국립국어원 2020-01-26

발 간 등 록 번 호
11-1371028-000840-01

구문 및 무형 대용어 복원 말뭉치 연구 분석

사업 책임자
곽 용 진

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘구문 및 무형 대용어 복원 말뭉치 연구 분석’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2020. 5. 27. - 2021. 1. 1.

2020년 12월 21일

사업 책임자: 곽용진((주)이르테크)

사업 수행자	주식회사 이르테크 연세대학교 산학협력단 충남대학교 산학협력단
사업 책임자	곽용진((주)이르테크)
사업 참여자	김한샘, 유현경, 봉미경, 이숙의, 김진수, 김재훈, 이공주, 김유섭, 김학수, 나승훈, 류법모, 정해영, 이지연, 박재은, 최지선, 장호림, 한문성, 이순미, 김상선, 김영환, 서보원, 이선덕, 이지나, 임수용, 홍은기, 김윤희, 명노현, 민경배, 권민수, 고준석, 임조은, 강혜린, 김교연, 김도경, 김보은, 김상민, 김은솔, 김지영, 남궁영, 민진우, 박서윤, 박성식, 박혜진, 신아영, 오태환, 윤나영, 윤호, 이다인, 이수빈, 이종현, 이찬영, 장연지, 정영석, 정주연, 정해윤, 천성호, 최민석, 최용석, 이정은, 장지현, 정민경, 김다은, 김지연, 김연주, 박지훈, 박채정, 신민선, 안세현, 안수빈, 오승혁, 오진우, 이수형, 이영서, 이예지, 임혜리, 조동민

<사업 수행자> 주식회사 이르테크 · 연세대학교 산학협력단 · 충남대학교 산학협력단

사업 책임자	곽용진((주)이르테크)
사업 참여자	정해영((주)이르테크)
	이지연((주)이르테크)
	박재은((주)이르테크)
	최지선((주)이르테크)
	장호림((주)이르테크)
	한문성((주)이르테크)
	이순미((주)이르테크)
	김상선((주)이르테크)
	김영환((주)이르테크)
	서보원((주)이르테크)
	이선덕((주)이르테크)
	이지나((주)이르테크)
	임수용((주)이르테크)
	홍은기((주)이르테크)
	김윤희((주)이르테크)
	명노현((주)이르테크)
	민경배((주)이르테크)
	권민수((주)이르테크)
	고준석((주)이르테크)
	임조은((주)이르테크)
	김한샘(연세대학교)
	유현경(연세대학교)
	봉미경(연세대학교)
	김재훈(한국해양대학교)
	이공주(충남대학교)
	김유섭(한림대학교)
	김학수(건국대학교)
	나승훈(전북대학교)
	류법모(부산외국어대학교)
	강혜린(연세대학교)

	김교연(연세대학교)
	김도경(한림대학교)
	김보은(강원대학교)
	김상민(연세대학교)
	김은솔(연세대학교)
	김지영(연세대학교)
	남궁영(한국해양대학교)
	민진우(전북대학교)
	박서윤(연세대학교)
	박성식(강원대학교)
	박혜진(연세대학교)
	신아영(연세대학교)
	오태환(연세대학교)
	윤나영(연세대학교)
	윤호(한국해양대학교)
	이다인(한림대학교)
	이수빈(연세대학교)
	이종현(전북대학교)
	이찬영(연세대학교)
	장연지(연세대학교)
	정영석(한림대학교)
	정주연(연세대학교)
	정해윤(연세대학교)
	천성호(연세대학교)
	최민석(한국해양대학교)
	최용석(충남대학교)
	이숙의(충남대학교)
	김진수(충남대학교)
	이정은(충남대학교)
	장지현(충남대학교)
	정민경(충남대학교)
	김다은(충남대학교)
	김지연(충남대학교)
	김연주(충남대학교)
	박지훈(충남대학교)

	박채정(충남대학교)
	신민선(충남대학교)
	안세현(충남대학교)
	안수빈(충남대학교)
	오승혁(충남대학교)
	오진우(충남대학교)
	이수형(충남대학교)
	이영서(충남대학교)
	이예지(충남대학교)
	임혜리(충남대학교)
	조동민(충남대학교)

<국문 초록>

구문 및 무형 대용어 복원 말뭉치 연구 분석

본 연구는 구문 분석 말뭉치 및 무형 대용어 복원 말뭉치를 구축하는 데 목적이 있다. 2019년 '구문 분석 말뭉치 구축' 및 '주격 무형 대용어 복원 말뭉치 구축'의 후속 사업으로 작년도 사업에서 다루지 않은 구어 구문 분석 말뭉치 연구 분석과 목적격 무형 대용어 복원 말뭉치 연구 분석에 중점을 두었다. 이에 따른 주요 과업과 연구의 성과는 다음과 같다.

○ **구어 구문 분석 말뭉치 지침 수립**

'2019년 국립국어원 구문 분석 말뭉치 지침'에 구어 지침을 추가하여 지침을 보완하였다. 한국정보통신기술협회(TTA)의 지침에 준하되 구어 말뭉치 특성에 맞추어 세부 분석 지침을 정비하고 전체 지침의 완성도를 제고하였다.

○ **구어 구문 분석 말뭉치 구축**

'2019년 국립국어원 구어 형태 분석 말뭉치' 약 100만 어절을 대상으로 구어 구문 분석을 진행하였다. 그 결과 총 1,006,448어절 규모의 구어 구문 분석 말뭉치를 구축하였다.

○ **목적격 무형 대용어 복원 말뭉치 지침 수립**

'2019년 국립국어원 주격 무형 대용어 복원 말뭉치 지침'과의 일관성을 유지하여 목적격 무형 대용어 복원 말뭉치 지침을 수립하였다. 문어 말뭉치와 구어 말뭉치에 대한 복원 지침을 두루 기술하였으며, 한국전자통신연구원(ETRI)의 기준에 준하였다.

○ **목적격 무형 대용어 복원 말뭉치 구축**

'2019년 국립국어원 주격 무형 대용어 복원 말뭉치' 약 300만 어절(문어 200만 어절, 구어 100만 어절)을 대상으로 목적격 무형 대용어 복원 분석을 진행하였다. 그 결과 문어 2,000,213어절, 구어 1,006,448어절, 총 3,006,661어절 규모의 목적격 무형 대용어 복원 말뭉치를 구축하였다.

○ **구문 분석 및 무형 대용어 말뭉치 연구 분석 검증**

납품 자료 전체에 대한 일관성과 정확성 검증 방안 및 검증 체계를 수립하였다. 오류율을 최소화하기 위하여 구축 말뭉치를 3차에 걸쳐 납품하고, 주관 기관의 피드백을 반영하여 품질을 향상시켰다. 또한 작업자-검수자 간 정답 세트 표본을 마련하여 주기적으로 오류 검증을 하였다. 작업자의 작업 결과는 보고서 형식으로 전달되었으며, 점수가 낮은 작업자에게는 개별 교육을 실시하여 말뭉치의 품질을 보증하였다.

인공 지능 산업 발전을 위한 대규모 고품질 우리말 자원 수요가 증대되고 있다. ‘구문 및 무형 대용어 복원 말뭉치 연구 분석’은 이에 대한 이해를 바탕으로 국어 자원의 활용도와 가치 제고를 위해 수행되었다. 본 연구의 결과물은 국내 표준화 작업에 기여하고 참조 기반 자료가 될 수 있는 고품질 언어 정보 부가 말뭉치로서 의의를 갖는다.

주요어 : 구문 분석, 구어 구문 분석, 무형 대용어 복원, 목적격 무형 대용어 복원, 말뭉치

차 례

제1장 사업 개요

제2장 사업 수행 내용

차 례

제3장 결론

차 례

<표 차례>

차 례

<그림 차례>

차 례

<수식 차례>

제 1 장

사업 개요

1. 사업 목적

본 사업은 구문 분석 말뭉치 및 무형 대용어 복원 말뭉치 구축을 목적으로 한다. 국립국어원은 2019년 구문 분석과 주격 무형 대용어 복원 말뭉치 구축 사업을 독립적으로 진행하였다. 2020년에는 관련 층위를 통합함으로써 분석의 효율성 및 일관성을 높이고자 하였다.

구문 분석 말뭉치의 경우, 구어 약 100만 어절을 대상으로 의존 구문 분석을 진행하였다. 이를 위해 기존의 구문 분석 지침에 구어 지침을 추가하여 지침을 정교화하였다. 무형 대용어 복원 말뭉치의 경우, 문·구어 통합 약 300만 어절을 대상으로 목적격 무형 대용어 복원 분석을 시행하였다. 주격 무형 대용어 복원과 일관성을 유지하되 목적격 복원 방식에 따라 별도의 지침을 수립하였다. 사업의 목적을 요약하면 아래와 같다.

○ 구문 분석 말뭉치 연구 분석
- 구어 구문 분석 말뭉치 구축
- 구어 말뭉치 특성에 맞추어 세부 구축 지침 정비

○ 무형 대용어 복원 말뭉치 연구 분석
- 목적격 무형 대용어 복원 말뭉치 구축
- 목적격 복원 방식에 따른 지침 수립

4차 산업 혁명 대비 인공 지능 산업 발전을 위한 대규모 우리말 자원 수요가 증대되고 있다. ‘구문 및 무형 대용어 복원 말뭉치 연구 분석’은 이에 대한 이해를 바탕으로 국어 자원의 활용도와 가치 제고를 위해 수행되었다. 본 사업의 결과물은 국내 표준화 작업에 기여하고 참조 기반 자료가 될 수 있는 고품질 언어 정보 부가 말뭉치로서 의의가 있다. 또한 민간에서 자유롭게 활용할 수 있는 국가 공공재로서 그 활용도와 가치가 매우 높다.

2. 사업 범위

본 사업은 말뭉치 구축을 위한 지침 수립, 대규모 말뭉치 구축, 구축 후 품질 검증의 단계로 진행되었다.

ㅇ 말뭉치 구축 지침 수립

- 한국정보통신기술협회(TTA), 한국전자통신연구원(ETRI) 등 관련 분야의 분석 지침을 검토하고, 내용과 예시를 보완하여 해당 말뭉치 구축에 최적화된 지침을 수립하였다.
- 구문 분석의 경우, 구어 지침을 추가하여 세부 구축 지침을 정비하고, 전체 지침의 완성도를 제고하였다.
- 무형 대용어 복원의 경우, 목적격 무형 대용어 복원 지침을 별도로 수립하였다. 문어 말뭉치와 구어 말뭉치에 대한 복원 지침을 두루 기술하였으며, 주격 무형 대용어 복원 지침과의 일관성을 유지하였다.

ㅇ 지침에 따른 말뭉치 가공

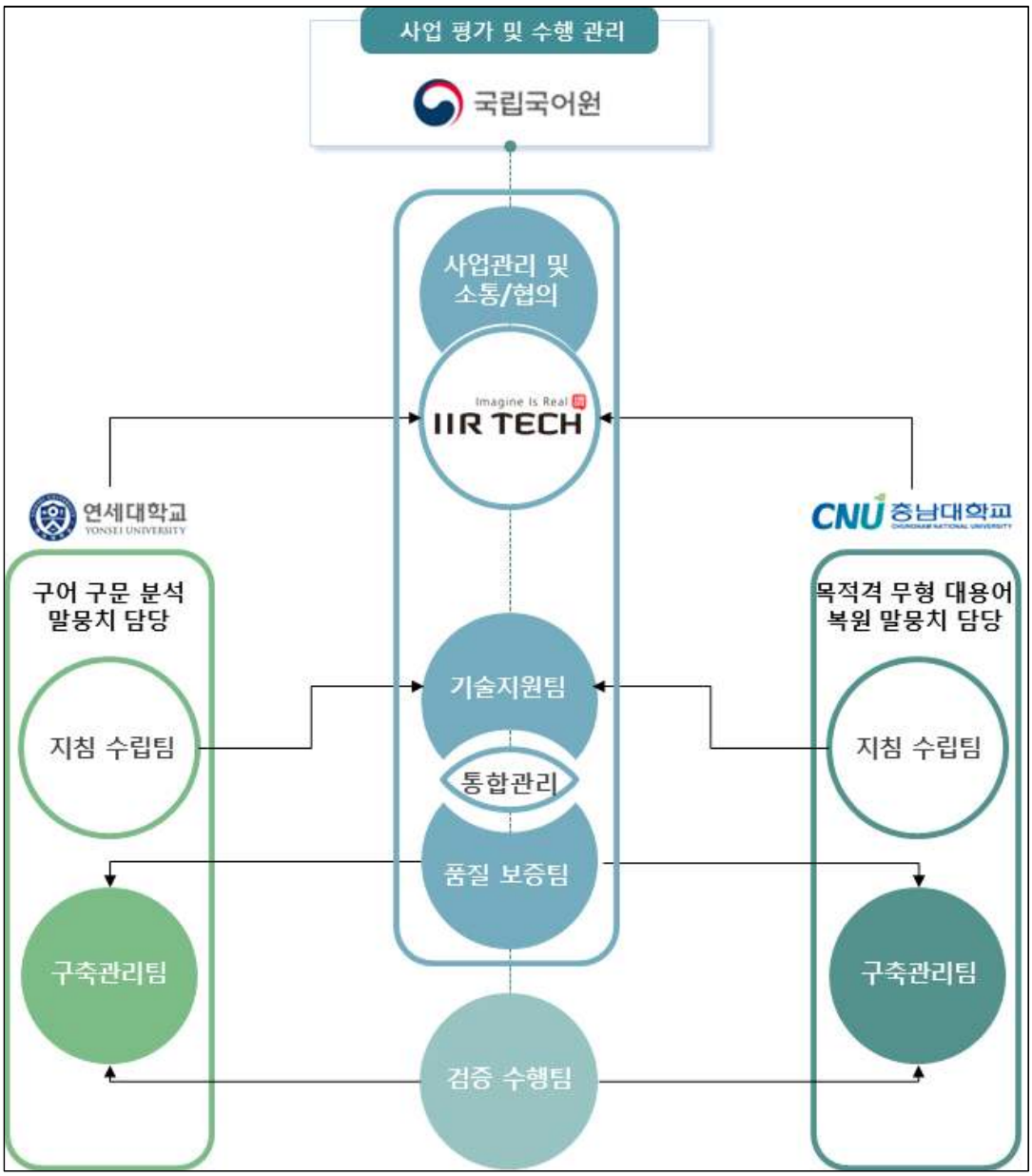
- 국내 표준화 작업에 기여하고 참조 기반 자료가 될 수 있는 정밀한 언어 정보 말뭉치를 구축하였다.
- 구문 분석은 ‘2019년 국립국어원 구어 형태 분석 말뭉치’ 약 100만 어절을 대상으로 진행되었다.
- 목적격 무형 대용어 복원 분석은 ‘2019년 국립국어원 주격 무형 대용어 복원 말뭉치’ 약 300만 어절(문어 200만 어절, 구어 100만 어절)을 대상으로 진행되었다.

ㅇ 말뭉치의 단계적 품질 검증

- 납품 자료 전체에 대한 일관성과 정확성 검증 방안 및 검증 체계를 수립하였다.
- 구축 기관 검증, 관리 기관 검증, 주관 기관 검증을 통해 단계적으로 품질을 점검하고 최종 결과물에 반영하였다.
- 오류율을 최소화하고자 구축 말뭉치를 3차에 걸쳐 납품하고, 주관 기관의 피드백을 반영하여 품질을 향상시켰다.
- 주기적 검증을 기반으로 한 통계적 검증을 실시하여 말뭉치의 품질을 보증하였다.

3. 사업 수행

본 사업은 교수진 11명을 포함하여 박사 및 석사 전공자 등 분야별 전문가의 협업으로 이루어졌다. 특히 2019년 문어 말뭉치 구문 분석, 주격 무형 대용어 복원 말뭉치 분석 수행 조직을 중심으로 인력을 구성하여 이전 구축 과정의 문제점과 비효율적인 수행 과정 등을 분석, 보완하고 전년도 말뭉치 구축 경험으로 효과적인 수행 체계를 구성하였다.

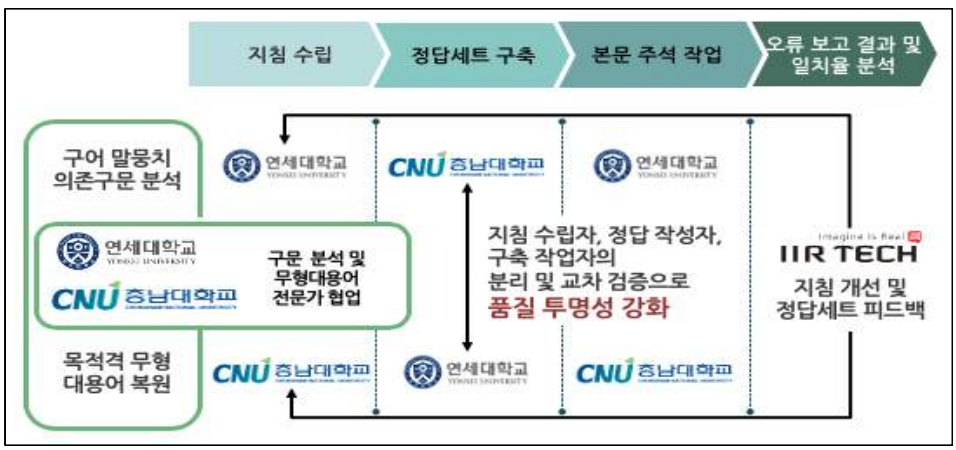


<그림 1> 이르테크 컨소시엄 사업 체계도

<그림 2> 사업 참여 인력 및 업무 분담

말뭉치 구축 시, 구문 분석 말뭉치를 먼저 구축하고, 어절 간 의존 관계를 바탕으로 목적격 무형 대용어를 복원함으로써 분석 층위 간 야기될 수 있는 해석 불일치도를 최소화하였다. 또한 구문 분석 말뭉치 구축 지침과 무형 대용어 복원 말뭉치 구축 지침을 각 구축 기관(연세대학교 산학협력단, 충남대학교 산학협력단) 및 관리 기관((주)이르테크), 주관 기관(국립국어원)에서 종합적으로 교차 검수하여 기본 원칙에 맞추어 통일성을 확보하였다.

