

TTA Standard

정보통신단체표준(국문표준)

TTAK.KO-10.0853

제정일: 2015년 12월 16일

의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법

Dependency Tag Sets and Dependency
Relation Establishment Methods for
Constructing Dependency Tagged Corpora



한국정보통신기술협회
Telecommunications Technology Association

의존 구문 분석 말뭉치 구축을 위한
의존 관계 태그 세트 및
의존 관계 설정 방법

Dependency Tag Sets and Dependency Relation
Establishment Methods for Constructing
Dependency Tagged Corpora



본 문서에 대한 저작권은 TTA 에 있으며, TTA 와 사전 협의 없이 이 문서의 전체 또는 일부를 상업적 목적으로 복제 또는 배포해서는 안 됩니다.

서 문

1. 표준의 목적

대규모 말뭉치(corpus)는 자연어 처리 분야에서 가장 기초적인 언어 자원으로 그 중요성이 높아지고 있다. 특히 말뭉치를 분석하여 가공한 의존 구문 분석 말뭉치(Dependency Tagged Corpus)는 의존 구문 분석이 개발에 있어서 필수적인 언어 자원으로 그 구축과 관리에 많은 비용이 소요되므로 연구 개발자 간에 공유가 가능해야 하고, 기존에 구축되어 있는 것을 재사용할 수 있어야 한다. 본 표준은 의존 구문 분석 말뭉치의 재사용 및 공유를 가능하게 하는 의존 관계 태그 세트 및 의존 관계 설정 방법의 표준을 제공하기 위해 작성되었다.

2. 주요 내용 요약

지난 2000년대 중반, 21세기 세종계획 연구 결과물인 세종 구구조 구문 분석 말뭉치가 구축 및 배포되어 많은 구구조 기반 한국어 구문 분석 연구가 가능하게 되었다. 그리고, 2000년대 중반 이후 의존구조 기반 구문 분석이 구문 분석 연구의 주요 흐름으로 자리잡으면서, 많은 구문 분석 연구자들이 세종계획 구구조 말뭉치를 의존구조 말뭉치로 변환하여 사용하게 되었다. 하지만, 아직 한국어 의존구조의 표현에 대한 명확한 기준 체계가 없어서, 각 연구자들마다 서로 다른 의존구조 변환 말뭉치를 이용하여 개별적인 연구를 수행하고 있어 연구의 효율성이 크게 떨어지는 문제가 발생하게 되었다.

이와 같은 문제를 해결하기 위하여 본 표준에서는 한국어 문장을 의존구조에 기반하여 구문 분석하기 위한 표준을 수립하고자 한다.

본 표준은 기구축된 대용량의 세종 구구조 구문 분석 말뭉치를 활용할 수 있도록 하기 위하여 세종 구구조 구문 분석 말뭉치와 동일한 태그 세트를 사용하였음을 밝힌다. 그리고, 해외 의존 구문 분석 연구의 결과물을 한국어에서도 활용할 수 있도록 영어권 연구에서 일반적으로 적용되는 단일 지배소 원칙(Single head constraint)과 투사성 원칙(Projective Constraint)을 같이 적용하였다. 마지막으로 한국어의 고유 특성을 반영하기 위하여 지배소 후위 원칙을 적용하였음을 밝힌다.

3. 표준 적용 산업 분야 및 산업에 미치는 영향

언어분석은 정보 추출, 정보 검색, 질의 응답 등의 자연어 처리 분야에서 기반이 되는 기술이며, 본 표준은 언어분석 과정 중 구문 분석에 대한 것이다. 구문 분석은 문장의 구조를 분석하는 기술로, 문장의 의미를 이해하기 위해 필수적인 기술이다. 정보 추출, 정보 검색 등의 응용 분야 기술은 구문 분석 결과에 의존적이므로 응용 분야의 필요에 부합되도록 구문 관계 태그 세트 및 구문 관계 설정 방법의 표준안을 마련할 필요가 있다. 응용 분야의 필요에 적합한 의존 구문 분석 말뭉치 구축 표준안이 나오면, 표준안에 따르는 의존 구문 분석 말뭉치가 다양한 기관에서 만들어지고 공유되면서 의존 구문 분석이기 개발 생태계가 자연스럽게 조성되고, 나아가서는 의존 구문 분석이기를 활용하는 응용 분야의 기술 발전에도 기여할 것이다.

4. 참조 표준(권고)

4.1. 국외 표준(권고)

- 해당 사항 없음.

4.2. 국내 표준

- 해당 사항 없음.



5. 참조 표준(권고)과의 비교

5.1. 참조 표준(권고)과의 관련성

- 해당 사항 없음.

5.2. 참조한 표준(권고)과 본 표준의 비교표

- 해당 사항 없음.

6. 지식 재산권 관련 사항

본 표준의 ‘지식 재산권 요약서’ 제출 현황은 TTA 웹사이트에서 확인할 수 있다.

※본 표준을 이용하는 자는 이용함에 있어 지식 재산권이 포함되어 있을 수 있으므로, 확인 후 이용한다.

※본 표준과 관련하여 접수된 확약서 이외에도 지식 재산권이 존재할 수 있다.

7. 시험 인증 관련 사항

7.1. 시험 인증 대상 여부

- 해당 사항 없음.

7.2. 시험 표준 제정 현황

- 해당 사항 없음.

8. 표준의 이력 정보

8.1. 표준의 이력

판수	제정·개정일	제정·개정 내역
제 1 판	2015.12.16.	제정 TTAK.KO-10.0853

8.2. 주요 개정 사항

- 해당 사항 없음.

Preface

1. Purpose of Standard

Large-scale corpora become more important in natural language processing. Specially, tagged corpora obtained by attaching dependency tags and relations to the lexical items in the raw corpora are essential language resources for dependency structure analysis and require tremendous cost for their development and management. As a result, it is desirable that the tagged corpora should be reusable and sharable among the research community. The purpose of this recommendatory standard is to provide commonly usable guidelines for the dependency tag sets and the relation establishment methods for reusing and sharing the tagged corpora.

2. Summary of Contents

In the mid-2000s, the phrase structure tagged corpus of 21st Century Sejong Project enables many phrase structure analysis studies for Korean. And, since the mid-2000s, dependency structure analysis became the mainstream of the structure analysis area, many researchers used the Sejong phrase structure tagged corpus after converting to a dependency tagged corpus. But, because there isn't a clear representation standard for the Korean dependency structure, each of the individual researchers have carried out research based on different dependency structures, and the research efficiency are significantly deteriorated.

