year government-sponsored Sejong project, and more than 80 million graphic-word Korean parsed corpus has been released to the public at the end of 2007. The purpose of this paper is to extract Context-Free Grammar (CFG) rules from SKT and to draw some linguistic generalizations based on the CFG rules. We introduce an extraction algorithm that was used in this study and prove that it meets the minimal requirements as an objective extraction method in terms of its precision and recall rates. Then our discussion of the extracted CFG rules proceed in terms of the minimal tree structure containing a mother node (MN) and its two daughter nodes (Left DN, Right DN). We arrive at various linguistic or stochastic generalizations restricting the distribution of the categories in the minimal tree structure for Korean, for example, one that states 'In more than 95% of the cases that involve S, VP, NP, VNP, and AP, MN and RDN share the same category.' We provide most of the detailed statistical information regarding the basic properties of SKT and the CFG rules derived from it.

논문투고일 : 2008년 1월 30일
심사완료일 : 2008년 2월 22일
게재확정일 : 2008년 2월 27일

제9호 언어정보

2008년 3월 31일 발행



고려대학교 언어정보연구소
136-701 서울특별시 성북구 안암동
5가 1번지
전화 : 02-3290-1648 (02)3290-1648
팩스 : 02-921-4376 (02)921-4376
홈페이지 : http://www.korea.ac.kr/~rili

발행인 김상열
발행처 책사랑
서울시 동대문구 용두동 767-1 (201호)
전화 (02)929-4547 / 팩스 (02)929-4548
E-mail;idid-yeol@hanmail.net
홈페이지:www.chaeksarang.co.kr
등록 제7-850

ISSN 1226-8011

값 10,000원