본 사업은 이러한 각 분야 선행 사업자 간 협업으로 통합 말뭉치 구축에 필요한 일관성을 확보하고, 말뭉치의 품질 향상에 이바지하였다.



<그림 3> 협업 체계

4. 사업 결과

본 사업의 결과는 다음과 같다.

- 분석 층위별 지침 수립 및 정교화
- 구어 구문 분석 말뭉치 구축(약 100만 어절)
- 목적격 무형 대용어 복원 말뭉치 구축(약 300만 어절)

○ 말뭉치 주석 결과

- 구어 구문 분석 말뭉치

<표 1> 구어 구문 분석 말뭉치 주석 결과

총 어절 수	총 문장 수	구문 태그 수	기능 태그 수	지배소를 갖는 어절 수	의존소를 갖는 어절 수
1,006,448	221,489	1,006,448	404,047	1,006,448	531,136

(지배소 값에는 ROOT값 포함)

- 목적격 무형 대용어 복원 분석 말뭉치

<표 2> 목적격 무형 대용어 복원 분석 말뭉치 주석 결과

	총 어절 수	총 문장 수	복원 대상 서술어 수	선행어 수
문어	2,000,213	150,082	40,446	40,446
구어	1,006,448	221,489	43,149	43,149
총합	**3,006,661**	**371,571**	**83,595**	**83,595**

제 2 장

사업 수행 내용

1. 사업 추진 일정

사업 추진 경과는 다음과 같다.

<표 3> 사업 추진 일정

작업 기간 / 작업 내용		6월	7월	8월	9월	10월	11월	12월
구축 상세 사항 협의		O	O	O	O	O	O	O
지침 수립 및 개정		O	O	O	O	O	O	O
작업자 교육			O	O	O	O	O	O
구축 환경 지원		O	O	O	O	O	O	O
말뭉치 구축 및 정제			△	O	O	O	O	O
결과물 납품	자동 구문 분석 결과		O					
	구어 구문 분석 말뭉치				O	O		O
	목적격 무형 대용어 복원 말뭉치				O	O		O
내부 검증(정답 세트 검증)				O	O	O	O	O
외부 검증(국어원 시스템 사용 작업자 교육)			O	O				
사업 보고	착수 보고	O						
	월간 보고		O	O	O	O	O	O
	중간 보고					O		
	최종 보고							O

(△ : 정답 세트만 구축)

2. 수행 환경 구성

2.1. 원시 데이터

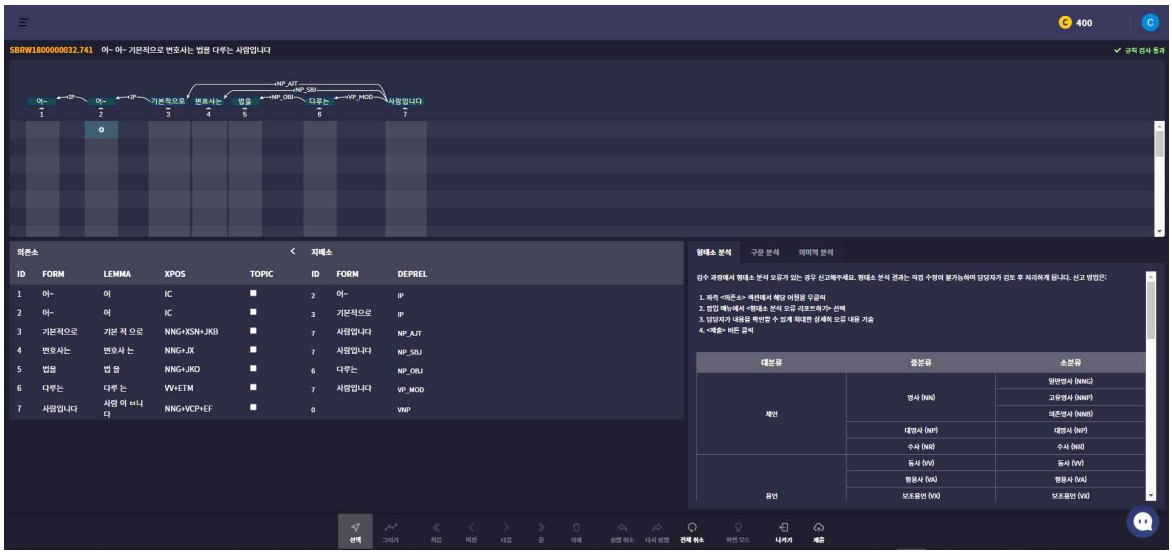
주관 기관인 국립국어원에서 제공한 원시 말뭉치 약 300만 어절(문어 2,000,213어절, 구어 1,006,448어절)을 대상으로 구문 및 무형 대용어 복원 말뭉치를 구축하였다. 구문 분석 말뭉치는 구어만 대상으로 하였으며, 목적격 무형 대용어 복원 말뭉치는 문어, 구

어를 모두 대상으로 하였다. 2019년에 구축된 문어 구문 분석 말뭉치, 구어 형태 분석 말뭉치, 주격 무형 대용어 복원 말뭉치, 상호 참조 해결 말뭉치를 국립국어원에서 제공받아 분석에 참고하였다.

2.2. 구축 도구

(1) 구어 구문 분석 말뭉치

구어 구문 분석 말뭉치 구축은 2019년 문어 구문 분석에 활용된 도구를 보완, 개선하여 진행하였다. 자동 주석 보팅(voting, 다수의 분석기가 예측한 값을 결과로 채택)에 따라 모듈을 개선하고 1차 주석 결과를 적재함으로써 작업자의 분석 시간을 단축하였다. 또한 검수자용 작업 도구를 개선하여 전체 작업 결과에 대한 검토가 가능하도록 하고, 정답 세트 구축을 위한 병행 프로젝트 운영도 가능하도록 하였다. 또한 납품 데이터 중간 생성 기능도 추가하여 사업 중반 단계별 납품 시에도 무결성을 검증한 데이터를 납품할 수 있게 하였다.



<그림 4> 구어 구문 분석 말뭉치 구축 도구 화면

해당 도구는 원시 문장이 자동 구문 분석된 CoNLL 형식의 구문 분석 결과를 파싱하여 제시한다.[1] 작업자들은 분석 내용을 확인한 뒤 오류를 수정하고, 검수자들이 결과를

1) 자동 구문 분석은 건국대학교, 전북대학교, 충남대학교가 보유, 개발한 3종의 분석기 결과를 비교하여 합치율에 따라 적재하였다. 구문 분석기에 관한 상세 설명은 2019년 '구문 분석 말뭉치 구축' 최종 보고서에 기술되어 있다.

재검수하여 내용을 최종 점검하였다. 유효성 검사를 실시하여 부정확한 분석 결과가 반영되지 않도록 사전에 방지하였으며, 어절 간 의존 관계를 그림으로 제시하여 작업자들의 직관적 판단이 가능하게 하였다. 또한 문의, 오류 신고 등 부가 기능을 활용해 효율적으로 의사소통하고, 진행 상황을 실시간으로 확인할 수 있도록 하였다.

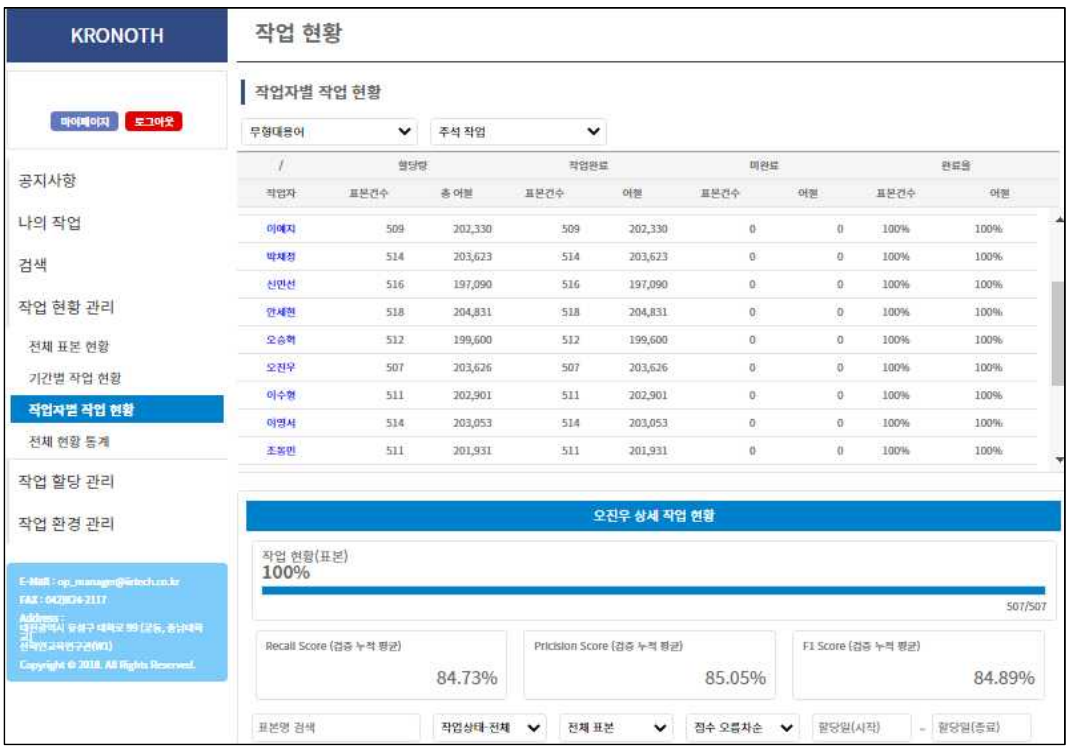
<그림 5> 구어 구문 분석 말뭉치 구축 현황 모니터링

(2) 목적격 무형 대용어 복원 말뭉치

목적격 무형 대용어 복원 말뭉치 구축은 관리 기관 ㈜이르테크가 개발한 주석 작업 도구(KRONOTH Annotation System)를 사용하여 진행되었다.

해당 도구로 말뭉치 구축과 제어 작업을 하였으며, 실시간으로 현황을 관리하였다. 목적격 복원 대상 서술어 후보군이 자동으로 제공되면 작업자가 이들을 바탕으로 서술어 추가 및 삭제 작업을 진행하였다. 보류 및 검토 요청 기능을 제공하여 자가 진단을 할 수 있게 하였으며, 작업 종료 전 미작업 내용을 팝업 창에 제시하여 오류를 최소화하는 방안을 마련하였다. 또한 인가자만이 시스템에 접근할 수 있게 설계하여 주관 기관의 보안 요구 사항을 준수하였다.

<그림 6> 목적격 무형 대용어 복원 분석 말뭉치 구축 도구 화면



<그림 7> 목적격 무형 대용어 복원 분석 말뭉치 작업 관리 및 품질 검증 관리

2.3. 작업자 교육

말뭉치 분석 작업에 대한 작업자의 이해도를 높이고, 말뭉치 구축 결과물의 품질 향상을 도모하기 위해 작업자 교육을 실시하였다. 작업자 교육은 주관 기관의 구축 시스템을 활용한 초기 교육과 구축 기관 내 정기 교육으로 구분된다.

(1) 주관 기관 교육

주관 기관인 국립국어원은 사업 초기 작업자들의 지침에 대한 이해도와 분석 일치도를 검사하였다. 구축 기관의 초기 지침 교육 후 모든 구축 인력(작업자와 검수자, 지침 담당자)이 일주일간 국립국어원의 말뭉치 구축 시스템에서 동일한 문서에 대해 주석 작업을 진행하였다. 주관 기관의 담당 연구원들과 분석 결과를 비교해 지침에 대한 개개인의 이해도를 평가하였다. 구축 말뭉치의 일관성 확보를 위해 구축 인력 내 일치도도 점검하였다. 국립국어원은 평가 결과를 작업자별로 우수, 양호, 미흡, 주의, 심각의 다섯 단계로 구분하여 통보하였고, 미흡 이하 작업자는 보충 교육을 받았다. 보충 교육을 실시한 후 동일한 과정으로 재검증하여 일정 수준 이상의 점수에 도달하면 교육을 해제하였다.

가. 구문 분석

- 교육 대상 인원 : 김한샘 외 18명

- 작업 범위 : 5,000어절/1인(1,000어절/1일)

- 교육 일정
  - 접속 계정 수령 및 테스트 : 2020년 7월 31일
  - 작업 및 제출 : 2020년 8월 6일 - 8월 12일
  - 작업 결과 확인 : 2020년 8월 13일 - 8월 14일
  - 재교육 및 인증 : 2020년 8월 17일 - 8월 21일

- 작업 후 조치 사항
  - 일정 점수 미만 작업자 재교육 또는 교체

- 작업자별 지침 숙지도 관리
- 불일치 유형 및 항목 정리
- 지침 보완 또는 불일치 빈발 사항 조치 방안 수립

- 보충 교육
  - 날짜 : 2020년 8월 25일 - 9월 4일
  - 작업자 재교육 : 미흡 이하 작업자 보충 교육 실시 후 ㈜이르테크 정답 세트 검증 결과 88% 이상 시 교육 해제

나. 무형 대용어 복원 분석

- 교육 대상 인원 : 이숙의 외 19명

- 작업 범위 : 10,000어절/1인(2,000어절/1일)

- 교육 일정
  - 접속 계정 수령 및 테스트 : 2020년 7월 29일
  - 작업 및 제출 : 2020년 7월 30일 - 8월 5일
  - 작업 결과 확인 : 2020년 8월 6일 - 8월 7일
  - 재교육 및 인증 : 2020년 8월 10일 - 8월 14일

- 작업 후 조치 사항
  - 일정 점수 미만 작업자 재교육 또는 교체
  - 작업자별 지침 숙지도 관리
  - 불일치 유형 및 항목 정리
  - 지침 보완 또는 불일치 빈발 사항 조치 방안 수립

- 보충 교육
  - 날짜 : 2020년 8월 11일 - 8월 19일
  - 재교육 방안
    1) 미흡 이하 작업자 보충 교육 실시 후 ㈜이르테크 정답 세트 검증 결과 85% 이상 시 교육 해제

2) 주관 기관과의 화상 회의 및 재교육 현황 인증

(2) 구축 기관 교육

말뭉치 구축 기관은 정기 교육을 실시하여 작업자의 지침 이해도를 높이고 작업 일관성을 확보하였다. 교육 내용은 지침 교육과 도구 사용 교육뿐 아니라 말뭉치 구축 제반에 관한 기본 내용을 포함하였다. 전공 교수와 박사 과정 참여자가 교육을 주도하여 작업의 전문성을 제고하였다. 정답 세트 검증 점수가 낮은 작업자를 대상으로 정기 교육 외에 수시 교육을 매주 실시하여 실시간 인력 관리를 지원하였다. 각 구축 기관의 교육 일시 및 내용은 아래와 같다.

가. 구문 분석

- 1차 교육
  - 날짜 : 2020년 7월 30일
  - 내용 : 작업 방식에 대한 전반적인 안내, 국어원 시스템 작업 방식 안내

- 2차 교육
  - 날짜 : 2020년 8월 18일
  - 내용 : 작업 방식에 대한 전반적인 안내, 국어원 시스템 작업 방식 안내

- 3차 교육
  - 날짜 : 2020년 9월 8일
  - 내용 : 수정된 지침 안내, 실제 작업 내용 공유 및 논의

- 4차 교육
  - 날짜 : 2020년 10월 26일
  - 내용 : 1-2차 프로젝트 빈발 오류 유형에 대한 추가 교육

- 5차 교육
  - 날짜 : 2020년 11월 23일
  - 내용 : 1차 프로젝트 정제 작업 방식 및 오류 유형 안내

나. 무형 대용어 복원 분석

- 1차 교육
  - 날짜 : 2020년 6월 23일
  - 내용 : 목적격 무형 대용어 복원 구축 작업을 위한 지침 및 도구 사용법 교육

- 2차 교육
  - 날짜 : 2020년 7월 10일
  - 내용 : 복원 대상 서술어에서 제외되는 보조 용언에 대한 교육

- 3차 교육
  - 날짜 : 2020년 7월 29일
  - 내용 : 국립국어원 검증 작업의 시스템 사용법 교육

- 4차 교육
  - 날짜 : 2020년 8월 4일
  - 내용 : 질의 사항에 대한 논의 및 답변

- 5차 교육
  - 날짜 : 2020년 9월 27일
  - 내용 : 문어 문서 전체 수정 전 전반적인 문어 지침 재교육

- 6차 교육
  - 날짜 : 2020년 10월 16일
  - 내용 : 구어 문서 작업 전 구어 지침 교육

- 7차 교육
  - 날짜 : 2020년 11월 30일
  - 내용 : 구어 문서 전체 수정 전 전반적인 구어 지침 재교육

※ 상시 교육 : 공유 문서를 통해 모든 작업자가 질의 사항을 공유 및 논의[2)]

2.4. 검증 정책

본 사업은 납품 결과물의 품질을 보장하고자 사업 수행 기간 동안 구축 기관 검증, 관리 기관 검증, 주관 기관 검증을 동시에 진행하였다. 또한 각 말뭉치의 층위별 교차 검증을 통해 분석의 일관성을 확보하였다.