In order to solve this problem, this standard is to establish dependency tag sets and dependency relation establishment methods for Korean.

Because the Sejong phrase structure tagged corpus is the largest structure tagged corpus for Korean, this standard used same tag sets with the sejong corpus to utilize it. In addition, to apply results of foreign dependency structure analysis studies to Korean, the single head constraint and the projective constraint, which are generally used constraints in English study, are applied. Lastly, in order to reflect the unique characteristics of Korean, the head final principle is also applied.

3. Applicable Fields of Industry and its Effect

Language analysis is the essential technology in the natural language processing areas for information retrieval, information search and question answering. The standard is related with the technology of dependency structure analysis, the core technology among various language analysis technologies. Since the application technologies such as information retrieval are heavily dependent on the results of dependency structure analysis, it is required to establish a standard for the dependency tag sets and dependency relation establishment methods to satisfy the needs of the applications. The standardized dependency tag sets and dependency relation establishment methods will contribute to the development of the eco system of the dependency structure analysis and technology enhancement of various application areas by producing and sharing dependency tagged corpus in many different institutes.

4. Reference Standards (Recommendations)

4.1. International Standards (Recommendations)

– None

4.2. Domestic Standards

– None

5. Relationship to Reference Standards (Recommendations)

5.1. Relationship of Reference Standards (Recommendations)

– None

5.2. Differences between Reference Standard(Recommendation) and this Standard

– None

6. Statement of Intellectual Property Rights

IPRs related to the present document may have been declared to TTA. The information pertaining to these IPRs, if any, is available on the TTA Website.

No guarantee can be given as to the existence of other IPRs not referenced on the TTA website.

And, please make sure to check before applying the standard.

7. Statement of Testing and Certification

7.1. Object of Testing and Certification

– None

7.2. Standards of Testing and Certification

– None

8. History of Standard

8.1. Change History

Edition	Issued date	Outline
The 1st edition	2015.12.16.	Established TTAK.KO-10.0853

8.2. Revisions

– None

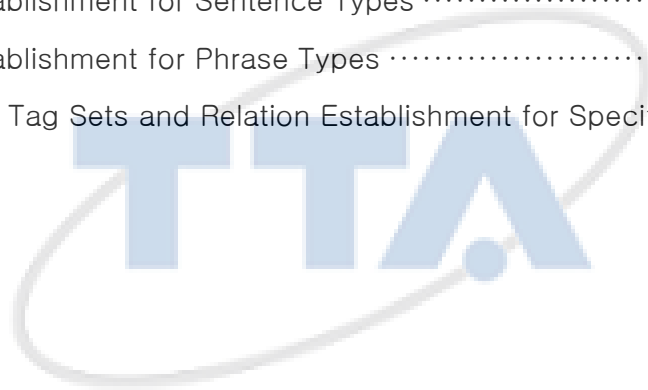
목 차

1. 개요	1
2. 표준의 구성 및 범위	1
3. 참조 표준(권고)	2
4. 용어 정의 및 약어	4
5. 의존 관계 태그 세트 및 의존 관계 설정 방법	5
5.1. 기본 원칙	5
5.2. 의존 관계 태그 세트 설정 방법	5
5.3. 문장 유형별 의존 관계 설정 방법	7
5.4. 세부 구별 태깅 가이드라인	13
5.5. 세부 유형별 가이드라인	19



Contents

1. Introduction	1
2. Constitution and Scope	1
3. Reference Standards(Recommendations)	2
4. Terms Definitions and Abbreviations	4
5. Dependency Tag Sets and Dependency Relation Establishment Methods	5
5.1. Basic Principles	5
5.2. Dependency Tag Sets	5
5.3. Relation Establishment for Sentence Types	7
5.4. Relation Establishment for Phrase Types	13
5.5. Dependency Tag Sets and Relation Establishment for Specific Cases	19



의존 구문 분석 말뭉치 구축을 위한

의존 관계 태그 세트 및 의존 관계 설정 방법

(Dependency Tag Sets and Dependency Relation Establishment Methods for Constructing Dependency Tagged Corpora)

1. 개요

대규모 말뭉치(corpus)는 자연어 처리 분야에서 가장 기초적인 언어 자원으로 그 중요성이 높아지고 있다. 특히 말뭉치를 분석하여 가공한 의존 구문 분석 말뭉치(Dependency Tagged Corpus)는 의존 구문 분석 시스템 개발에 있어서 필수적인 언어 자원이다.

의존 구문 분석 말뭉치는 그 중요성이 큰 만큼 구축과 관리에 있어서 많은 비용이 소요되므로 연구 개발자 간에 공유가 가능해야 하고, 기존에 구축되어 있는 것을 재사용할 수 있어야 한다. 의존 구문 분석 말뭉치의 공유와 재사용이 가능하기 위해서는 말뭉치에 쓰이는 의존 관계 태그 세트와 의존 관계 설정 방법에 대한 표준화가 선행되어야 한다.

표준 의존 관계 태그 세트 및 의존 관계 설정 방법이 관련 연구 개발자 누구나 수용할 수 있는 체계로 되어 있고, 표준에 따라 구축된 의존 구문 분석 말뭉치가 구축된다면, 기존 서로 다른 의존 관계 설정 체계를 기반으로 연구하던 연구 내용들을 하나의 동일한 기준을 이용하여 비교할 수 있어서 의존 구문 분석 분야의 기술 개발 효율성이 증대될 것으로 기대된다.

2. 표준의 구성 및 범위

본 표준은 말뭉치를 이용하여 문장의 의존 구문구조를 분석할 때, 각 어절에 부착되는 태그 세트 및 의존 관계 설정 방법에 대한 기준을 제공한다. 의존 관계 태그 세트에 대한 전체 체계를 제시하고, 한국어 문장 유형별로 의존 관계 설정 방법에 대한 기준을 기술한다.