(1) 구축 기관 검증: 정답 세트 활용

각 구축 기관은 전체 구축 어절 가운데 약 10%의 어절을 대상으로 작업자-검수자 간 정답 세트 문서를 마련하였다. 정답 세트 활용 검증은 2019년 주관 기관의 검증 방식으로 이미 신뢰성이 부여된 방식이다.

구축량의 일부를 정답 세트로 지정해 검수자들이 분석을 진행하였다. 작업자들의 주간 작업이 완료되면 관리 기관은 작업자와 검수자 간 비교 분석을 실시하였다. 분석 결과는 검증 보고서 형태로 구축 기관에 전달하였다. 검증 보고서는 작업자별 오류 상세 결과 및 검증 점수를 포함하였다. 각 구축 기관의 검수자는 점수가 낮은 작업자를 대상으로[3] 개별 교육을 실시하여 말뭉치의 품질을 상시 관리하고 일관성을 확보하였다.

2) 공유 문서에는 접속 제한을 두어 권한이 부여된 사용자만 문서에 접근할 수 있도록 하였다. 무형 대용어 복원 분석 작업자만 문서에 접근할 수 있도록 보안 환경을 유지하였다.

3) 구문 분석 말뭉치 분석의 경우 88점, 무형 대용어 복원 말뭉치 분석의 경우 85점 이하 작업자에게 재교육을 실시하였다.

```
12. 검증문장:::::SBRW1800000034.443 그러지 말고 딴 거를 했으면

{'word_id': 1, 'word_form': '그러지', 'head': 2, 'label': 'VP', 'dependent': []}
{'word_id': 2, 'word_form': '말고', 'head': 5, 'label': 'VP', 'dependent': [1]}
{'word_id': 3, 'word_form': '딴', 'head': 4, 'label': 'DP -> VP_MOD', 'dependent': []}
{'word_id': 4, 'word_form': '거를', 'head': 5, 'label': 'NP_OBJ', 'dependent': [3]}
{'word_id': 5, 'word_form': '했으면', 'head': 0, 'label': 'VP', 'dependent': [2, 4]}


13. 검증문장:::::SBRW1800000033.741 저~ 티비를 통해서도 이제 얘기는 하지만

{'word_id': 1, 'word_form': '저~', 'head': 2, 'label': 'IP', 'dependent': []}
{'word_id': 2, 'word_form': '티비를', 'head': 3, 'label': 'NP_OBJ', 'dependent': [1]}
{'word_id': 3, 'word_form': '통해서도', 'head': 6, 'label': 'VP', 'dependent': [2]}
{'word_id': 4, 'word_form': '이제', 'head': '6 -> 5', 'label': 'AP -> IP', 'dependent': []}
{'word_id': 5, 'word_form': '얘기는', 'head': 6, 'label': 'NP_OBJ', 'dependent': '[] -> [4]'}
{'word_id': 6, 'word_form': '하지만', 'head': 0, 'label': 'VP', 'dependent': '[3,4,5] -> [3,5]'}

<<<<<<<<<<정답 스코어>>>>>>>>>>
recall : 81.57894736842105
precision : 81.57894736842105
F1 Score : 81.57894736842105
```

<그림 8> 구어 구문 분석 말뭉치의 정답 세트 검증 보고서 예시

```
파일명:::::::NWRW1800000024.212.json


----------Predicate Miss----------

1.
predicate  : NWRW1800000024.212.6.2 - 아예 한글을 모르면 기역부터 [시작한다].
antecedent : NWRW1800000024.212.6.1 - [수업]은 한글 실력에 따라 2개 반으로 나뉜다.
2.
predicate  : NWRW1800000024.212.10.3 - 김 할머니는 애초 문중 소유의 초가집에 살았으나 [지은] 지 100년이 넘어 붕괴 위험이
                                       있고 비가 샜다.
antecedent : NWRW1800000024.212.10.3 - 김 할머니는 애초 문중 소유의 [초가집]에 살았으나 지은 지 100년이 넘어 붕괴 위험이
                                       있고 비가 샜다.


----------Antecedent Diff----------

1.
predicate  : NWRW1800000024.212.9.1 - 공 회장은 "빨리 [익히도록] 일부러 숙제를 많이 내주는데도 할머니들이 거르는 일이
                                      거의 없다"며 "2시간30분동안 계속되는 수업시간이 힘들기도 하실 텐데 화장실도 가지
                                      않고 자리를 지키시는 모습을 보면서 되레 배우는 것이 많다"고 밝혔다.
antecedent : NWRW1800000024.212.7.1 - 할머니들은 [한글]만큼 어려운 수학도 배운다.
error_ant  : -1 - 한글

----------Score----------
recall : 93.75
precision : 75.0
F1 Score : 83.33333333333333
```

<그림 9> 목적격 무형 대용어 복원 분석 말뭉치의 정답 세트 검증 보고서 예시

(2) 관리 기관 검증

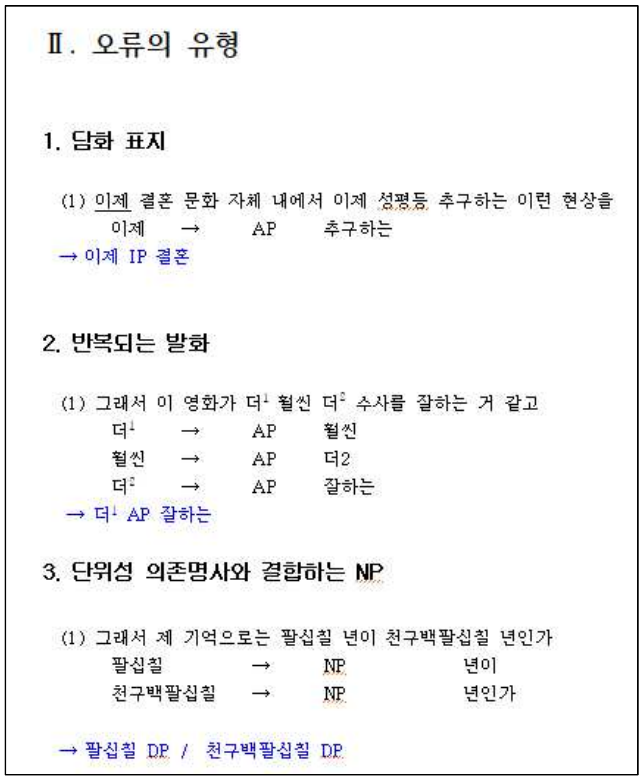
관리 기관인 ㈜이르테크는 구어 구문 분석 말뭉치 및 목적격 무형 대용어 복원 말뭉치 구축 결과물을 수시로 모니터링하여 오류 내용을 구축 기관에 전달하였다. 이를 통해 동일 오류가 반복 생성되는 것을 방지하고 말뭉치 품질을 개선하고자 하였다.

가. 분석 층위별 검수

관리 기관은 말뭉치 구축이 일정 기간 이루어진 이후, 구문 분석 및 목적격 무형 대용어 복원 말뭉치 주석 결과를 자체적으로 검수하였다. 작업이 완료된 문서 가운데 샘플 문서를 뽑아 주석 오류가 있는지 검수하였으며, 동일 오류 재발을 방지하고자 주석 오류를 유형화하여 구축 기관에 전달하였다. 특히 정답 세트 검증 점수가 낮은 작업자의 문서를 위주로 검수하였으며, 지침이 개정될 때마다 최신 지침에 맞게 작업이 이루어지는지 확인하였다.

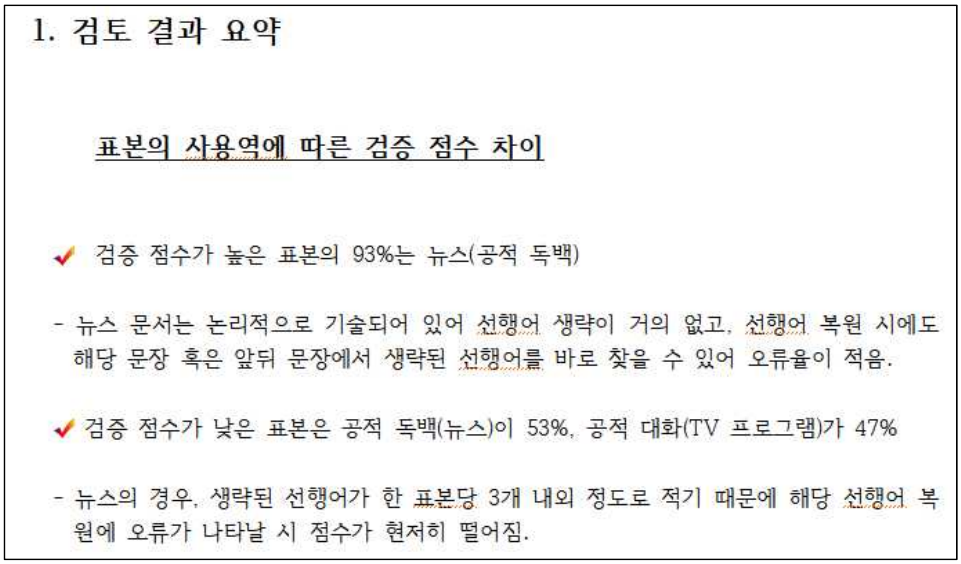
문장	작업자 작업 결과	정답	오류 유형
미용실 좀 다녀본 분들은 다 이제 얘기를 들어봤을 법한 건데	이제 AP 들어봤을	이제 IP 얘기를	구문태그, 의존관계
아우 저 머리 쪼금만 다듬어 주세요. 예.	쪼금만 NP_AJT 다듬어	쪼금만 AP 다듬어	구문태그, 기능태그
왜 미용실만 가면 머리가 그렇게 개털이에요?	미용실만 NP_OBJ 가면	미용실만 NP_AJT 가면	기능태그
이거는 뭐 컷트로 한다고 해서 될 게 아니라	이거는 NP_SBJ 아니라	이거는 NP_SBJ 될	의존관계
아니 그니까 파마 안 하고 컷트만 하고 가면	그니까 AP 하고	그니까 AP 가면	의존관계
꼭 해야지 머리를 잘라도 느낌이 좀 날 거 같애요.	해야지 VP 잘라도	해야지 VP 날	의존관계
왜냐면 파마를 안 하실 거니까.	왜냐면 AP 하실	왜냐면 AP 거니까.	의존관계
파마는 지금 나 못 시켜주지.	나 NP_OBJ 시켜주지	나 NP_SBJ 시켜주지	기능태그
근데 지금 할 수 있는 상황이 아니래니까?	지금 AP 할	지금 AP 아니래니까?	의존관계
트리트먼트는 뭐 얼마 하지도 않으니까	트리트먼트는 NP_OBJ 하지도	트리트먼트는 NP_SBJ 하지도	기능태그
트리트먼트만 받고 가요 오늘은.	오늘은 NP_AJT	오늘은 AP	구문태그, 기능태그

<그림 10> 구어 구문 분석 오류 내용 예시



<그림 11> 구어 구문 분석 오류 유형 예시

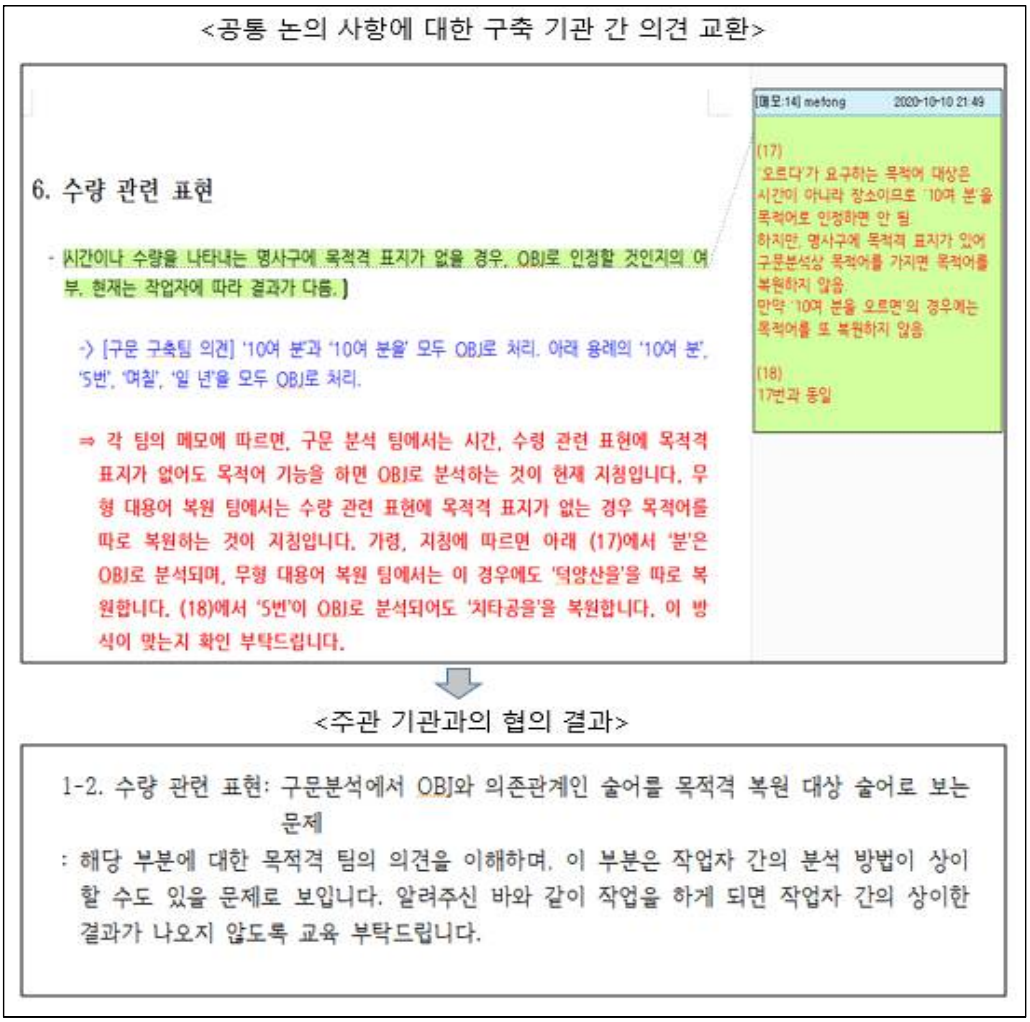
작업자의 정답 세트 검증 점수가 큰 폭으로 변화한 경우에는 해당 원인을 찾아 분석하였다. 가령, 구어를 대상으로 한 목적격 무형 대용어 복원 말뭉치에서 검증 점수가 큰 폭으로 변화한 경우가 있었는데, 확인 결과 생략된 목적어의 개수가 3개 이하로 적은 공적 독백 문서에서 해당 선행어 결정에 오류가 있을 시, 검증 점수가 큰 폭으로 하락했던 것으로 나타났다.



<그림 12> 검수 결과 분석의 예시

나. 공통 논의 사항 검수

구어 구문 분석 및 목적격 무형 대용어 복원 말뭉치 구축 과정에서 분석 층위 간 해석 불일치도를 최소화하고자 지침을 교차로 검수하였지만, 실제 구축 과정에서 주석 결과에 대해 구축 기관 간 확인이 필요한 경우가 있다. 따라서 공통 논의가 필요한 부분은 해당 사항을 정리하여 구축 기관에 각각 전달하였다. 구축 기관은 서로의 의견을 공유하고, 이에 대한 해결 방안을 모색한 뒤 주관 기관과 협의하여 결론을 도출하였다.



<그림 13> 공통 논의 사항에 대한 협의 과정 예시

(3) 주관 기관 검증

가. 지침 내용 검증

주관 기관은 구축 지침이 개정될 때마다 내용을 확인하고, 구축 기관에 피드백을 송부하였다. 구축 기관은 지침 내용에 대해 주관 기관과 상시 소통하였으며, 해당 내용을 정리하여 지침에 최종 반영하였다. 그 결과, 구문 분석 말뭉치 구축 지침은 0.1.8버전으로, 목적격 무형 대용어 구축 지침은 0.3.0버전으로 지침이 완성되었다.

나. 주석 작업 내용 검증

주관 기관은 납품된 말뭉치를 대상으로 국립국어원의 담당 연구원들이 구축한 정답 세트와 비교 검증을 실시하였다. 이 결과를 토대로 오류 유형을 분석하여 각 구축 기관에 전달하였다. 구축 기관은 주관 기관의 피드백을 바탕으로 해당 오류가 재발하지 않도록 작업자 교육 및 주석 작업을 진행하였다.

(4) 말뭉치 층위 간 검증

무형 대용어 복원 분석은 구문 분석 결과에 의존하며, 분석 결과는 상호 참조 해결 분석과 연관된다. 이에 본 사업에서는 원시 말뭉치가 동일한 구문 분석 말뭉치와 상호 참조 해결 말뭉치를 활용하여 무형 대용어 복원 분석 결과를 검증하였다.

가. 구문 분석 말뭉치와 무형 대용어 복원 말뭉치 간 검증

무형 대용어 복원 분석 시 구문 분석 결과를 기준으로 생략어 복원 대상 서술어 여부를 결정하였다. 이에 검증에서도 구문 분석 말뭉치와 무형 대용어 복원 분석 말뭉치의 결과를 비교하였다. 구문 분석상 타동사가 지배하는 목적어가 없을 때 생략어 복원을 실시하였는지, 타동사가 지배하는 목적어가 있는데 생략어를 복원하지는 않았는지 확인하였다. 이를 위해 구문 분석 말뭉치를 먼저 구축하고 해당 구축 결과를 참고하여 목적격 무형 대용어 복원 말뭉치를 구축하였다.

문어 말뭉치 무형 대용어 복원 작업은 전년도에 구축한 구문 분석 말뭉치를 바탕으로 실시하였다.

(가) 하찮은 재료를 긁어모아 엮을 때 얼마나 아름다운 꽃이 **피어나는지** 관객들과 함께 **확인하고** 싶어요.

(나) 노동자의 삶을 담기 위해 2년에 걸쳐 **5번을 방문했다**.

예를 들어 (가)에서 '확인하고'는 구문 분석 말뭉치에서 '피어나는지'를 VP_OBJ로 분석한다. 따라서 '확인하고'는 목적격 복원 대상 술어로 보지 않는다. (나)의 '방문했다'도 구문 분석 말뭉치에서 '5번을'이 NP_OBJ로 분석되므로 목적격 복원 대상이 아니다.

이와 같이 어절 간 의존 관계를 고려하여 목적격 무형 대용어를 복원함으로써 분석 층위 간 일관성을 확보하였으며, 작업 지침을 분석 층위 간에 교차 검토하여 구축에 관한 기본 원칙을 확인하였다.

나. 무형 대용어 복원 말뭉치와 상호 참조 해결 말뭉치 간 검증

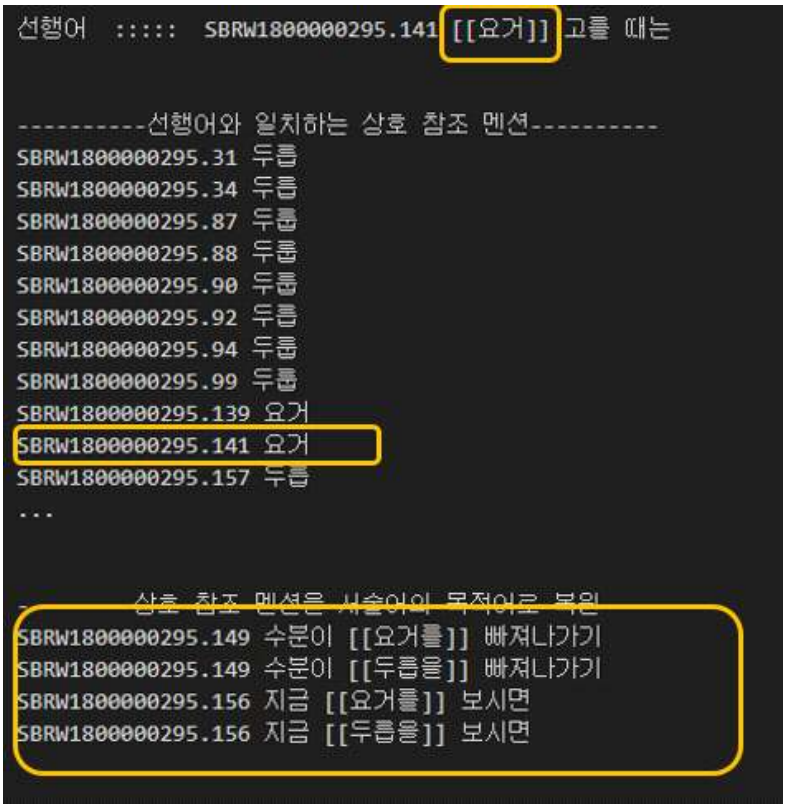
무형 대용어 복원 말뭉치와 상호 참조 해결 말뭉치는 대용어의 개념을 공유하고 있다. 무형 대용어 복원은 생략된 목적어의 선행어를 찾아 복원하고, 상호 참조 해결은 동일 문서 내에서 서로를 대신할 수 있는 개체명을 찾아 동일 집합으로 연결한다.

상호 참조 해결 말뭉치의 분석 결과를 활용하여 무형 대용어 복원 말뭉치 분석에서 선행어 분석이 지침에 맞게 이루어졌는지, 적합한 의미의 선행어로 생략어가 복원되었는지 등을 검증할 수 있다. 본 사업의 무형 대용어 복원 분석에서는 여러 개의 선행어 후보 중 서술어와 가장 가까이 위치한 전방 선행어를 우선시한다. 그러나 동일 문장 내에 선행어가 존재하지 않아 문장 밖에서 선행어를 탐색할 경우, 작업자는 서술어에서 가장 가까이 위치한 선행어를 제대로 탐지하지 못할 가능성이 있다.

1	이제 절단이
2	되어 있는 상태에서부터
3	수분이 <1211:요거를> 빠져나가기
4	시작하면서
5	향기가 날라갑니다

<그림 14> 목적격 무형 대용어 복원 분석 말뭉치의 선행어 복원 예시

<그림 14>의 예시에서 ‘빠져나가기’의 생략된 목적어의 선행어는 문장 내에 없고, 선행어 후보로 다른 문장의 ‘요거’와 ‘두릅’, ‘두룹’을 생각할 수 있다. 이때 분석 지침에 따라 ‘빠져나가기’와 가장 가까이 위치한 ‘요거’로 선행어를 선택한다.



<그림 15> 목적격 무형 대용어 복원 분석 말뭉치와 상호 참조 해결 말뭉치 간 검증 예시

선행어와 일치하는 상호 참조 멘션 목록에서 ‘수분이 빠져나가기’와 가장 가까이 위치한 멘션이 ‘요거’임을 확인할 수 있다. 이에 해당 문장은 지침에 맞게 선행어를 분석하였다고 판단된다. 따라서 선행어 복원은 지침에 맞게 이루어졌다. 또한 선택된 선행어가 의미적으로 적합한지도 확인할 수 있다. ‘요거’와 동일한 상호 참조 멘션 집합으로 묶인 ‘두릅’으로 해당 문장을 교체 출력해도 그 의미가 자연스럽게 해석되므로 적절한 선행어로 복원되었다고 판단된다.[4)]

4) 상호 참조 해결 말뭉치는 2019년에 구축되어 2020년에 유지 보수가 이루어졌다. 상호 참조 해결 말뭉치의 유지 보수 기간과 목적격 무형 대용어 복원 말뭉치 구축 기간이 맞물려, 해당 층위의 검증은 작업 초기 검증, 교육 기간에 한해서만 활용되었다.

3. 지침

3.1. 구축 지침 정비

본 사업은 2019년 ‘구문 분석 말뭉치 구축’ 및 ‘주격 무형 대용어 복원 말뭉치 구축’의 후속 사업으로 전년도 지침과의 일관성을 유지하되 새롭게 구축하는 말뭉치의 특성에 맞도록 지침을 개정·보완하였다.

구어 구문 분석의 경우, 구어의 특성에 맞추어 2019년 구문 분석 말뭉치 지침의 세부 분석 지침을 정비하였다. 목적격 무형 대용어 복원 분석의 경우, 2019년 주격 무형 대용어 복원 말뭉치 지침과의 일관성을 유지하여 목적격 복원 분석에 관한 지침을 새롭게 수립하였다.[5] 두 분석 층위 모두 기본적으로 한국정보통신기술협회(TTA) 및 한국전자통신연구원(ETRI)의 기준을 준수하였다.

지침 개정 시 구축 기관, 관리 기관, 주관 기관이 모두 지침을 공유하여 교차 확인하였으며, 특히 세부 분석 지침을 새로 정할 때에는 사례를 기반으로 주관 기관과 협의하여 최종 결정하였다.

본 사업에서 수립한 구문 분석 지침 및 목적격 무형 대용어 복원 분석 지침을 3.2.와 3.3.에 각각 요약 제시하였다.[6]

3.2. 구문 분석 지침[7]

1. 기본 원칙

(1) 자연 언어 처리를 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 분석

5) 목적격 무형 대용어 복원 분석 시 구문 분석 결과를 기준으로 생략어 복원 대상 서술어 여부를 결정하지만, 이 경우 주격 무형 대용어 복원 지침과의 일관성이 유지되어야 한다는 전제가 우선한다.
6) 지침에서 용례를 제시할 때 말뭉치상에서 나타난 띄어쓰기 오류 등의 정서법 오류는 수정하지 않고 그대로 제시하였다.
7) 한국정보통신기술협회(TTA) 표준 지침인 ‘의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법’의 표현 및 예문을 수정하거나 세부 지침을 추가하는 방식으로 지침을 기술하였다.

한다.

(5) 각 어절은 1개의 지배소를 가진다.(Single-Head Constraint)

(6) 각 어절 및 지배소 쌍은 서로 교차하지 않는다.(Projective Constraint)

(7) 보어와 부가어를 구분하되 보어의 범위를 엄격히 제한한다.

- 보어 CMP는 보격 조사가 부착된 NP, 용언구, 절, 그리고 인용절 보문의 용언구와 절에 한해서 분석한다.
- 조사가 생략되거나 보조사가 부착된 명사구 또는 이에 상응하는 용언구와 절도 서술어 구문 틀에 따라 보격 조사로 대치 가능하면 CMP로 분석한다.

(가) 그가 돌아왔다고(VP_CMP) 그녀가 알려줬어.(TTA, 9쪽)

(나) 그녀가 그 일을 했다고(VP_CMP) 스스로 말했다.(TTA, 10쪽)

(다) 비평가 칼라일이 “인도와도 바꿀 수 없다.”고(VP_CMP) 말하였다.(TTA, 13쪽)

(라) 물이 얼음이(NP_CMP) 되었다.

(마) 철수가 발이 아프다고(VP) 훈련을 빠졌다.(VP)

(바) 마법사가 와인을 물로(NP_AJT) 바꾸었다.

(사) 철수가 영희가(NP_SBJ) 보고 싶다.

(8) 원칙적으로 접속과 내포를 구별하지 않으며, 접속절은 모두 부사절로 분석한다.(다만, 명사구 접속은 인정함)

(9) 하나의 성분이 모문과 내포문 모두에 관련되어 있으면 내포절의 유형에 따라 해당 주어의 지배소를 결정한다.

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)

2.1. 구문 태그 세트

<표 4> 구문 태그 세트

구문 태그	의미	예시
NP	체언(명사, 대명사, 수사) 또는 외국 문자, 숫자를 비롯한 기능을 알 수 없는 미등재어	-**비가**(NP) 와서 **우산을**(NP) 샀다. -공신의 **딸을**(NP) **부인으로**(NP) 삼았다. -**그는**(NP) **셋을**(NP) 세었다. - “닥쳐(Shut up)!” (NP)
VP	용언(동사, 형용사, 보조 용언)	-비가 **와서**(VP) 우산을 **샀다.**(VP) -피부가 몹시 **건조하다.**(VP)
AP	부사구	-전화 **잠깐**(AP) 써도 될까요? -피부가 **몹시**(AP) 건조하다.
VNP	긍정 지정사구(명사+이다)	-이게 **우리집이야.**(VNP) -할머니는 **걱정이셨다.**(VNP)
DP	관형사구	-벌써 **새**(DP) 학기가 되었다. -**한**(DP) **마흔**(DP) 살쯤 되어 보인다.
IP	감탄사구(호칭 및 대답 등의 표현)	-**아이고!**(IP) 이 일을 어쩔까. -**선생님!**(IP) 질문이 있습니다.
X	의사 구(pseudo phrase, 조사 단독 어절 또는 기호 등)	
L	부호(왼쪽 괄호 및 따옴표)	
R	부호(오른쪽 괄호 및 따옴표)	

2.2. 기능 태그 세트

<표 5> 기능 태그 세트

기능 태그	의미	예시
SBJ	주어	-**비가**(SBJ) 와서 우산을 샀다. -**그는**(SBJ) 셋을 세었다.
OBJ	목적어	-비가 와서 **우산을**(OBJ) 샀다. -그는 **셋을**(OBJ) 세었다.
MOD	관형어(체언 수식어)	-**낯익은**(MOD) 자동차 한 대가 내려왔다. -어제 **본**(MOD) 영화는 무척 재미있었다.
AJT	부사어(용언 수식어)	-나는 **서울에**(AJT) 산다. -과일을 **칼로**(AJT) 잘랐다.
CMP	보어	-나는 **학생이**(CMP) 아니다. -얼음이 **물이**(CMP) 되었다. -나는 **싫다고**(CMP) 말했다.
CNJ	접속어(와/과)	-**철수와**(CNJ) 영희는 친구이다. -**개와**(CNJ) 고양이의 관계.

- 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

3. 문장 유형별 의존 관계 설정 방법

- 의존 관계 설정을 위한 문장 유형 구분은 일반 언어학 통사론의 기준을 따라 서술어가 한 개인 홑문장과 서술어가 두 개 이상인 겹문장으로 문장을 분류하여 분석한다.
- 겹문장은 다시 접속문과 내포문으로 구분되지만, 기본 원칙 (8)에 따라 접속문을 따로 분류하지 않고, 접속문을 부사절 내포문으로 분류하여 분석한다.
- 내포문은 명사절, 관형절, 부사절, 인용절 내포문의 네 가지 유형으로 분류하여 분석한다.

내포문의 유형별 예문은 아래와 같다.

- 명사절: 우리는 **시민의 관전태도도 그만큼 성숙했음**을 잊지 말아야 한다.
- 관형사절: **내가 좋아하는** 꽃은 들국화다. (관계 관형절)
  내가 사진을 좋아하는 사실을 친구들은 다 안다. (동격 관형절)
- 부사절: **멜라닌은 자외선을 차단해서** 자외선으로부터 피부를 보호해 준다.
  비가 와서 땅이 미끄럽다.
- 인용절: 그는 **그녀가 그 일을 해냈다고** 말했다. (간접 인용절)
  그는 **"그녀가 그 일을 해냈어." 라고** 말했다. (직접 인용절)

3.1. 홑문장(단문) 분석 방법

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

(1) 주어는 SBJ 기능 태그를 가지고, 홑문장 서술어에 의존하도록 분석한다.
(2) 목적어는 OBJ 기능 태그를 가지고, 홑문장 서술어에 의존하도록 분석한다.
(3) 관형어는 MOD 기능 태그를 가지고, 수식하는 명사구에 의존하도록 분석한다.
(4) 부사어는 AJT 기능 태그를 가지고, 홑문장 서술어에 의존하도록 분석한다.

- 단문 내에서 부사어로 기능하는 용언구의 경우 VP_AJT로 분석한다. 즉, 부사절 내포문의 서술어('눈에 띄게')에 AJT를 부여하지 않는다. '-게' 부사형 어미가 지배하는 구성이 단독 어절이거나('작게') 수식어만 취할 때('정말 예쁘게') AJT를 부여한다. 또한 보조 용언 구성('-게 하다')에도 AJT를 부여하지 않는다.

(가) 현주는 작게 한숨을 내쉬었다.

- 현주는 → NP_SBJ 내쉬었다
- **작게 → VP_AJT 내쉬었다.**
- 한숨을 → NP_OBJ 내쉬었다.
- 내쉬었다. → VP ROOT

(나) 최근에는 눈에(→ NP_AJT 띄게) **띄게(→ VP 늘고)** 늘고 이러한 추세가 가속화되는 경향이 있다.

(다) 저는 이런 자리에(→NP_AJT 참여하게) **참여하게(→ VP 되어서)** 되어서 매우 영광입니다.

→ '-게 되다'는 보조 용언 구성으로 처리하지 않고, 다른 '-게' 부사절과 동일하게 '-게' 성분이 단일 어절이면 VP_AJT, '-게' 성분이 다어절이면 VP로 처리한다.

(라) 이번 쿠키는 정말(→AP 예쁘게) **예쁘게(→VP_AJT 됐어요~)** 됐어요~

→ '정말', '아주' 등의 부사가 '-게' 성분을 수식하는 경우는 '-게' 성분을 VP_AJT로 분석한다. 즉, '-게' 성분에 SBJ, OBJ, AJT가 의존하는 경우에는 이를 부사절로 보아 VP로 처리하고 부사가 의존하는 경우에는 부사어로 보아 VP_AJT로 처리한다.