본 표준에서 제시한 의존 관계 태그 세트 및 의존 관계 설정 방법은 전산처리적 관점을 기본으로 설정하였기 때문에, 국어학이나 언어학에서 연구하고 있는 통사론과는 그 목적과 용도가 분명히 다를 것을 밝힌다. 본 권고안이 국어학이나 언어학에서 연구한 결과를 가능한 많이 참고하지만, 실제 실행되는 시스템들의 일반적인 환경적 제약과 개발에 드는 노력 및 자원의 제한을 고려한 전산 처리의 효율성을 고려하는 관점에서 의존 관계 태그 세트 및 설정 방법 기준을 설정하였다. 각 문장 및 구문 유형별로 제시되는 예제는 해당 유형의 모든 가능한 예제가 아니라 대표적인 유형만을 예시한 것이다.

본 표준에서 제시하는 의존 관계 태그 세트 및 의존 관계 설정 방법의 가장 주된 용도는 자동 의존 구문 분석 시스템을 개발하기 위한 학습자료로 활용되는 말뭉치 구축이다. 이러한 용도는 의미론적 관점에서 문장을 분석하는 연구자들의 분류체계와 관점이 다르다는 것을 전제로 한다.

일반적인 자연어처리 분석 과정은 형태소 분석, 구문 분석, 의미분석의 파이프라인을 따라 진행된다. 즉, 구문 분석은 형태소 분석 결과가 입력되고, 구문 분석 결과를 의미분석의 입력으로 제공하게 된다. 따라서, 형태소 분석에서 분석하지 않는 태그 세트를 구문 분석의 태그 세트에서 제시한다면 자연어처리 파이프라인이 원활하게 동작하지 않을 것이다. 또한 의미분석에서 필요로 하는 구문구조 결과를 구문 분석에서 제시하지 못한다면 의미분석 기술 개발의 한계로 작용하게 된다. 따라서, 의존 관계 태그 세트 및 의존 관계 설정 방법 기준은 한국어 형태소 분석 태그 세트와의 호환성과 의미분석에서의 활용성을 고려하여 기준이 설정되어야 한다는 점을 주지해야 한다.

본 표준의 구성은 다음과 같다. 우선, 전산처리학적 관점에서 설정한 한국어 의존 관계 분석을 위한 기본원칙을 소개한다. 그리고, 의존 관계 태그 세트 및 그 분석 예를 설명한다. 다음으로, 한국어 문장 유형별 의존 관계 설정 방법과 구문 유형별 의존 관계 설정 방법을 설명한다. 마지막으로, 기타 세부 유형별 의존 관계 설정 방법을 소개한다.

끝으로 이 권고안을 활용하는 사람들은 통계적 자연언어 처리의 기본 개념과 제약사항 등을 숙지하고 있음을 가정한다.

3. 참조 표준(권고)

3.1. 국외 표준(권고)

- 해당 사항 없음.

3.2. 국내 표준

- 해당 사항 없음.

4. 용어 정의 및 약어

4.1. 언어 자원(Language Resource)

최근에 유행하는 경험적이고 관찰적인 자연언어처리 기법에서 귀납적으로 언어지식을 추출하기 위해 필요한 대량의 다양한 언어현상을 담고 있는 기초자료이며, 전자사전, 말뭉치, 품사 부착 말뭉치, 구문구조 부착 말뭉치, 의미역 부착 말뭉치 등을 말한다. 이 중, 본 표준안과 직접적으로 관련이 있는 ‘말뭉치’, ‘의존 구문 분석 말뭉치’에 대해서만 정의를 내리기로 한다.

4.2. 말뭉치(Corpus)

신문, 소설, 교과서 등 각 분야의 저서나 출판물을 정해진 저장양식과 표현양식에 따라 전자화한 혹은 기계가독형(machine readable) 텍스트 모음을 말한다.

4.3. 의존 구문 분석(Dependency Parsing)

자연어 문장의 구문구조를 “지배소와 피지배소”의 의존 관계로 분석하는 구문 분석 방법론을 말한다. 예를 들어, “백설공주가 예쁜 사과를 먹었다.” 문장을 의존 구문 분석 하면 {(백설공주가 →NP_SBJ 먹었다), (예쁜 →VP_MOD 사과를), (사과를 → NP_OBJ 먹었다.)} 와 같은 결과가 나온다. 이를 통해서 컴퓨터 시스템은 먹었다는 이벤트의 주격 및 목적격을 분석할 수 있고, 예쁘다라는 관형어가 수식하는 대상이 백설공주인지 사과인지 여부를 구분할 수 있다.

4.4. 의존 구문 분석 말뭉치(Dependency Tagged Corpus)

미리 정의된 의존 관계 태그 세트 및 의존 관계 설정 기준에 따라 각 문장의 어절 별로 지배소를 설정하고, 의존 관계를 부착한 형태의 텍스트 모음을 말한다.

4.5. 의존 관계 태그 세트(Dependency Relation Tag Sets)

문장을 의존 구문 분석 하였을 때, 각 어절이 지배소와 가질 수 있는 의존 관계의 유형을 정리한 체계이다. 의존 관계 태그 세트는 구문 태그 세트와 기능 태그 세트를 조합하여 사용한다. 구문 태그 세트는 대상 어절이 체언인지, 용언인지, 관형사인지, 부사인지 등을 구분하고, 기능 태그 세트는 지배소와 가지는 관계가 주격인지, 목적격인지, 관형격인지 등을 구분한다.

4.6. 의존 관계 설정 방법(Dependency Relation Establishment Methods)

문장을 의존 구문 분석 하였을 때, 각 어절의 지배소를 설정하기 위한 방법을 말한다. 일반적으로 주어, 목적어, 부사어 등은 서술어에 의존하도록 의존 관계를 설정한다. 다만, 한국어에서 빈번한 생략현상 때문에, 한 어절이 두 개 이상의 어절을 지배소로 가질 수 있는 경우가 발생한다. (예를 들어, “백설공주가 사과를 먹고 잠이 들었다.”와 같은 경우이다.) 이와 같은 경우, 미리 정의한 의존 관계 설정 기준에 따라 지배소를 설정한다.

4.7. 전산 처리적 관점과 용도

전산처리적인 관점이라는 것은 좀더 엄밀하게 말하면, 사람들에게 설명을 위한 형태로 언어현상의 특징을 설명하는 형태론, 구문론이 아니다. 형태소 분석기, 구문 분석기 등과 같은 프로그램으로 하여금 주어진 컴퓨팅 환경과 제한된 메모리와 계산 능력을 활용하여 관찰 가능하고 처리 가능한 범위 내의 문장을 입력으로 받아 정해진 형태의 출력을 제한된 시간 내에 출력해야 한다는 제약적 용도를 고려한 관점이다.