(5) 보어는 CMP 기능 태그를 가지고, 홑문장 서술어에 의존하도록 분석한다. (보어의 범위는 기본 원칙 (7)을 참조)

(6) 홑문장의 경우, 서술어는 문장의 가장 뒤에 위치하며, ROOT 어절에 의존하도록 분석한다.

- 홑문장의 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

3.2. 겹문장(복문) 분석 방법

- 겹문장은 명사절 내포문, 관형절 내포문, 부사절 내포문, 인용절 내포문의 네 가지 유형으로 분류하여 분석한다.

3.2.1. 명사절, 부사절 및 간접 인용절 내포문 분석 방법

3.2.1.1. 모문과 내포문의 주어 및 서술어가 각각 다른 경우

- 모문의 주어는 모문의 서술어로, 내포문의 주어는 내포문의 서술어로 연결한다.
- 내포문의 서술어는 수식하는 모문의 해당 어절에 의존하도록 분석한다.
- 명사절, 부사절 및 간접 인용절별 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 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

3.2.1.2. 모문과 내포문의 주어가 다르고, 서술어가 동일한 경우[8)]

- 모문과 내포문의 주어가 다르고 서술어가 동일한 경우 내포문의 서술어가 생략된 것으로 간주한다. 내포문의 주어는 내포문의 가장 마지막 어절에 연결하고, 내포문의 가장 마지막 성분을 모문의 서술어에 연결한다.

(가) 나는 과자를, 동생은 빵을 먹었다.

- **나는 → NP_SBJ 과자를**
- **과자를 → NP_OBJ 먹었다.**
- 동생은 → NP_SBJ 먹었다.
- 빵을 → NP_OBJ 먹었다.
- 먹었다. → VP ROOT

(나) 티셔츠는 2002년 1500만장, 2006년 1000만장이나 팔려나갔다.

- **티셔츠는 → NP_SBJ 1500만장,**
- 2002년 → NP_AJT 1500만장,
- **1500만장, → NP_SBJ 팔려나갔다.**
- 2006년 → NP_AJT 팔려나갔다.
- 1000만장이나 → NP_SBJ 팔려나갔다.
- 팔려나갔다. → VP ROOT

- 명사구 접속 구조와 혼동되는 경우가 있는데, 비슷한 구조에서 접속 조사로 이어진 명사구만 'CNJ'(명사구 접속)로 처리함에 유의한다.

(다) 2006년 연구와 2008년 연구에서는 연구팀의 주장이 유지되었다.

- 2006년 → NP 연구와
- **연구와 → NP_CNJ 연구에서는**

8) 모문과 내포문이 공유하는 성분에 대한 분석은 주어를 기준으로 세부 지침을 기술하였고, 실제 분석에서는 목적어 등의 다른 문장 성분에 확대 적용하였다. 이때 교차 의존 금지 제약을 준수하도록 주의하였다.

- 2008년 → NP 연구에서는
- 연구에서는 → NP_AJT 유지되었다.
- 연구팀의 → NP_MOD 주장이
- 주장이 → NP_SBJ 유지되었다.
- 유지되었다. → VP ROOT

3.2.1.3. 모문과 내포문의 주어가 같고, 서술어가 다른 경우

- 기본 원칙 (9)에 따라 모문과 내포문의 관계에 따라 주어의 지배소를 결정한다.[9)]
- 명사절, 부사절의 경우 주어가 내포문의 서술어를 지배소로 가지도록 분석한다.
- 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.
- 명사절, 부사절 예문은 아래와 같다.

(가) 그는 목표를 이루었음에 매우 감사했다.

- **그는** → **NP_SBJ** **이루었음에**
- 목표를 → NP_OBJ 이루었음에
- **이루었음에** → **VP_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”며’의 경우 ‘~다고 하며’가 줄어든 꼴로 보아 부사절과 동일하게

9) 간접 인용절의 경우 인용절 분석 지침에서 함께 다룬다.

처리한다. 즉, 모문의 주어는 ‘“~다”며’ 어절에 의존하고, ‘“~다”며’ 어절은 모문의 서술어에 VP로 의존한다.

(다) 위원장은 “아직 준비가 미흡하다”며 “빠른 시일 안에 준비를 마치겠다”고 강조했다.

- **위원장은 → NP_SBJ 미흡하다”며**
- “아직 → AP 미흡하다”며
- 준비가 → NP_SBJ 미흡하다”며
- **미흡하다”며 → VP 강조했다.**
- “빠른 → VP_MOD 시일
- 시일 → NP 안에
- 안에 → NP_AJT 마치겠다”고
- 준비를 → NP_OBJ 마치겠다”고
- 마치겠다”고 → VP_CMP 강조했다.
- 강조했다. → VP ROOT

- 특히, ‘대하여/대해’, ‘통하여/통해’, ‘관하여/관해’, ‘위하여/위해’에 선행하는 주어 성분은 이들 서술어에 의존하는 것으로 분석하여야 함에 유의해야 한다. (그러나 교차 금지 제약에 해당하는 경우 모문 서술어에 의존할 수 있음)

(다) ○○○ **의원은(→NP_SBJ 통해)** 15일 시정 질의를 통해 “어려운 경제 여건과 1000억 원 이상 드는 야구장 신축이 어렵다면 기존 ○○○ 야구장을 리모델링하는 방안을 검토해야 한다”고 주장했다.[10)]

3.2.2. 관형절 내포문 분석 방법

3.2.2.1. 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 연결한다.
- 내포문의 서술어는 수식하는 모문의 어절에 의존하도록 분석한다.

10) 개인 정보 비식별화 등을 고려해 말뭉치의 예문에 등장하는 개체명은 OOO로 바꿔 표기하였다.

(가) 내가 좋아하는 꽃은 들국화이다.

- **내가 → NP_SBJ 좋아하는**
- 좋아하는 → VP_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

3.2.2.2. 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([A는 ~~한 B이다.])이 이 유형에 속한다.
- 예문은 아래와 같다.

(가) 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

(나) 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

(다) 이 사료는 가축들이 먹는 음식이다.

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

3.2.3. 내포문이 3개 이상 포함되어 중의적으로 해석되는 경우

- 문장 좌측에서 우측 방향으로 순서대로 의존 관계를 설정하여 분석한다.
- 복문의 해석이 중의적일 때에는 가능한 의미 중에서 가장 가까운 서술어에 의존한다.

(가) 김민이 돈을 잘 써서 멋진 선배로 통한다.

- **김민이 → NP_SBJ 써서**
- 돈을 → NP_OBJ 써서
- 잘 → AP 써서
- **써서 → VP 멋진**
- 멋진 → VP_MOD 선배로
- 선배로 → NP_AJT 통한다.

- 통한다. → VP ROOT

- 복문의 해석이 단일할 때에는 해당 해석에 따라 구조를 분석한다.

(가) ○○○은 신작에서 판타지의 비중은 줄이고 사회를 비판한 내용을 부각했다.

- **줄이고 → VP 부각했다.**

3.2.4. 인용절 분석 방법

- 기본적으로 문장 부호에 의하여 구분된 단위를 준수하여 분석한다.
- 인용 부호가 있는 직접 인용문은 모문과 내포문을 인용 부호에 의해서 구분할 수 있으므로, 모문은 모문 구조 내의 주어와 서술어 간의 의존 관계를 연결하고 내포문은 내포문 내의 주어와 서술어 간의 의존 관계를 연결하여 분석한다.
- 보어의 범위를 엄격하게 제한하고 있기 때문에(기본 원칙 (7) 참조) 기능 태그로 CMP를 부여하는 인용절의 범위에 유의하여야 한다.

■ 인용 술어(모문 서술어)가 <표준국어대사전>에서 '-고' 성분을 격틀로 제시하고 있는 경우: 기능 태그로 CMP를 부여함.

■ 인용 술어가 <표준국어대사전>에서 '-고' 성분을 격틀로 가지고 있지 않은 경우: 기능 태그로 AJT를 부여함.

- '-다고'류 절이 이유를 나타내는 연결 어미인 경우에는(서로 잘 아는 친구 사이라고 무례하게 대해서는 안 된다.) 인용절이 아닌 일반 부사절로 처리해야 함에 유의한다. 또한 흔히 속담과 같은 관용구를 인용하는 경우에도(가지 많은 나무 바람 잘 날 없다고 우리 부모님 마음 편할 날이 없으셨지.) 일반 부사절로 처리한다.

3.2.4.1. 간접 인용의 처리

- 간접 인용은 따옴표 없이 '-다고/냐고/자고/라고'로 인용된 것으로, 부사절과

동일하게 처리하되 인용 술어(모문의 서술어)가 <표준국어대사전>에서 격틀 정보로 '-고'를 가지고 있을 경우에만 기능 태그로 CMP를 부여하고, 격틀 정보로 '-고'를 가지고 있지 않은 경우에는 AJT를 부여한다.

- 예문은 아래와 같다.

(가) 그녀가 그 일을 했다고 스스로 말했다.

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

('말하다'의 격틀에 '-고' 있음)

(나) 형사가 용의자에게 사건 시각에 어디에 있었느냐고 신문하고 있었다.

- 형사가 → NP_SBJ 신문하고
- 용의자에게 → NP_AJT 신문하고
- 사건 → NP 시각에
- 시각에 → NP_AJT 있었느냐고
- 어디에 → NP_AJT 있었느냐고
- **있었느냐고 → VP_AJT 신문하고**
- 신문하고 → VP 있었다.
- 있었다. → VP ROOT

('신문하다'의 격틀에 '-고' 없음)

3.2.4.2. 직접 인용의 처리

- 직접 인용은 따옴표를 사용하여 뒤에 '이라고/라고/고' 등이 결합한 것으로, 따옴표 밖에 있는 주어는 따옴표 안의 서술어의 주어와 동일하더라도 모문의 서술어에 연결한다.
- 즉, 직접 인용에서 모문의 주어와 내포문의 주어가 일치하는 경우 주어를 모문의 서술어에 연결하여 분석한다.

- 예문은 아래와 같다.

(가) 비평가 칼라일이 “인도와도 바꿀 수 없다”고 말하였다.

- 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

(‘말하다’의 격틀에 ‘-고’ 있음)

(나) 그는 “불이야!”라고 소리쳤다.

- **그는 → NP_SBJ 소리쳤다.**
- **“불이야!” 라고 → VNP_AJT 소리쳤다.**
- 소리쳤다. → VP ROOT

(‘소리치다’의 격틀에 ‘-고’ 없음)

- 또한, 인용된 부분이 명사로 끝나는 경우 VNP가 아닌 NP로 처리해야 함에 유의해야 한다.

(다) 그는 “○○○은 그룹 내에서 가장 중요한 사업장 중 하나”라며 “제조 역량이 뛰어나고 개발과 디자인 역량도 충분히 갖추고 있어 그룹의 미래에도 큰 영향을 미칠 **것” 이라고(NP_CMP)** 평가했다.

- 또한 ‘-다” 며’로 인용된 절은 ‘-다고 하며’가 줄어든 것으로 보고 일반적인 부사절과 동일한 원칙으로 처리한다. (3.2.1.3. 참조)

(라) 앞서 서 판사는 “도망칠 우려가 있다”며 지난달 11일 구속영장을 발부했다.[11]

- 앞서 → AP 발부했다.

11) 그러나 교차 금지 제약에 적용이 되는 경우 교차 금지 원칙을 우선적으로 고려하여 의존 관계를 설정해야 한다.
예) 앞서 서 판사는(→ NP_SBJ 발부했다.) 지난달 11일 구속영장을 “도망칠 우려가 있다”며 발부했다.

- 서 → NP 판사는
- **판사는 → NP_SBJ 있다” 며**
- “도망칠 → VP_MOD 우려가
- 우려가 → NP_SBJ 있다” 며
- **있다” 며 → VP 발부했다.**
- 지난달 → NP 11일
- 11일 → NP_AJT 발부했다.
- 구속영장을 → NP_OBJ 발부했다.
- 발부했다. → VP ROOT

3.2.5. 이중 주어문 분석 방법

- 이중 주어문은 각 구성 성분별로 동일한 서술어에 의존하도록 분석한다.
- 이때 일반적인 이중 주어문뿐만 아니라 [NP1이 NP2가 VP]의 격틀 구조를 가지는 일부 용언 분석에도 동일하게 적용한다.
- 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. 세부 구별 태깅 가이드라인

4.1. [관형어+명사+명사] 유형 분석

- 관형어가 명사구를 수식할 때, 명사구 중 관형어의 지배소는 관형어가 의미적으로 수식하는 어휘를 지배소로 분석한다.
- 여러 개의 명사들이 특별한 수식 관계 없이 나열되어 있는 경우에는, 핵어 명사를 제외한 수식 명사들을 모두 기능 표지 없이 NP로 처리하고, 각각 바로 다음 어절(오른쪽 명사)에 의존하는 것으로 분석한다.
- 예문은 아래와 같다.

(가) 무분별한 포획 문제

- **무분별한 → VP_MOD 포획**
- **포획 → NP 문제**

(나) 고려의 충신 정몽주

- **고려의 → NP_MOD 충신**
- **충신 → NP 정몽주**

(다) 조선의 제3대 임금

- **조선의 → NP_MOD 임금**
- **제3대 → NP 임금**

(라) 변형생성문법은 1987년에 출판된 노엄 촘스키의 저서이다.

- 변형생성문법은 → NP_SBJ 저서이다.
- 1987년에 → NP_AJT 출판된
- **출판된 → VP_MOD 저서이다.**
- **노엄 → NP 촘스키의**
- 촘스키의 → NP_MOD 저서이다.
- 저서이다. → VNP ROOT

(마) 결명자의 한자 뜻인 ‘눈을 밝게 틔우는 씨앗’이라는 이름대로

- **결명자의 → NP_MOD 뜻인**
- **한자 → NP 뜻인**
- **뜻인 → VNP_MOD 씨앗'이라는**
- ‘눈을 → NP_OBJ 틔우는
- 밝게 → VP_AJT 틔우는
- **틔우는 → VP_MOD 씨앗'이라는**
- **씨앗'이라는 → NP_MOD 이름대로**

(바) 레오나르도 다빈치는 역사상 가장 위대한 천재 중 하나로 기억된다.

- **레오나르도 → NP 다빈치는**
- 다빈치는 → NP_SBJ 기억된다.
- 역사상 → NP_AJT 위대한
- 가장 → AP 위대한
- **위대한 → VP_MOD 천재**
- **천재 → NP 중**
- **중 → NP 하나로**
- 하나로 → NP_AJT 기억된다.
- 기억된다. → VP ROOT

- 여러 개의 명사로 이루어진 명사구의 내부에 수식 구조가 존재하는 경우에는, 각 명사구의 수식 구조를 고려하여 의존 관계를 설정한다.

(사) 프랑스 파리 루브르 박물관

- **프랑스 → NP 파리**
- **파리 → NP 박물관**
- 루브르 → NP 박물관
- 박물관 → NP ROOT

(아) ○○○ ○○당 대표가 정계에 복귀했다.

- **○○○ → NP 대표가**
- **○○당 → NP 대표가**
- 대표가 → NP_SBJ 복귀했다.

- 정계에 → NP_AJT 복귀했다.
- 복귀했다. → VP ROOT

- 명사구와 수량사구가 연달아 나타나는 경우, 명사구 앞의 수식어는 명사구에 연결하고 명사구가 수량사구에 의존하는 것으로 처리한다.

(자) 보유하고 있는 소방헬기는 1995년에 **구입한(→VP_MOD 8인승)** 8인승 1대뿐

4.2. 명사구 접속 유형 분석

- 복수 개의 명사구가 접속 또는 나열된 경우, 가장 마지막 명사구에 의존하도록 분석한다.
- 이때, 접속 조사로 인정하는 것은 <표준국어대사전> 기준으로 12개이다. (고24/이고5, 과12/와3, 나9/이나2, 니1/이니, 다4/이다4, 랑4/이랑2, 며1/이며, 면9/이면3, 에4, 이랴2, 하고5, 하며)[12] 이들 조사가 아닌 경우('(이)라든지', '(이)라든가' 등)는 접속으로 보지 않는다.
- '및', '또는', '그리고' 등에 의해 명사구가 접속 또는 나열된 경우, 이들 접속 부사는 후행하는 명사구에 의존하는 것으로 분석한다. 그리고 이때 구문 표지만 부착하고, 기능 표지는 부착하지 않는다.
- 명사구 접속의 예는 아래와 같다.

(가) 비단 피부뿐만이 아니라 털, 눈, 귀, 심지어 뇌에도 존재한다.
- 비단 → AP 아니라
- 피부뿐만이 → NP_CMP 아니라
- 아니라 → VP 존재한다.
- **털, → NP_CNJ 뇌에도**
- **눈, → NP_CNJ 뇌에도**
- **귀, → NP_CNJ 뇌에도**
- 심지어 → AP 뇌에도
- 뇌에도 → NP_AJT 존재한다.

12) 조사 뒤의 숫자는 <표준국어대사전>에서 제시하는 의미 번호이다.

- 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

- 조사 '와/과'로 나타나는 명사구는 'N과 N' 순서인 경우에 명사구 접속으로, 'N이 N과'와 같은 순서인 경우에는 부사어로 처리한다.

(라) 철수와 영희가 만났다.

- **철수와 → NP_CNJ 영희가**
- 영희가 → NP_SBJ 만났다.
- 만났다. → VP ROOT

(마) 철수가 영희와 헤어졌다.

- 철수가 → NP_SBJ 헤어졌다.
- **영희와 → NP_AJT 헤어졌다.**
- 헤어졌다. → VP ROOT

- 조사 '(이)라든가', '(이)라든지'는 나열 기능을 하는 조사로, '명사구+(이)라든가/라든지 명사구+(이)라든가/라든지 하는' 등의 구성으로 자주 나타난다.

- 이때의 ‘명사구+(이)라든가/라든지’ 성분은 각각 NP_AJT로 서술어에 연결한다.

(바) 그는 **돈이라든지(→NP_AJT 하는) 명예라든지(→NP_AJT 하는)** 하는 것에 연연해하지 않는다.

4.3. [용언+용언] 유형 분석

4.3.1. 본용언+본용언

- 본용언이 연속적으로 나타날 경우, 주어를 앞에 위치한 서술어에 연결한다.

(가) 해구보다는 폭이 넓고 얕다.

- 해구보다는 → NP_AJT 넓고
- **폭이 → NP_SBJ 넓고**
- 넓고 → VP 얕다.
- 얕다. → VP ROOT

4.3.2. 본용언+보조 용언

- 본용언과 보조 용언이 연속하여 두 개 이상 나올 때는 주어를 본용언에 연결하고 본용언은 보조 용언에 연결한다.
- 주어 외에도 의존소들을 연결하고 있는 필수 성분들이 일차적으로 본용언에 연결되고, 본용언이 보조 용언에 의존하는 방식으로 처리한다.
- 본용언과 보조 용언이 붙여쓰기로 제시되어 있는 경우에는 해당 형식을 하나의 용언으로 처리한다.
- 보조 용언 구성이 두 개 이상 연속될 때에도 마찬가지로 본용언 → 보조 용언1 → 보조 용언2의 순으로 연결한다.

(가) 멜라닌은 물에는 용해되지 않는다.

- **멜라닌은 → NP_SBJ 용해되지**
- 물에는 → NP_AJT 용해되지
- **용해되지 → VP 않는다.**

- 않는다. → VP ROOT

(나) 그들의 화려함 속에 감춰진 갈등과 속내가 공개돼 화제를 모으고 있다.

- 갈등과 → NP_CNJ 속내가
- **속내가 → NP_SBJ 공개돼**
- **공개돼 → VP 있다.**
- 화제를 → NP_OBJ 모으고
- **모으고 → VP 있다.**
- 있다. → VP ROOT

- 문장 부사 또한 본용언이 아닌 보조 용언에 연결한다. 의사 보조 용언의 경우에도 동일하게 적용한다.

(다) **그러나(→AP 있었다.)** 이번 분기는 생각보다 순조롭게 진행되고 있었다.

4.3.3. 의존 명사 구성(의사 보조 용언 구성)

- 주 서술어 다음에 보조 용언은 아니지만 서법을 나타내는 의존 명사가 포함된 구성이 오는 경우, 해당 서술어와 의존 명사 구성을 별개의 단위로 처리하여 분석하고 의존 관계를 연결한다.[13]
- 의존 명사에 '이다', '하다' 등이 결합해 있는 경우에는 그 자체를 각각 VNP, VP 등으로 처리한다.
- 이때 문장의 주어와 의존 명사의 관계는 해당 의존 명사가 내용 명사로 대체 가능한 경우와 불가능한 경우로 구분하여 분석한다. 의존 명사가 주어와 공지시(co-reference)되는 경우에는 주어와 의존 명사 어절을 연결하고, 그렇지 않은 경우에는 주어와 서술어를 연결한다.
- 표면적으로는 '-ㄴ 것이다'와 같이 의사 보조 용언 구성에 해당되더라도, 의존 명사나 명사가 서법을 나타내지 않는 경우는 의사 보조 용언 구성에 해당되

13) 본 지침은 한국전자통신기술협회(TTA) 의존 구문 분석 가이드라인을 따르고 있으므로 한국전자통신기술협회(TTA)에서 제시하는 의사 보조 용언의 목록을 따른다. 그 목록은 다음과 같다.
'-ㄹ 수/리(가) 있다/없다', '-ㄴ/ㄹ+의존 명사(또는 일반명사)+이다'(것/터/뿐/따름/모양/지경/참/중/노릇/예정/길), '-ㄹ {만/법/듯}하다', '-는 말이다', '-ㄴ/ㄹ 듯(도) 하다', '-ㄹ 것 같다', '-ㄹ 것을(걸) 그랬다', '-어서는 안된다', '-고 해서', '-든지 하다'
각각의 예시는 한국전자통신기술협회(TTA) 지침을 참고하라.

지 않는다.

(가) 그는 일어날 수 없었다.

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

(나) 나는 곧 밥을 먹을 것이다.

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

(다) 이 사료는 가축들이 먹는 것이다.

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

→ 이때의 ‘것’은 ‘사료, 식품’을 뜻하므로 ‘사료는’이 ‘것이다’에 의존함.

(라) 내가 놀란 것은 철수가 천재라는 것이다.

- 내가 → NP_SBJ 놀란
- 놀란 → VP_MOD 것은
- **것은 → NP_SBJ 것이다.** (것=사실 → 사실=사실)
- 철수가 → NP_SBJ 천재라는
- 천재라는 → VNP_MOD 것이다.
- 것이다. → VNP ROOT

→ 이때의 ‘것’은 ‘사실’을 뜻하므로 ‘것은’이 ‘것이다’에 의존

하는 것으로 처리함.

(마) 곧 비가 올 것 같다.

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

4.3.4. ‘NP 중이다’ 구문

- ‘NP 중이다’ 구문에서 NP의 논항이 되는 성분은 ‘중이다’가 아닌 해당 NP에 의존한다.

(가) 구치소는 사건경위를 조사한 뒤 손씨에게 규율 위반 행위에 상응하는 징벌을 내리는 **방안을(→ NP_OBJ 검토)** 검토 중이다.

4.4. 부호

- 따옴표 (‘ ’ / “ ”) 및 괄호(< > / [ ]) 등과 같이 좌우 짝이 있는 부호에 한정하여 각각 L, R 표지를 부착하고, 그 외의 경우에는 일괄적으로 X를 부착한다. (괄호가 수리 기호나 층위 표시로 쓰일 때에도 X로 분석함)
- L은 R에 의존하도록 분석한다.

(가) “나는 너를 좋아해 ” 라고 말했다 .

- “ → L ”
- ” → R 라고

(나) “ 나는 너를 좋아해”라고 말했다.

- “ → L 좋아해”라고
- 좋아해”라고 → VP_CMP 말했다.

(다) 과일(사과, 배 등)의 등급은

- 과일(사과, → NP_CNJ 배
- 배 → NP 등
- 등 → NP)의
-)의 → X_MOD 등급은

(라) 과일 (사과, 배 등)의 등급은

- 과일 → NP)의
- (사과, → NP_CNJ 배
- 배 → NP 등
- 등 → NP)의
-)의 → X_MOD 등급은

(마) 과일 (사과, 배 등) 의 등급은

- 과일 → NP 의
- (→ L)
- 사과, → NP_CNJ 배
- 배 → NP 등
- 등 → NP)
-) → R 의
- 의 → X_MOD 등급은

- 단락 기호는 기호가 속하는 최상위 핵에 의존하도록 분석한다.

(바) 철수의 버릇 : 다리 꼬기, 이갈기

- 버릇 → NP :
- : → X 이갈기

(사) 철수의 버릇: 다리 꼬기, 이갈기

- 버릇: → NP 이갈기
- 꼬기, → NP_CNJ 이갈기

(아) - 철수의 버릇 : 다리 꼬기, 이갈기

- - → X 이갈기
- 버릇 → NP :
- : → X 이갈기

(자) ● 내년 주요 핵심안은 예산 결의 문제

- ● → X 문제

- 그 외: 후행 어절에 의존하도록 분석한다.

(차) 세종(1418 ~ 1450)은 조선전기 제4대 왕이다.

- 세종(1418 → NP ~
- ~ → X 1450)은
- 1450)은 → NP_SBJ 왕이다.

- 삽입구가 복수의 어절로 구성되어 있을 경우, 기호로 결합된 복합 형태소는 선행 성분을 기준으로 구문 태그를 결정하고, 후행 성분을 기준으로 기능 태그를 결정한다.

(카) 전문위원의 임기는 3년을 보장한다(1차에 한하여 연임 가능).

- 보장한다(1차에 → VP_AJT 한하여

4.5. 외국 문자/외국어 처리 방법

- 외국 문자, 숫자를 비롯한 기능을 알 수 없는 미등재어의 구문 태그는 NP이다.

(가) “닥쳐(Shut up)!”

- “닥쳐(Shut → VP up)!”
- up)!” → NP ROOT

- 외국어는 바로 다음 요소에 의존하도록 분석한다.

(나) 아이 러브 유

- 아이 → NP 러브
- 러브 → NP 유

(다) I love you

- I → NP love
- love → NP you

4.6. 띄어쓰기 오류 처리 방법

- 어절 내부가 분할되어 있는 경우, 구문 태그와 기능 태그는 형태 분석 결과를 기준으로 정하고, 의존 관계는 바른 띄어쓰기를 기준으로 정한다.[14] (2019 국립국어원 형태 분석 말뭉치 참조)
- 절단 어절의 형태 분석 결과와 구문 분석 태그의 대응표는 다음과 같다.

<표 6> 절단 어절의 형태 분석 결과와 구문 분석 태그 대응표

형태 분석 태그	구문 분석 태그
NNG, NNP, NNB, NP, NR, XSN, XR, NF	NP
VV, VA, VX, VCN, EP, EF, EC, ETN, XSV, XSA, NV	VP
MMA, MMD, MMS	DP
MAG, MAJ	AP
IC	IP
JKS, JKC, JKG, JKO, JKB, JKV, JKQ	X
SF, SP, SS, SE, SO, SW, SL, SH, NA	NP

- 기능 태그는 맨 마지막 분할 어절에 부여한다. (조사 생략 경우도 마찬가지임)
- 원래 어절 내부에서는 다음 분할 어절에 의존하고, 원래 어절의 분할 어절은 지배소 어절(의 최종 분할 어절)에 의존한다.

(가) 마음 씨 가 중요하 다.

- 마음 → NP 씨
- 씨 → NP 가

14) ‘6.5.1. 절단된 어절의 처리’ 방침을 함께 참고할 수 있다.

- 가 → X_SBJ 다.
- 중요하 → VP 다.
- 다 → VP ROOT

5. 세부 유형별 가이드라인

5.1. 의존 관계 태그 부착 세부 유형 가이드라인

5.1.1. 보조사적 쓰임을 보이는 '이/가', '을/를'의 주석

- 본용언에 화용적 기능을 가지는 조사 {가/를}이 붙은 경우는 기능 표지를 부착하지 않는다. 즉, 명사형(-음, -기)을 제외한 용언의 활용형에 붙은 조사 {가/를}은 무시한다.

(가) 철수가 밥을 이틀내 먹지를 않았다.

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

(나) 그 산은 그리 높지가 않다.

- 그 → DP 산은
- 산은 → NP_SBJ 높지가
- 그리 → AP 높지가
- **높지가 → VP 않다.**
- 않다. → VP ROOT

- '-기 바라다', '-기 시작하다' 등의 '-기' 절을 요구하는 서술어의 경우 뒤에 격 조사가 붙는 경우와 붙지 않는 경우 모두 기능 태그를 부착하여 VP_OBJ로 태깅한다.

(다) 나는 네가 빨리 오기(를) 바란다.

- 나는 → NP_SBJ 바란다.
- 네가 → NP_SBJ 오기(를)
- 빨리 → AP 오기(를)
- **오기(를) → VP_OBJ 바란다.**
- 바란다. → VP ROOT

5.1.2. ‘~즈음’, ‘~쯤’ 부사구의 주석

- ‘~즈음’, ‘~쯤’ 등과 같이 의미적으로 시간과 공간을 의미하고 조사가 없는 경우 AJT 기능 태그를 부착한다.

(가) 광복절 즈음 해서 독립기념관을 찾았다.

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

- 그러나 ‘~즈음’, ‘~쯤’ 뒤에 다른 격 조사가 쓰이는 경우 격 조사를 고려하여 주석한다.

(나) 다섯 명 즈음의 학생이 길에 서 있었다.

- 다섯 → DP 명
- 명 → NP 즈음의
- **즈음의 → NP_MOD 학생이**
- 학생이 → NP_SBJ 서

5.1.3. 격 조사가 붙은 수량 관련 표현의 주석

- 시간이나 거리 등 수량이나 단위를 나타내는 명사구에 목적격 조사 ‘을/를’이 붙은 경우에는 목적어로 분석한다.

(가) 나는 학교까지 다섯 시간을 걸었다.

- 나는 → NP_SBJ 걸었다.
- 학교까지 → NP_AJT 걸었다.
- 다섯 → DP 시간을
- **시간을 → NP_OBJ 걸었다.**
- 걸었다. → VP ROOT

- ‘을/를’이 결합하지 않았더라도 ‘을/를’ 이외의 다른 격 조사가 결합할 수 없는 개체/수량/횟수/시간/거리 표현의 경우 OBJ로 분석한다.

(나) 노동자들의 삶을 담기 위해 2년에 걸쳐 **5번(→NP_OBJ 방문했다.)** 방문했다.

(다) 운동장을 세 **바퀴(→NP_OBJ 뛰었다.)** 뛰었다.

5.1.4. 품사와 문장 성분이 일치하지 않는 경우

- 품사로는 부사나 활용형 등이 쓰였지만 문장 내에서 인용이 된 것처럼 쓰인 경우에는 해당 문장 성분을 기준으로 기능 태그를 부여하되, 구문 태그 분석은 단어의 본래 품사를 기준으로 한다. 다만, 형태소의 일부가 잘린 경우 미등재어로 보아 NP를 부여한다.

(가) ‘거꾸로’는 ‘거꾸’와 ‘로’로 분석할 수 있을까?

- **‘거꾸로’는 → AP_OBJ 분석할**
- **‘거꾸’와 → NP_CNJ ‘로’로**
- **‘로’로 → NP_AJT 분석할**
- 분석할 → VP_MOD 수
- 수 → NP_SBJ 있을까?
- 있을까? → VP ROOT

5.2. 세부 유형 가이드라인

5.2.1. 서술어의 역할('~이다. 그리고')을 하는 '~으로' 의 주석

- '~으로' 가 의미적으로 '~이다. 그리고' 와 같이 사용되었다면, 예외적으로 서술어로 인정한다. 즉, 주어 논항을 가질 수 있다.

(가) 모나리자는 레오나르도 다빈치가 그린 초상화로, 현재 프랑스 파리 루브르 박물관에 전시되어 있다.
- 모나리자는 → NP_SBJ 초상화로,
- 초상화로, → NP_AJT 있다.

- '~으로' 가 쉼표로 연결되어 있지 않아도 의미적으로 '~이다. 그리고' 에 대응된다면 마찬가지로 서술어로 인정한다.

(나) 모나리자는 레오나르도 다빈치가 그린 초상화로 현재 프랑스 파리 루브르 박물관에 전시되어 있다.
- 모나리자는 → NP_SBJ 초상화로
- 초상화로 → NP_AJT 있다.

5.2.2. 부사 '없이', '같이' 의 주석

- '없이' 나 '같이' 와 같은 부사(용언의 활용형이 아님에 유의)는 서술어와 마찬가지로 논항을 취할 수 있다. 특히 부사 '같이' 는 앞에 '~와' 에 해당하는 명사구가 나타나는 경우 '~와' 명사구를 '같이' 의 부사어로 처리한다.

(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.

- 나섰다. → VP ROOT

(나) 예상한 바와 같이 주가가 크게 떨어졌다.

- 예상한 → VP_MOD 바와
- **바와 → NP_AJT 같이**
- **같이 → AP 떨어졌다.**
- 주가가 → NP_SBJ 떨어졌다.
- 크게 → VP_AJT 떨어졌다.
- 떨어졌다. → VP ROOT

5.3. 연결된 부사구 세부 유형 가이드라인: '~부터 ~까지', '~에서 ~으로'의 주석

- '~부터', '~까지' (또는 '~에서', '~로')와 같은 부사구는 각각을 지배소에 연결한다.

(가) 그는 작년부터 지금까지 열심히 일했다.

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

(나) 부산에서 서울로 가는 표 한 장 있나요?

- **부산에서 → NP_AJT 가는**
- **서울로 → NP_AJT 가는**
- 가는 → VP_MOD 표
- 표 → NP 장
- 한 → DP 장
- 장 → NP_SBJ 있나요?
- 있나요? → VP ROOT

- 그러나 만일 '~부터 ~까지를', '~에서 ~로의' 등과 같이 '~부터', '~까지'

(또는 ‘~에서’, ‘~로’) 부사구가 하나의 단위로 묶이는 경우 선행 성분을 후행 성분의 부사어(AJT)에 연결한 후, 후행 성분을 지배소에 연결한다.