5. 의존 관계 태그 세트 및 의존 관계 설정 방법

본 표준의 의존 관계 태그 세트 및 의존 관계 설정의 방법은 세종계획에서 구축된 구구조 말뭉치를 의존구조 말뭉치로 변환하여 활용이 가능하도록 하고, 한국어 의미 처리를 위하여 문장의 구문구조를 의존 관계로 효율적으로 표현할 수 있도록 하는데 목적을 두었다. 더불어, 영어권의 의존구문 분석 연구동향을 참고하여 영어권의 구문 분석 연구 결과를 한국어 의존구문 분석 연구에도 같이 활용할 수 있도록 노력하였다.

본 표준 의존 관계 태그 세트 및 의존 관계 설정의 기본 원칙은 21 세기 세종계획의 결과물인 “구문 분석 말뭉치 분석 지침(Ver. 2003-1)”을 참고했음을 밝힌다.

5.1. 기본 원칙

- (1) 자연언어처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서도 크게 벗어나지 않도록 한다.
- (2) 문장의 표층 구조를 중시하여 분석한다.
- (3) 의존 관계 분석의 기본 단위로 어절을 사용한다.
- (4) 지배소 후위 원칙에 따라 각 어절의 지배소는 자신보다 뒤에 위치하도록 분석한다.
- (5) 각 어절은 1 개의 지배소를 가진다. (Single-Head Constraint)
- (6) 각 어절 및 지배소쌍은 서로 교차하지 않는다. (Projective Constraint)
- (7) 보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.
- (8) 원칙적으로 접속과 내포를 구별하지 않으며, 접속절은 모두 부사절로 분석한다. (다만, 명사구 접속은 인정한다.)
- (9) 하나의 주어가 모문과 내포문 모두에 관련되어 있으면 모문과 내포문의 관계에 따라 해당 주어의 지배소를 결정한다.

5.2. 의존 관계 태그 세트 설정 방법

- 각 어절은 자신 어절과 지배소 어절 사이의 관계를 표현하는 의존 관계 태그를 가진다.
- 의존 관계 태그는 아래 구문 태그와 기능 태그를 결합하여 사용한다. (예: NP_SBJ, VP_MOD 등)

5.2.1. 구문 태그 세트

구문 태그	의미
NP	체언 (명사, 대명사, 수사)
VP	용언 (동사, 형용사, 보조용언)
AP	부사구
VNP	긍정 지정사구 (명사+이다)
DP	관형사구

IP	감탄사구 (호칭 및 대답 등의 표현)
X	의사 구 (pseudo phrase, 조사 단독 어절 또는 기호 등)
L	부호 (왼쪽 괄호 및 따옴표)
R	부호 (오른쪽 괄호 및 따옴표)

5.2.2. 기능 태그 세트

기능 태그	의미
SBJ	주어
OBJ	목적어
MOD	관형어 (체언 수식어)
AJT	부사어 (용언 수식어)
CMP	보어
CNJ	접속어 (~와)

- 의존 관계 설정 및 의존 관계 태그 부착 결과는 아래와 같다.

(가) 멜라닌은 사람의 피부색을 결정하는 중요한 요소이다.

- 멜라닌은 →NP_SBJ 요소이다.
- 사람의 →NP_MOD 피부색을
- 피부색을 →NP_OBJ 결정하는
- 결정하는 →VP_MOD 요소이다.
- 중요한 →VP_MOD 요소이다.
- 요소이다. →VNP ROOT

(나) 역린은 용 목에 거꾸로 난 비늘을 의미한다.

- 역린은 →NP_SBJ 의미한다.
- 용 →NP 목에
- 목에 →NP_AJT 난
- 거꾸로 →AP 난
- 난 →VP_MOD 비늘을
- 비늘을 →NP_OBJ 의미한다.
- 의미한다. →VP ROOT

5.3. 문장 유형별 의존 관계 설정 방법

- 의존 관계 설정을 위한 문장 유형 구분은 일반 언어학 통사론의 기준을 따라 ‘주어-서술어’ 관계가 한 개인 흠문장과 ‘주어-서술어’ 관계가 2 개 이상인 겹문장으로 문장을 분류하여 분석한다.
- 겹문장은 다시 접속문과 내포문으로 구분한다. 다만, 기본원칙 (8) 에 따라 접속문을 따로 분류하지 않고, 접속절을 부사절 내포문으로 분류하여 분석한다.
- 내포문은 명사절 내포문, 관형절 내포문, 부사절 내포문, 서술절 내포문, 직접 인용절 내포문, 간접 인용절 내포문의 6 가지 유형으로 분류하여 분석한다.¹

5.3.1. 흠문장(단문) 분석 방법

- 흠문장을 이루는 문장의 구성 성분은 크게 주어, 목적어, 관형어, 부사어, 보어, 서술어로 구분한다.

- (1) 주어는 SBJ 기능 태그를 가지고, 흠문장 서술어에 의존하도록 분석한다.
- (2) 목적어는 OBJ 기능 태그를 가지고, 흠문장 서술어에 의존하도록 분석한다.
- (3) 관형어는 MOD 기능 태그를 가지고, 수식하는 명사구에 의존하도록 분석한다.
- (4) 부사어는 AJT 기능 태그를 가지고, 흠문장 서술어에 의존하도록 분석한다.
- (5) 보어는 CMP 기능 태그를 가지고, 흠문장 서술어에 의존하도록 분석한다.
- (6) 흠문장의 경우, 서술어는 문장의 가장 뒤에 위치하며, ROOT 어절에 의존하도록 분석한다.

¹ 내포문의 유형별 예문은 아래와 같다.

- 명사절: 우리는 시민의 관전태도도 그만큼 성숙했음을 잊지 말아야 한다.
- 관형사절: 내가 좋아하는 꽃은 들국화다. (관계관형절),
내가 사진을 좋아하는 사실을 친구들은 다 안다. (동격관형절)
- 부사절: 멜라닌은 자외선을 차단해서 자외선으로부터 피부를 보호해 준다.
비가 와서 땅이 미끄럽다.
- 인용절: 그는 그녀가 그 일을 해냈다고 밝혔다. (간접 인용절)
그는 “그녀가 그 일을 해냈어.”라고 말했다. (직접 인용절)
- 서술절: 꼬끼리가 코가 길다.
할머니께서 치아가 아프시다.