(다) 평균 90점부터 100점까지를 모두 금상으로 처리한다.

- 평균 → NP 90점부터
- **90점부터 → NP_AJT 100점까지를**
- **100점까지를 → NP_OBJ 처리한다.**

(라) 이것은 바로 개인주의에서 단체주의로의 전환을 의미했다.

- 이것은 → NP_SBJ 의미했다.
- 바로 → AP 의미했다.
- **개인주의에서 → NP_AJT 단체주의로의**
- **단체주의로의 → NP_MOD 전환을**

5.4. 장형 사동 구문 유형 세부 가이드라인

- ‘-게 하다’의 장형 사동 구문은 다른 보조 용언과 동일하게 처리한다. 즉, ‘-게 하다’에서 선행하는 용언(‘-게’)에 다른 성분들을 모두 연결한다.
- 또한 ‘A가 B가/B를/B에게 V-게 하다’와 같은 장형 사동 구문의 B성분은 격조사에 의존하여 기능 태그를 부착한다. 즉, ‘B가’는 SBJ, ‘B를’은 OBJ, ‘B에게’는 AJT로 처리한다.

(가) 그가 철수를 집에 가게 하였다.

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

(나) 그가 철수에게 집에 가게 하였다.

- 그가 → NP_SBJ 가게
- 철수에게 → NP_AJT 가게

(다) 그가 철수가 집에 가게 하였다.

- 그가 → NP_SBJ 가게
- 철수가 → NP_SBJ 가게

5.5. 여러 개의 문장 처리[15)]

- 여러 개의 문장이 분할되지 않은 상태로 제시되어 있는 경우, 각 문장의 서술어가 순차적으로 의존하도록 처리한다.

(가) “철수는 집에 갔어. 영희는 모르겠어. 민지는 집에 있겠지.” 라고 말했다.

- 갔어. → VP 모르겠어.
- 모르겠어. → VP 있겠지.” 라고
- 있겠지.” 라고 → VP_CMP 말했다.

5.6. 명사-부사 통용어 또는 체언 수식 부사의 처리

- 명사-부사 통용어 중 뒤에 격 조사가 붙지 않고 서술어에 의존하는 경우(‘오늘’ 등) 또는 본래 부사이지만 체언을 수식하는 용법으로 쓰이는 경우에는 (‘가장’, ‘아주’, ‘바로’ 등) 이들의 품사적 지위를 기준으로 하여 AP로 처리한다.

(가) 친구네 집은 우리 집 바로 뒤에 있어.

- 친구네 → NP 집은
- 집은 → NP_SBJ 있어.
- 우리 → NP 집
- 집 → NP 뒤에
- **바로 → AP 뒤에**
- 뒤에 → NP_AJT 있어.
- 있어. → VP ROOT

15) 이 항목은 문어 말뭉치에만 적용된다. 구어 말뭉치에서는 문장 분할의 기준이 다르기 때문이다.

(나) **오늘(→AP 해야)** 해야 할 일을 다음 날로 미루어서는 안 된다.

5.7. 술어 생략 세부 유형 가이드라인

- 내포문의 서술어가 생략되었다고 판단되는 경우 내포문의 표층에 나타난 마지막 어절에 선행하는 성분들을 의존하게 하고 마지막 어절을 모문의 서술어에 의존하게 한다.
- 내포문의 서술어가 생략된 경우 형태 기준(조사)으로 분석한다.

(가) 영희가 철수를 조건으로 매장을 임시로 설치하였다.

- 영희가 → NP_SBJ 조건으로
- 철수를 → NP_OBJ 조건으로
- 조건으로 → NP_AJT 설치하였다.

(나) 나는 여섯 살, 이름은 철수예요.

- 나는 → NP_SBJ 살,
- 여섯 → DP 살,
- **살, → NP 철수예요.**
- 이름은 → NP_SBJ 철수예요.
- 철수예요. → VNP ROOT

→ 계사 '이다'가 생략되는 경우 형태를 기준으로 '살,'을 NP로 주석하여 후행절 서술어에 연결하도록 한다.

- 모문의 서술어가 생략된 경우 형태 기준(조사)으로 분석한다.

(다) 과거 하계 올림픽의 정식 종목으로 맞는 **것은?(→NP_SBJ)**

(라) **밥을.(→NP_OBJ)**

(마) 철수가 밥을.

- 철수가 → NP_SBJ 밥을.
- 밥을. → NP_OBJ ROOT

6. 구어 구문 분석을 위한 추가 지침

6.1. 감탄사의 범위와 처리

- 본 구문 분석에서의 ‘감탄사’는 <표준국어대사전>에서 품사가 감탄사로 되어 있는 것(응답, 감탄 등)과 머뭇거림, 발화의 시작, 화제 전환 등에 쓰이는 이른바 ‘담화 표지’를 포함하는 개념을 뜻한다. 이들 감탄사에 대해서는 IP(감탄사구) 표지를 부여한다.

6.1.1. 부사, 대명사와 감탄사의 구별

- 부사나 대명사인지 감탄사인지가 모호한 경우에는 감탄사(IP)로 처리한다. 특히, 부사 또는 대명사로 해석하여 의존소와 의존 관계를 설정하였을 때 교차 의존 금지 제약을 어기게 되는 경우에는 감탄사(IP)로 처리하여 바로 다음 어절에 의존하도록 분석한다.

  (가) 외국인 유학생들한테 들어가는(→VP_MOD 비용들이) **이제(→IP 추가)** 추가 비용들이 많이 증가하고 있어요.

  → ‘들어가는’이 ‘비용들이’에 의존하므로 ‘이제’가 ‘증가하고’에 의존할 경우 교차 의존 금지 제약을 위반하게 된다. 따라서 이때 ‘이제’는 감탄사로 보아 IP로 처리해서 ‘추가’에 의존하도록 한다.

6.1.2. 감탄사의 주석

- 문장 부호의 유무와 관계없이 감탄사는 IP로 바로 다음 어절에 의존하도록 처리한다.

  (가) **예.(→IP 간장하구요)** 간장하구요 설탕 또 요렇게 넣으시고.

  (나) **그래?(→IP 나도)** 나도 모르겠던데

6.2. 주제어의 주석

- 주제어(문장의 서술어와 호응을 보이지 않고 문장의 화제를 제시하는 기능만 하는 경우)는 형태를 기준으로(격 조사) 처리하고 '은/는'으로 나타나는 경우, 대응되는 문장 성분(주어, 목적어, 부사어 등)이 있으면 해당 기능 태그를 부여하고, 문장 안의 서술어를 기준으로 대응되는 문장 성분이 없는 경우에는 일괄적으로 SBJ 태그를 부여한다.

  (가) 앞 **부분은(→NP_OBJ 질문하고)** 걔가 막 질문하고 넘어가고 넘어가고 그러던데

  (나) **논어는(→NP SBJ)** 흔히 잔소리가 많은 그러한

6.3. 호칭어의 주석

- 호칭어는 IP로 지배소 서술어에 의존하도록 처리한다. 또한 선행하는 수식 요소들은 일반적인 구문 분석 지침에 따라 처리한다.

  (가) 임 **선생님(→IP 있다구오?)** 요걸 또 쉽구 간편하구 맛있게(→VP_AJT 만드는) 만드는 방법이 있다구오?

6.4. 불완전한 문장의 주석

- 개별 어절에는 문제가 없지만 비문에 가까운 문장에 대해서는 다음과 같이 의존 관계를 설정한다.
- 해당 어절이 자연스럽게 이해될 수 있는 방안을 상정하여 그에 맞춰 처리하지 않고, 있는 그대로의 표면형을 중시하여 처리한다.
- 감탄사, 절단된 어절이 아닌 완전한 어절을 이루는 경우, 해당 어절이 의존할 수 있는 어절이 문장 안에 있으면 최대한 의존하도록 처리하되 의존할 수 있는 어절이 문장 내에 없는 것으로 판단되면 문장의 최상위 지배소에 의존하도록 처리한다.

(가) **아무리(→AP 되잖아요.)** 추석 하면은 명절 때는 전이랑 적이 있어야 되잖아요.

→ 위의 예문에서 '아무리'는 '아무래도'로 대치하면 자연스럽게 이해될 수 있지만, 이러한 과정을 거치지 않고 있는 그대로의 표면형을 중시하여 처리한다.

(나) 뭐 들은 얘기가 그 **비슷한(→VP_MOD 소리인지)** 그때 그게 무슨 소리인지 모르는데

(다) **그래야지(→VP 어우러져서)** 양념이(→NP_SBJ 섞인) 서로(→AP 섞인) 섞인(→VP_MOD 어우러져서) 어우러져서

- 또한, 어절의 기능을 판단하기 어려운 경우에는 'NP'로 처리하고 바로 다음 어절에 의존하도록 한다.

(라) 뒤집어가지고 또 이 쪽도 막 **저이(→NP 뚜들겨요)** 뚜들겨요 인제.

- 문장의 마지막 어절이 '은/는' 등의 보조사로 끝나는 경우, 기능 태그는 SBJ를 부여한다.

(마) 남 대표가 원화 작가 고용을 확대한 **첫걸음은(→NP_SBJ)**

- 또한, 문장의 마지막 어절이 '와/과'로 끝나는 경우, 기능 태그를 AJT로 분석한다.

(바) 아 그래서 우리가 아 지금의 **상황과(→NP_AJT)**

- '(이)나'와 같이 보조사로도 해석이 가능하고, 접속 조사로도 해석이 가능한 조사가 문장의 맨 마지막 어절로 나타나는 경우 보조사로 해석하는 것을 우선으로 하여 기능 태그를 SBJ로 부여한다.

(사) 그렇게 보면 이런 **경우나(→NP_SBJ)**

- ‘-하고 -하고’, ‘-랑 -랑’과 같이 접속 조사가 반복해서 쓰였는데 그 상태로 문장이 끝나는 경우, 마지막 어절의 기능 태그는 AJT가 된다.

(아) 빵하고(→NP_CNJ) 밥하고(→NP_AJT)

6.5. 불완전한 어절의 주석(절단된 어절, 발화 수정된 어절 등)

6.5.1. 절단된 어절의 처리

- 완전한 어절을 이루지 않고 어절이 절단된 경우 ‘NP’로 처리하고 다음 어절에 의존하도록 한다.

(가) 그 붙(→ NP 그) 그 아이부터 세 명이었죠.

- 완전한 어절이 띄어쓰기 오류로 인해 절단된 경우 형태 분석 결과에 따라 구문 태그를 설정하는데, 자세한 내용은 4.7.을 참고한다.
- 조사가 절단되어 명사(구)가 분리된 경우, 명사(구)의 수식어는 명사에 의존하게 하고, 명사를 조사에 의존하게 한다.

(나) 거기에 애매한(→VP_MOD 부분) 부분(→NP 이) 이(→X_SBJ 있는데) 있는데

- 완전한 어절이 절단되어 나타난 경우 기능 태그는 맨 마지막 분할 어절에 부여하며 의존 관계는 바른 띄어쓰기를 기준으로 한다. 특히 용언의 활용형이 절단된 경우 완전한 어절이 무엇인지 고려하여 처리함에 유의한다.

(다) 밥을(→NP_OBJ 구나) 먹는(→VP 구나) 구나(→VP ROOT)
→ 이때 ‘먹는’이 VP_MOD로 보일 수 있는 형태이지만, 뒤따르는 어절 ‘구나’와 본래 한 어절이었음을 알 수 있다.

(라) 불조절이(→NP_SBJ 요.) 징짜(→AP 요.) 중요하겠어(→VP 요.) 요.(→X ROOT)

→ (라)에서 '중요하겠어요'가 본래 하나의 어절로 나타나야 하지만 절단되어 나타났다. '불조절이', '징짜'는 절단된 어절의 맨 마지막인 '요.'에 의존하도록 한다. '-어요'는 어말어미 '-어'에 보조사 '요'가 결합한 것이므로(즉 '-어'만으로도 어절이 완전히 성립하므로) '중요하겠어'는 VP, '요.'는 X로 처리한다.

6.5.2. 발화 수정의 처리

- 발화 수정에서 각 어절이 완전한 경우(형태를 기반으로 구문, 기능 태그의 설정이 가능한 경우)에는 구문, 기능 태그를 형태를 기반으로 부여하고 일반적인 구문 분석의 원칙에 따른다.
- 그러나 해당 어절의 기능을 알기 어려운 경우에는 'NP'로 바로 다음 어절에 의존하도록 처리한다.

(가) 두 번째 실제로 무슨 말을 어떤 톤으로(→NP_AJT 할) 할(→VP_MOD 하는가) 하는가 이게 좀 중요하게 생각했는데

(나) 그 대표분하고 서기 뭐 하는 분이랑 그 라디오 방송에 하시는걸(→NP_OBJ 듣고) 보(→NP 듣고) 듣고

6.5.3. xx, xxx 등의 처리

- 구어 전사에서 xx, xxx 등은 잘 들리지 않는 부분을 전사한 것이다. 이들 어절은 다른 불완전한 어절과 동일하게 NP로 처리하고, 조사나 어미 등이 부착되지 않아 그 기능을 알기 어려운 경우에는 바로 다음 어절에 의존하도록 한다.

(가) xx(→NP 그럴만한) 그럴만한 얘긴가?

(나) xx가(→NP_SBJ 얘기했다고) 그렇게 얘기했다고 합니다.

6.6. 반복되는 어절, 공지시어의 주석, 어미 반복 구성의 주석

6.6.1. 반복되는 어절의 처리

- 형태적, 기능적으로 완전한 어절이 여러 번 반복되는 경우 각각 지배 어절에 의존하도록 처리한다.

  (가) 요렇게 **딱(→AP 해)** **딱(→AP 해)** 준비를 해 놓는 거예요 인제.

6.6.2. 공지시어의 의존 관계 설정

- 문장 내 공지시어는 각각 서술어에 의존하도록 한다.

  (가) 누름적을 할 때 밀가루와 달걀을 입힌 **거(→NP_OBJ 부친다.)** 옷을 입힌 **걸(→NP_OBJ 부친다.)** 기름에 부친다.

- 그러나 '사과 배 바나나 이런 거를 먹었다.' 에서의 'N 이런 거', 즉 '등'과 같이 앞의 예시를 묶어서 다시 언급하는 경우, 앞의 명사구를 '이런 거'에 NP로 연결하고 '이런 거를'을 서술어에 연결한다.[16]

  (나) **사과(→NP_CNJ 바나나)** **배(→NP_CNJ 바나나)** **바나나(→NP 거를)** **이런 거를(→NP_OBJ 먹었다.)** 먹었다.

- '이런 거' 앞에 용언구가 나오는 경우에도 위와 동일하게 처리한다.

  (다) 이제 보험은 **고의성으로(→NP_AJT 했냐)** 누가 잘못을 **했냐(→VP 있었냐)** 고의성이 **있었냐(→VP 거를)** 뭐~ 이런 거를 따지는데

- '이런 거'의 선행 명사구가 접속 조사와('(이)나' 등) 함께 나타나는 경우, '이런 거'에 선행 명사구들이 직접 접속하는 것으로 처리한다.

16) 그러나 '그래서 사과한다 이렇게 생각했어요.'와 같이 명사구 나열의 예시가 아닌 경우 일반적인 의존 구문 분석과 동일하게 처리한다.
그래서 사과한다(→VP_CMP 생각했어요.) 이렇게(→VP_AJT 생각했어요.) 생각했어요.

(라) 사과나(→NP_CNJ 거를) 배나(→NP_CNJ 거를) 바나나나(→NP_CNJ 거를) 이런 거를

- '(이)라든지, (이)라든가' 의 경우 접속 조사가 아니므로 아래와 같이 처리한다.

(마) 담겨진 어떤 뜻이라든지(→NP_OBJ 보면서) 이런 걸 보면서

6.6.3. 어미 반복 구성의 처리

- 어미 반복 구성은 기본적으로 모문의 서술어에 각각 연결해야 하지만 서술어가 한 문장 안에 나타나지 않는 경우 첫 번째 구성을 두 번째 구성에 의존하도록 한다. 이때 기능 태그 CNJ를 사용해서는 안 된다.(CNJ는 명사구 접속에만 제한적으로 사용됨)

(가) 수준별 수업이라든지(→NP_SBJ 수업이라든지) 무슨 이동식 수업이라든지

6.7. 여러 문장이 한 주석 단위에 포함되는 경우

- 구어 자료에서는 하나의 주석 단위(발화 단위)를 크게 한 문장과 동등하게 한다. 즉, 한 주석 단위 안에 마침표가 있더라도 별개의 문장으로 보지 않는다.
- 다만, 별개의 문장임이 확실한 경우에는 문어와 마찬가지로 한 문장의 최상위 지배소를 다음 문장의 최상위 지배소에 연결한다.

(가) 통계방법이 뭔지를 찾아내라.(→VP 를) 를(→X_OBJ 줬어요.) 숙제로 줬어요.