- 흘문장의 구문 분석 예는 아래와 같다.

(가) 멜라닌은 사람의 피부색을 결정한다.

- 멜라닌은 →NP_SBJ 결정한다.
- 사람의 →NP_MOD 피부색을
- 피부색을 →NP_OBJ 결정한다.
- 결정한다. →VP ROOT

5.3.2. 겹문장(복문) 분석 방법

- 겹문장은 명사절 내포문, 관형절 내포문, 부사절 내포문, 서술절 내포문, 직접 인용절 내포문, 간접 인용절 내포문의 6 가지 유형으로 분류하여 분석한다.

(1) 명사절, 부사절 및 간접 인용절 내포문 분석 방법

(1.1) 모문과 내포문의 주어 및 서술어가 각각 다른 경우

- 모문의 주어는 모문의 서술어로, 내포문의 주어는 내포문의 서술어로 분석한다.
- 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.
- 명사절, 부사절 및 간접 인용절 별 예문은 아래와 같다.

(가) 우리는 시민의 관전태도도 그만큼 성숙했음을 잊지 말아야 한다.

- 우리는 →NP_SBJ 잊지
- 시민의 →NP_MOD 관전태도도
- 관전태도도 →NP_SBJ 성숙했음을
- 그만큼 →AP 성숙했음을
- 성숙했음을 →VP_OBJ 잊지
- 잊지 →VP 말아야
- 말아야 →VP 한다.
- 한다. →VP ROOT

(나) 그가 돌아와줘서 우리는 정말 고마웠어.

- 그가 →NP_SBJ 돌아와줘서
- 돌아와줘서 →VP_AJT 고마웠어.
- 우리는 →NP_SBJ 고마웠어.
- 정말 →AP 고마웠어.

- 고마웠어. →VP ROOT

(다) 그가 돌아왔다고 그녀가 알려줬어.

- 그가 →NP_SBJ 돌아왔다고
 - 돌아왔다고 →VP_CMP 알려줬어.
 - 그녀가 →NP_SBJ 알려줬어.
 - 알려줬어. →VP ROOT

(1.2) 모문과 내포문의 주어가 다르고, 서술어가 동일한 경우

- 모문의 주어와 내포문의 주어가 동일한 서술어에 의존하도록 분석한다.

(1.3) 모문과 내포문의 주어가 같고, 서술어가 다른 경우

- 기본원칙 (9)에 따라 모문과 내포문의 관계에 따라 주어의 지배소를 결정한다.
 - 명사절, 부사절 및 간접 인용절 내포문의 경우, 주어가 내포문의 서술어를 지배소로 가지도록 분석한다.
 - 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.
 - 명사절, 부사절 및 간접 인용절 별 예문은 아래와 같다.

(가) 그는 목표를 이루었음에 매우 감사했다.

- 그는 →NP_SBJ 이루었음에
 - 목표를 →NP_OBJ 이루었음에
 - 이루었음에 →NP_AJT 감사했다.
 - 매우 →AP 감사했다.
 - 감사했다. →VP ROOT

(나) 멜라닌은 자외선을 차단해서 자외선으로부터 피부를 보호해 준다.

- 멜라닌은 →NP_SBJ 차단해서
 - 자외선을 →NP_OBJ 차단해서
 - 차단해서 →VP 보호해
 - 자외선으로부터 →NP_AJT 보호해
 - 피부를 →NP_OBJ 보호해
 - 보호해 →VP 준다.
 - 준다. →VP ROOT

(다) 그녀가 그 일을 했다고 스스로 밝혔다.

- 그녀가 →NP_SBJ 했다고
- 그 →DP 일을
- 일을 →NP_OBJ 했다고
- 했다고 →VP_CMP 밝혔다.
- 스스로 →AP 밝혔다.
- 밝혔다. →VP ROOT

(2) 관형절 내포문 분석 방법

(2.1) 모문과 내포문의 주어 및 서술어가 다른 경우

- 명사절, 부사절 및 간접 인용절 내포문 분석 방법과 동일하게 분석한다.
- 모문의 주어는 모문의 서술어로, 내포문의 주어는 내포문의 서술어로 분석한다.
- 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.

(가) 내가 좋아하는 꽃은 들국화이다.

- 내가 →NP_SBJ 좋아하는
- 좋아하는 →NP_MOD 꽃은
- 꽃은 →NP_SBJ 들국화이다.
- 들국화이다. →VNP ROOT

(나) 내가 사진을 좋아하는 사실을 친구들은 다 안다.

- 내가 →NP_SBJ 좋아하는
- 사진을 →NP_OBJ 좋아하는
- 좋아하는 →VP_MOD 사실을
- 사실을 →NP_OBJ 안다.
- 친구들은 →NP_SBJ 안다.
- 다 →AP 안다.
- 안다. →VP ROOT

(2.2) 모문과 내포문의 주어가 다르고, 서술어가 동일한 경우

- 명사절, 부사절 및 간접 인용절 내포문 분석 방법과 동일하게 분석한다.

(2.3) 모문과 내포문의 주어가 같고, 서술어가 다른 경우

- 관형절은 모문 내의 체언을 수식하는 성격이 더욱 강하기 때문에, 동일 주어를 모문의 서술어로 연결하도록 한다.
- 일반적으로 종차 개념에 의한 정의 구문 유형([A 는 ~~한 B 이다.])이 이 유형에 속한다.
- 예문은 아래와 같다.

(가) 모나리자는 레오나르도 다빈치가 그린 초상화이다.

- 모나리자는 →NP_SBJ 초상화이다.
- 레오나르도 →NP 다빈치가
- 다빈치가 →NP_SBJ 그린
- 그린 →VP_MOD 초상화이다.
- 초상화이다. →VNP ROOT

(나) 세포벽은 식물 세포의 가장 바깥층을 에워싸고 있는 약간 두꺼운 막이다.

- 세포벽은 →NP_SBJ 막이다.
- 식물 →NP 세포의
- 세포의 →NP_MOD 바깥층을
- 가장 →AP 바깥층을
- 바깥층을 →NP_OBJ 에워싸고
- 에워싸고 →VP 있는
- 있는 →VP_MOD 막이다.
- 약간 →AP 두꺼운
- 두꺼운 →VP_MOD 막이다.
- 막이다. →VNP ROOT

(다) 멜라닌은 사람의 피부색을 결정하는 주요 요소이다.

- 멜라닌은 →NP_SBJ 요소이다.
- 사람의 →NP_MOD 피부색을
- 피부색을 →NP_OBJ 결정하는
- 결정하는 →VP_MOD 요소이다.
- 주요 →NP 요소이다.
- 요소이다. →VNP ROOT

(라) 이 사료는 가축들이 먹는 음식이다.²

- 이 →DP 사료는
- 사료는 →NP_SBJ 음식이다.
- 가축들이 →NP_SBJ 먹는
- 먹는 →VP_MOD 음식이다.
- 음식이다. →VNP ROOT

(3) 서술절 내포문 분석 방법

- 이종주어 구문과 동일하게 각 구성성분 별로 동일한 서술어에 의존하도록 분석한다.
- 예문은 아래와 같다.

(가) 나는 그 시계가 필요했다.

- 나는 →NP_SBJ 필요했다.
- 그 →DP 시계가
- 시계가 →NP_SBJ 필요했다.
- 필요했다. →VP ROOT

(나) 그는 그녀가 자랑스러웠다.

- 그는 →NP_SBJ 자랑스러웠다.
- 그녀가 →NP_SBJ 자랑스러웠다.
- 자랑스러웠다. →VP ROOT

(4) 직접 인용절 분석 방법

- 기본적으로 문장 부호에 의하여 구분된 단위를 준수하여 분석한다.
- 인용부호가 있는 직접 인용문은 모문과 내포문을 인용부호에 의해서 구분할 수 있으므로, 모문은 모문 구조 내의 주어와 서술어 간의 의존 관계를 연결하고 내포문은 내포문 내의 주어와 서술어 간의 의존 관계를 연결하여 분석한다.
- 예문은 아래와 같다.

(가) 비평가 칼라일이 "인도와도 바꿀 수 없다"고 말하였다.

² “이 사료는 가축들이 먹는 것이다.”와 같은 예문의 경우, “먹는 것이다”를 의사 보조용언 구성으로 보고 관형절 내포문으로 분석하지 않는다. 즉, “사료는 →NP_SBJ 먹는”, “먹는 →VP_MOD 것이다.”와 같이 분석한다.

- 비평가 →NP 칼라일이
- 칼라일이 →NP_SBJ 말하였다.
- “인도와도 →NP_AJT 바꿀
- 바꿀 →VP_MOD 수
- 수 →NP_SBJ 없다.”고
- 없다.”고 →VP_CMP 말하였다.
- 말하였다. →VP ROOT

5.4. 세부 구별 태깅 가이드라인

5.4.1. [관형어 + 명사 + 명사] 유형 분석

- 관형어가 명사구를 수식할 때, 명사구 중 관형어의 지배소는 관형어가 의미적으로 수식하는 어휘를 지배소로 분석한다.

- 예문은 아래와 같다.

(가) 무분별한 포획 문제

- 무분별한 →VP_MOD 포획
- 포획 →NP 문제

(나) 고려의 총신 정몽주

- 고려의 →NP_MOD 총신
- 총신 →NP 정몽주

(다) 조선의 제 3 대 임금

- 조선의 →NP_MOD 임금
- 제 3 대 →NP 임금

(라) 변형생성문법의 이론은 1987 년에 출판된 노엄 촘스키의 저서이다.

- 변형생성문법의 →NP_MOD 이론은
- 이론은 →NP_SBJ 저서이다.
- 1987 년에 →NP_AJT 출판된
- 출판된 →NP_MOD 저서이다.
- 노엄 →NP 촘스키의
- 촘스키의 →NP_MOD 저서이다.

- 저서이다. →VNP ROOT

(마) 결명자의 한자 뜻인 ‘눈을 밝게 띄우는 씨앗’이라는 이름대로

- 결명자의 →NP_MOD 뜻인
- 한자 →NP 뜻인
- 뜻인 →VNP_MOD 씨앗’이라는
- ‘눈을 →NP_OBJ 띄우는
- 밝게 →VP_AJT 띄우는
- 띄우는 →VP_MOD 씨앗’이라는
- 씨앗’이라는 →VNP_MOD 이름대로

5.4.2. 명사구 접속 유형 분석

- 복수 개의 명사구가 접속 또는 나열된 경우, 가장 마지막 명사구에 의존하도록 분석한다.
- 명사구 접속의 예는 아래와 같다.

(가) 비단 피부뿐만이 아니라 털, 눈, 귀, 심지어 뇌에도 존재한다.

- 비단 →AP 피부뿐만이
- 피부뿐만이 →NP_CMP 아니라
- 아니라 →VP 뇌에도
- 털, →NP_CNJ 뇌에도
- 눈, →NP_CNJ 뇌에도
- 귀, →NP_CNJ 뇌에도
- 심지어 →AP 뇌에도
- 뇌에도 →NP_AJT 존재한다.
- 존재한다. →VP ROOT

(나) 매리너스 협곡과 극관이 존재한다.

- 매리너스 →NP 협곡과
- 협곡과 →NP_CNJ 극관이
- 극관이 →NP_SBJ 존재한다.
- 존재한다 →VP ROOT

(다) 주로 왕과 왕세자의 강론 및 정책토론을 주관했다.

- 주로 →AP 주관했다.
- 왕과 →NP_CNJ 왕세자의
- 왕세자의 →NP_MOD 정책토론을
- 강론 →NP_CNJ 정책토론을
- 및 →AP 정책토론을
- 정책토론을 →NP_OBJ 주관했다.
- 주관했다. →VP ROOT

5.4.3. [용언 + 용언] 유형 분석

(1) 본용언 + 본용언

- 본용언이 연속적으로 나타날 경우, 주어를 앞에 위치한 서술어에 연결한다

(가) 그 아내는 들고 오던 공후를 끌어 잡아 타면서

- 그 →DP 아내는
- 아내는 →NP_SBJ 끌어
- 들고 →VP 오던
- 오던 →VP_MOD 공후를
- 공후를 →NP_OBJ 끌어
- 끌어 →VP 잡아
- 잡아 →VP 타면서

(나) 해구보다는 폭이 넓고 얇다.

- 해구보다는 →NP_AJT 넓고
- 폭이 →NP_SBJ 넓고
- 넓고 →VP 얇다.
- 얇다. →VP ROOT

(2) 본용언+보조용언

- 본용언과 보조용언이 연속하여 두 개 이상 나올 때는 주어는 본용언에 연결하고 보조용언은 본용언에 연결한다.

(가) 멜라닌은 물에는 용해되지 않는다.

- 멜라닌은 →NP_SBJ 용해되지
- 물에는 →NP_AJT 용해되지

- 용해되지 →VP 않는다.
- 않는다. →VP ROOT

(3) 의사보조용언 구성

- 주 서술어 다음에 보조용언은 아니지만 서법을 나타내는 언어 단위들이 올 경우, 의사보조용언 구성으로 보고 분석한다.³
- 주어는 주 서술어와 의존 관계를 연결하고, 뒤따르는 언어 단위들은 주서술어의 뒤에 차례대로 의존 관계를 연결한다.⁴
- 의사보조용언 구성의 목록은 아래와 같다.

(가) ‘-ㄹ 수/리(가) 있다/없다’

1. 그는 일어날 수 없었다.

- 그는 →NP_SBJ 일어날
- 일어날 →VP_MOD 수
- 수 →NP_SBJ 없었다.
- 없었다. →VP ROOT

(나) ‘-ㄴ/ㄹ + 의존명사(또는 일반명사)+이다’

의존명사: 것/터/뿐/따름/모양/지경/참/중

일반명사:노릇/예정/길

1. 나는 곧 밥을 먹을 것이다.

- 나는 →NP_SBJ 먹을
- 곧 →AP 먹을
- 밥을 →NP_OBJ 먹을
- 것이다. →VNP ROOT

2. 그가 고향이 그리운 모양이다.

- 그가 →NP_SBJ 그리운
- 고향이 →NP_SBJ 그리운

³ 의사 보조용언 구성은 언어학에서는 양태 표현에 포함되는 개념으로, 추측, 바람, 판단, 행동지시, 의도, 의지 표현 등의 표현이 있다. 양태 표현은 명사, 동사, 어미 등이 함께 어울려서 나타낼 수 있다. 세부적으로 {관형형+의존명사+용언}, {관형형+의존명사+조사+용언}, {연결어미+부정부사+보조용언}, {명사형 전성 어미 -기+용언}, {명사형 전성어미 -기+조사+보조용언}, {부사+(연결어미, 부정부사)+용언} 등의 형태를 가진 것을 의사 보조용언 구성으로 볼 수 있다.

⁴ 의사 보조용언 구성에서 언급되지 않은 의존명사는 의사 보조용언 구성에 해당하지 않고 일반적 구조 분석과 동일하게 분석한다.

- 그리운 →VP_MOD 모양이다.
 - 모양이다. →VNP ROOT
3. 나는 외갓집에 심부름을 갔다 오는 길이다.
- 나는 →NP_SBJ 갔다
 - 외갓집에 →NP_AJT 갔다
 - 심부름을 →NP_OBJ 갔다
 - 갔다 →VP 오는
 - 오는 →VP_MOD 길이다.
 - 길이다. →VNP ROOT

(다) ‘-ㄹ {만/법/듯}하다’

1. 그 영화는 아이들과 같이 볼 만하다.
- 그 →DP 영화는
 - 영화는 →NP_SBJ 볼
 - 아이들과 →NP_AJT 볼
 - 같이 →AP 볼
 - 볼 →VP_MOD 만하다.
 - 만하다. →VP ROOT

(라) ‘-는 말이다’

1. 정녕 일을 그르쳤단 말이냐?
- 정녕 →AP 그르쳤단
 - 일을 →NP_OBJ 그르쳤단
 - 그르쳤단 →VP_MOD 말이냐?
 - 말이냐? →VNP ROOT

(마) ‘-ㄴ/ㄹ 듯(도) 하다’

1. 따스한 손길로 머리를 어루만져 주시는 듯 했다.
- 따스한 →VP_MOD 손길로
 - 손길로 →NP_AJT 어루만져
 - 머리를 →NP_OBJ 어루만져
 - 어루만져 →VP 주시는
 - 주시는 →VP_MOD 듯
 - 듯 →NP_AJT 했다.

- 했다. →VP ROOT

(바) ‘ㄹ 것 같다’ 구성

1. 곧 비가 올 것 같다.

- 곧 →AP 올
 - 비가 →NP_SBJ 올
 - 올 →VP_MOD 것
 - 것 →NP 같다.
 - 같다. →VP ROOT

(사) ‘ㄹ 것을(걸) 그랬다’

1. 거기 나가서 동창들한테 협찬금 받아낼 걸 그랬잖아.

- 거기 →NP_AJT 나가서
 - 나가서 →VP 받아낼
 - 동창들한테 →NP_AJT 받아낼
 - 협찬금 →NP_OBJ 받아낼
 - 받아낼 →NP_MOD 걸
 - 걸 →NP_OBJ 그랬잖아.
 - 그랬잖아. →VP ROOT

(아) ‘-어서는 안된다’

1. 우리는 경계를 늦추어서는 안된다.

- 우리는 →NP_SBJ 늦추어서는
 - 경계를 →NP_OBJ 늦추어서는
 - 늦추어서는 →VP 안된다.
 - 안된다. →VP ROOT

(자) ‘-고 해서’

1. 시간도 없고 해서 그는 친척집에 들르지 않았다.

- 시간도 →NP_SBJ 없고
 - 없고 →VP 해서
 - 해서 →VP 들르지
 - 그는 →NP_SBJ 들르지
 - 친척집에 →NP_AJT 들리지

- 들르지 →VP 않았다.
- 않았다. →VP ROOT

(차) ‘-든지 하다’

1. 모처럼 왔으면 따뜻한 밥 한 그릇 먹고 가든지 해야지.

- 모처럼 →AP 왔으면
- 왔으면 →VP 먹고
- 따뜻한 →VP_MOD 밥이나
- 밥 →NP 그릇
- 한 →DP 그릇
- 그릇 →NP_OBJ 먹고
- 먹고 →VP 가든지
- 가든지 →VP 해야지.
- 해야지. →VP ROOT

5.5. 세부 유형별 가이드라인

5.5.1. 의존 관계 태그 부착 세부 유형 가이드라인

- 1) 본용언에 화용적 기능을 가지는 조사 {가/를}이 붙은 경우는 기능표지를 부착하지 않는다.
- 2) “~즈음, ~쯤” 등과 같이 의미적으로 시간과 공간을 의미하고 조사가 없는 경우 AJT 기능 태그를 부착한다.
- 예문은 아래와 같다.

(1.가) 철수가 밥을 이틀내 먹지를 않았다.

- 철수가 →NP_SBJ 먹지를
- 밥을 →NP_OBJ 먹지를
- 이틀내 →NP_AJT 먹지를
- 먹지를 →VP 않았다.
- 않았다. →VP ROOT

(1.나) 그 산은 그리 높지가 않다.

- 그 →DP 산은
- 산은 →NP_SBJ 높지가

- 그리 →AP 높지가
- 높지가 →VP 않다.
- 않다. →VP ROOT

(2.가) 광복절 즈음 해서 독립기념관을 찾았다.

- 광복절 →NP 즈음
- 즈음 →NP_AJT 해서
- 해서 →VP 찾았다.
- 독립기념관을 →NP_OBJ 찾았다.
- 찾았다. →VP ROOT

5.5.2. 의존 관계 설정 세부 유형 가이드라인

- 1) “~으로”가 의미적으로 “~이다. 그리고”와 같이 사용되었다면, 예외적으로 서술어로 인정한다.
- 2) 부사구 중 "없이"나 '같이'와 같은 부사는 서술어와 마찬가지로 논항을 취할 수 있다.
- 예문은 아래와 같다.

(1.가) 세포벽(細胞壁)은 식물 세포의 가장 바깥층을 에워싸고 있는 약간 두꺼운 막으로, 변형균을 제외한 모든 식물 세포에 존재한다.

- 세포벽(細胞壁)은 →NP_SBJ 막으로,
- (이하 생략)

(1.나) 모나리자는 레오나르도 다 빈치가 그린 초상화로, 현재 프랑스 파리 루브르 박물관에 전시되어 있다.

- 모나리자는 →NP_SBJ 초상화로,
- (이하 생략)

(2.가) 철수는 아무 생각도 없이 길을 나섰다.

- 철수는 →NP_SBJ 나섰다.
- 아무 →DP 생각도
- 생각도 →NP_SBJ 없이
- 없이 →AP 나섰다.
- 길을 →NP_OBJ 나섰다.

- 나셨다. →VP ROOT

(2.나) 난 너와 같이 살고 싶어.

- 난 →NP_SBJ 살고
 - 너와 →NP_AJT 같이
 - 같이 →AP 살고
 - 살고 →VP 싶어.
 - 싶어. →VP ROOT

5.5.3. 부사구 연결 세부 유형 가이드라인

- 1) ‘~부터’ 와 ‘~까지’ (또는 ‘~로’)와 같은 부사구는 각각을 지배소에 연결한다.
- 2) ‘~부터 ~까지의’, ‘~에서 ~에로의’ 와 같이 후행 어절에 조사가 부착된 경우, 선행 어절을 후행 어절에 연결시킨 후, 후행어절을 지배소에 연결한다.
- 예문은 아래와 같다.

(1.가) 그는 작년부터 지금까지 열심히 일했다.

- 그는 →NP_SBJ 일했다.
 - 작년부터 →NP_AJT 일했다.
 - 지금까지 →NP_AJT 일했다.
 - 열심히 →AP 일했다.
 - 일했다. →VP ROOT

(2.가) 한나라의 건국부터 왕망의 치세기까지의 기간이 서한, 혹은 전한으로 알려져 있다.

- 한나라의 →NP_MOD 건국부터
 - 건국부터 →NP_AJT 치세기까지의
 - 왕망의 →NP_MOD 치세기까지의
 - 치세기까지의 →NP_MOD 기간이
 - (이하 생략)

5.5.4. 장형 사동 구문 유형 세부 가이드라인

- “~게 하다”의 장형 사동 구문은 하나의 단위로 분석한다.
- 예문은 아래와 같다.

(가) 그는 철수를 집에 가게 하였다.

- 그는 →NP_SBJ 가게
- 철수를 →NP_OBJ 가게
- 집에 →NP_AJT 가게
- 가게 →VP 하였다.
- 하였다. →VP ROOT

(나) 그를 죽게 하였다.

- 그를 →NP_OBJ 죽게
- 죽게 →VP 하였다.
- 하였다. →VP ROOT



표준 작성 공헌자

표준 번호 : TTAK.KO-10.0853

이 표준의 제정·개정 및 발간을 위해 아래와 같이 여러분들이 공헌하였습니다.

구분	성명	위원회 및 직위	연락처 (E-mail 등)	소속사
표준(과제) 제안	최미란	PG606 위원	miranc@etri.re.kr	ETRI
표준 초안 작성자	최미란	PG606 위원	miranc@etri.re.kr	ETRI
	임준호	선임연구원	joonho.lim@etri.re.kr	ETRI
	임수종	책임연구원	isj@etri.re.kr	ETRI
	이충희	선임연구원	forever@etri.re.kr	ETRI
	김윤정	연구원	jungi0006 @mail.ulsan.ac.kr	울산대학교
표준 초안 에디터	이창한	PG606 위원	leech@kdb.or.kr	한국데이터베이스진흥원
표준 초안 검토	서태설	메타데이터 프로젝트그룹 의장	tsseo@kisti.re.kr	KISTI
		외 프로젝트그룹 위원		
표준안 심의	박승민	소프트웨어/콘텐츠 기술위원회 의장	minpark@etri.re.kr	ETRI
		외 기술위원회 위원		
사무국 담당	김영화	-	ykim@tta.or.kr	TTA
	이상호	-	sangho.lee@tta.or.kr	TTA



정보통신단체표준(국문표준)

의존 구문 분석 말뭉치 구축을 위한

의존 관계 태그 세트 및 의존 관계 설정 방법

(Dependency Tag Sets and Dependency Relation Establishment
Methods for Constructing Dependency Tagged Corpora)

발행인 : 한국정보통신기술협회 회장

발행처 : 한국정보통신기술협회

463-824, 경기도 성남시 분당구 분당로 47

Tel : 031-724-0114, Fax : 031-724-0109

발행일 : 2015.12.