(나) 자(→IP 고깁니다.) 고깁니다.(→VNP 그리고.) 네.(→IP 그리고.) 그리고.(→AP ROOT)

6.8. 인용 표지가 없는 인용의 주석

- 명확한 인용 표지(이라고, 라고, 다고 등)가 없는 경우이더라도, 인용 구문임이 언어적으로 표현되면 해당 기능 태그로 분석한다. 이는 후행하는 부분에 '하고, 하면, 하는' 등 인용 내용을 요구하는 술어가 있고 인용 내용 없이는 완전한 구성을 이루지 못하는 경우에 적용된다.

  (가) 근데 인제 산적용 **주세요(→VP_CMP 하면)** 하면(→VP 밀어서) 이케 한 번 쭉 밀어서

- 완전한 문장이 인용되는데 인용 표지가 없고 후행하는 부분이 완전한 문장인 경우에는 여러 문장이 나열되는 것으로 처리한다.

  (나) 우리가 할 수 있지 않을까?(→VP_CMP 있습니다) 이렇게 생각을 하고 있습니다.

6.9. 구조적 중의성이 있는 명사구의 주석(명사구 나열 vs 하나의 명사구)

- 쉼표, '또는', '및', '이나', '하고' 등의 외현적인 접속 표지가 없더라도 명확하게 명사구가 나열된 경우에는 CNJ 태그를 부여한다.

  (가) **두 사람(→NP_CNJ 사람)** **네 사람(→NP_CNJ 사람)** 여덟 사람(→NP 식으로) 이런 식으로 앉잖아 보통[17]

- 명사구의 나열인지 하나의 명사구인지 알기 어려운 경우에는 하나의 명사구로 보고 처리한다.

  (나) 예를 들면 **국민(→NP 고객)** 고객 중심의 업무추진 관련된 부분.

17) '두 사람', '네 사람' 모두 '여덟 사람'의 '사람'에 연결한다.

6.10. 한 어절 안에 여러 단위가 포함되어 있는 경우

- 띄어쓰기 분할이 제대로 되어 있지 않아 한 어절 안에 여러 기능 단위가 포함되어 있는 경우, 맨 마지막 단위가 해당 어절의 핵임을 고려하여 구문 및 기능 태그를 부여한다.

(가) 유아교육을 전공한 원장의 관리 하에 운영되는 **공립유치원이다보니(→VP 높습니다.)** 학부모들의 만족도가 높습니다.

→ '공립유치원이다(VNP) 보니(VP)'와 같이 두 어절이어야 하는 부분이 한 어절로 나타난 경우, '보니'의 구문 태그와 기능 태그를 기준으로 하여 VP로 주석한다.

6.11. 문장의 마지막 어절이 '-다고' 류 어미로 끝나는 경우

- 마지막 어절이 '-다고'류 어미로 끝나는 경우, 해당 어절이 인용이 분명한 경우에는 VP_CMP를 부여하고, 인용인지 이유를 나타내는 부사절인지 모호한 경우에는 VP를 부여한다.

(가) 유아~ 그것이 상당히 일본 학계에 파장을 **던졌다고(→VP)** 어~

→ 이 경우 '던졌다고'의 '-다고'가 연결 어미인지 인용 표지인지 확실하지 않다. : VP

(나) 일본 팬들을 생각해서 영화에 참여하기 좀 **어려웠겠구나라고(→VP_CMP)**

→ 이 경우 '~ 어려웠겠구나'와 같이 하나의 완전한 문장이 종결어미와 함께 인용된 것이므로 인용이 확실하다. (다만 인용 술어가 드러나지 않아 CMP나 AJT 분석이 불분명할 때는 '하다, 말하다, 생각하다' 등의 기본 술어로 가정한다) : VP_CMP

6.12. '에'로 전사된 관형격 조사 '의'의 처리

- 관형격 조사로 보이는 것이 '에'로 전사되어 나타난 경우, 관형격 조사로서의 쓰임이 명확한 경우에는 NP_MOD로 수식 명사에 의존하게 하되 그 쓰임이 명확

하지 않은 경우 표층형을 따라 NP_AJT로 해당 절의 서술어에 의존하게 한다.

(가) 평균 옷값을 가지고 이 **옷에(→NP_MOD 가격을)** 가격을 추론한 거예요.

6.13. 미등재어의 처리

- <표준국어대사전> 기준 미등재어 중, 기능을 파악할 수 있는 미등재어는 해당 기능에 따라 구문, 기능 태그를 부여한다.

(가) 주인공이 그냥 **졸라(→AP 성공하는)** 성공하는 내용이에요. 승승장구해.
→ '졸라'는 미등재어이나 부사의 기능을 하고 있으므로 AP로 처리한다.

(나) **오우(→IP 역시나)** 역시나 그분이 나오시길래 무슨 일인가 했더니
→ '오우'는 미등재어이나 감탄사의 기능을 하고 있으므로 IP로 처리한다.

(다) 걔는 **완전(→AP 사랑꾼이지)** 사랑꾼이지(→VNP ROOT)
→ '완전'은 <표준국어대사전> 명사로만 등재되어 있으나 이 문장에서는 '정말', '진짜' 등의 부사 용법과 동일하게 사용되고 있고, <우리말샘>에서는 부사 용법이 등재되어 있으므로 AP로 처리한다.

- 그러나 기능을 파악할 수 없는 미등재어는 NP로 처리하여 바로 다음 어절에 연결한다.

3.3. 목적격 무형 대용어 복원 분석 지침

1. 개념 및 기본 원칙

1.1. 개념

(1) 목적격 무형 대용어란 서술어가 요구하는 논항 중 문장에서 생략된 목적격 명사구(이하 생략어 또는 생략된 목적어)이다.
(2) 선행어란 생략어를 의미적으로 복원하는 표현을 말한다.
(3) 목적격 무형 대용어 복원에서는 목적격 명사구가 생략된 경우, 구문 분석과 의미를 고려하여 선행어를 찾아 생략된 목적어를 복원한다.

1.2. 기본 원칙

(1) 자연 언어 처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서 크게 벗어나지 않도록 생략된 목적어를 복원한다.
(2) 의존 구문 분석 말뭉치 구축을 위한 지침 수립과 최대한의 일관성을 유지하도록 하며, 두 층위에서 기술적으로 다를 수밖에 없는 부분은 지침에 제시한다.
(3) 2019년 주격 무형 대용어 복원 지침과 일관성을 유지한다.
(4) 서술어가 구성하는 문장의 격틀은 국립국어원 <표준국어대사전>, <우리말샘>을 기반으로 한다.
(5) 대상 서술어가 속한 문장 내에 선행어가 존재하면, 해당 선행어를 우선적으로 목적어로 복원한다.
(6) 대상 서술어가 속한 문장 내에 선행어가 존재하지 않으면, 전방에 위치한 선행어로 생략어를 복원한다.
(7) 대상 서술어가 속한 문장이나 전방에 선행어 후보가 존재하지 않으면, 후행하는 문장에서 생략된 목적어의 복원 대상 후보를 선택한다.
(8) 선행어 후보가 문서 내에 존재하지 않으면, 비지시적 대명사(박진호, 2007: 117)로 복원한다.[18)]

18) 비지시적 대명사란, 뭔가를 지시하는 것이 주목적이 아닌 대명사를 일컫는다. 이러한 비지시적 대명사로 의문대명사, 비한정/부정 대명사를 분류하여 제시하였는데, 본 지침에서는 '누구를' 또는 '무엇을'에 해당한다.(박진호(2007),「유형론적 관점에서 본 한국어 대명사 체계의 특징」,『국어학』50, 국어학회, 115-147쪽.)

2. 목적격 무형 대용어 복원을 위한 문어 지침

2.1. 기본 원칙

- 문어에서의 목적격 무형 대용어 복원을 위한 대상 서술어, 선행어 선택의 기본 원칙과 유의점은 다음과 같다.

2.1.1. 대상 서술어

(1) 문장 내 모든 본용언을 탐색 대상으로 선정한다.
(2) 서술어의 격틀을 중심으로 목적어가 생략된 서술어를 탐지한다.
(3) 지정사 구문은 복원 대상 서술어에서 제외한다.
- ‘명사+이다’ 형: 것이다, 학생이다, 예정이다 등

→ 단, 지정사구의 명사는 선행어 후보가 될 수 있다.

(4) 서술성 명사가 서술어 기능을 하더라도 복원 대상 서술어에서 제외한다.
- 서술성 명사: 공부 중, 진행 중 등

→ 단, 서술성 명사는 선행어 후보가 될 수 있다.

(5) 보조 용언과 의사 보조 용언은 복원 대상 서술어에서 제외한다.[19)]
- ‘-(으)ㄹ 수/리(가) 있다/없다’, ‘-(으)ㄴ/ㄹ {것¦터¦뿐¦모양¦참¦노릇¦예정¦길}이다’, ‘-(으)ㄹ {만¦법¦듯}하다’, ‘-(으)ㄴ 말이다’, ‘-(으)ㄴ/ㄹ 것 같다’, ‘-(으)ㄹ 것을(걸) 그랬다’, ‘-아/어서는 (안) 되다’, ‘-고 해서’, ‘-든지 하다’ 등

(가) 시간을 절약하기 위해 미리 만든 김치를 준비했다가 손님들이 준 <u>김치통</u>에 채워 <u>주도록</u> 선배 직원들한테 배웠다.

→ ‘-아/어 주다’에서는 본용언과 보조 용언을 문맥에 따라 구분한다. 위의 문장에서 ‘채워 주다’는 ‘채우다’와 ‘-아/어 주다’의 보조 용언 구문이 아니라, ‘김치를 채우다 + -아/어서 + 김치통을 주다’이다. 따라서 ‘주다’는 목적어가 필요한 본용언이다.

19) 의사 보조 용언은 한국전자통신기술협회(TTA)의 ‘의존 구문 분석 가이드라인’에서 제시하는 목록을 따른다.

→ 서술어 ‘주도록’의 목적어를 ‘김치통’으로 복원한다.

(6) 서술어는 어절 단위로 선택한다.[20]

(가) 대리점에서 고심 끝에 **구입한** 것은 중형 세단이다.
→ 서술어 단위는 어절 단위로 선택하므로, ‘구입한’을 서술어로 선택한다.

(나) 언론사는 “김철수 사장이 백화점에서 **구입했다**”고 보도했다.
→ 서술어 단위는 기본적으로 어절 단위라는 점에서 ‘구입했다”고’를 서술어로 선택하는 것이 맞으나 어절의 외곽에 위치한 기호 및 조사는 삭제한다.

(다) 반발이 나올 것을 우려해 **‘물타기한’** 것이라는 풀이가 나온다.
→ 한 어절 내에 쌍으로 들어 있는 기호와 삽입어 구 등 부가 요소는 한꺼번에 서술어에서 배제한다.
→ “‘물타기한’”은 기호(‘’)는 제외하고, “물타기한”만을 서술어로 선택한다.

(라) 반발이 나올 것을 우려해 **‘물타기’ 한** 것이라는 풀이가 나온다.
→ ‘물타기’와 ‘한’ 사이의 작은따옴표는 삭제할 수 없다. 따라서 앞의 작은따옴표도 유지하여 “‘물타기’ 한”을 모두 서술어로 선택한다.

(마) **홍길동전**/김철수 **지음**·김영희 **옮김**/304쪽·1만3000원/출판사
→ 명사형 전성 어미가 부착된 동사는 복원 대상 술어이다. 단, 부가 요소는 서술어에 포함하지 않는다는 원칙에 따라 ‘지음·김영희’나 ‘옮김/304쪽’과 같이 가운뎃점, 또는 빗금으로 서술어와 다른 품사가 함께 있으면 어절 단위가 아닌 대상 서술어만 선택하여 목적어를 복원한다.
→ 서술어로 ‘지음’, ‘옮김’만 선택하고, 각각의 목적어를 ‘홍길동전’으로 복원한다.

(7) 주로 부사절에 쓰여 독립절을 구성하기 어려운 술어는 복원 대상 서술어에서

20) 단, 어절 외곽의 부가 요소는 서술어에 포함하지 않는다.

제외한다.(관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고 등)

(8) 한 단어지만 띄어쓰기로 나뉜 경우 어절 단위 서술어 태깅에서 예외 처리한다. 복원 대상 서술어인 단어가 '명사+하다'의 구조이고, 그 한 단어 안에 괄호 내용이 들어가 '명사'와 '하다'가 분리되면 마지막 서술어 부분인 '하다'만 서술어로 선택한다.

(가) 회사는 비교적 문맥 속에서 매끄럽게 풀이(모든 과학 개념은 ... 수학적으로 표현하고)**했다**.

→ '풀이했다'라는 서술어를 목적격 무형 대용어 복원 대상 서술어로 삼을 때, 마지막 서술어 부분 '했다'처럼 서술어의 지배소 어절을 복원 대상 서술어로 삼는다. 즉, 삽입구 '표현하고)'를 제외하고 '했다'만을 서술어로 선택한다.

→ 서술어 '했다'의 목적어를 문제, 개념 등과 같은 명사로 복원하고, 다른 선행어 후보가 없을 시 '무엇을'로 복원한다.

(9) 서술어의 띄어쓰기가 잘못된 경우 원문의 오류를 교정하지 않고 생략된 목적어를 복원한다.

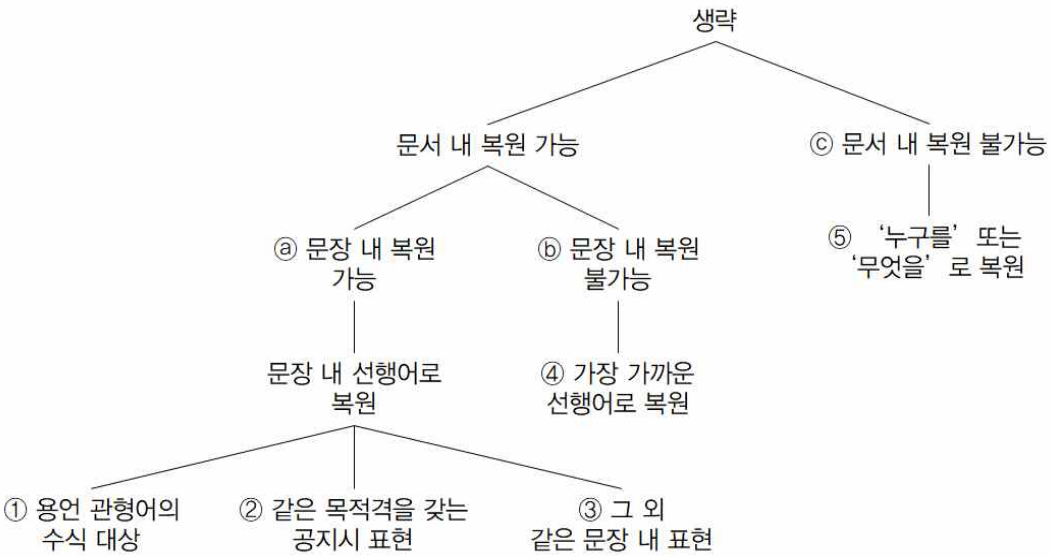
(가) 영화제들은 풍부한 자금력을 바탕으로 스타들을 초청해 **관객**을 **끌어 모으고** 있다.

→ 서술어 '끌어 모으다'는 '끌어모으다'의 띄어쓰기 오류이다. 그러나 어절 단위에 따라 '끌다'와 '모으다'를 각각의 서술어로 보고 격틀에 의해 목적격이 필요하면 목적어를 복원한다.

→ 서술어 '모으고'의 목적어를 '관객'으로 복원한다.

2.1.2. 목적격 무형 대용어 복원

(1) 선행어 선택의 순서



<그림 16> 목적격 무형 대용어 복원 분석의 선행어 선택 순서

- ①→②→③→④→⑤ 순서로 적용한다.
  단, 문장 내 복원이 불가능할 때, 선행어 후보는 선행 표현에서 우선 탐색한다.
- ④ 적용 시 한 문장 내에 복수의 선행어 후보가 있으면, ②→③ 순서로 선행어를 결정한다.
- ⓐ, ⓑ, ⓒ에 대한 세부 순서는 아래와 같다.
- ⓐ - 1) 문장 내 가까운 전방 선행어
- ⓐ - 2) 문장 내 가까운 후방 선행어
- ⓑ - 1) 문장 외 가까운 전방 선행어
- ⓑ - 2) 문장 외 가까운 후방 선행어
- (단, 삽입구(괄호 안에 있는 표현 등)도 선행어 후보가 될 수 있다.)
  ⓒ 문서 전방/후방에 선행어가 존재하지 않는 경우, 비지시적 대명사(누구를, 무엇을)

(2) 목적격 무형 대용어 복원 시 표준국어대사전의 격틀을 참고한다.

<퇴원하다> 【…에서】【…을】 일정 기간 병원에 머물던 환자가 병원에서 나오다. · 그 환자는 수술 후 경과가 좋아 곧 병원에서 퇴원하게 될 것이다. · 병원을 퇴원하면 곧 다시 전선으로 배치되나요?

(가) 이와 함께 **중환자실**에 함께 있다가 **퇴원한** 4명 가운데 1명이 감기 증상을, 다른 병원으로 옮겨진 8명 가운데 1명이 기력저하를 보여 관찰 중이다.

→ 목적격과 부사격을 바꿔 쓸 수 있는 술어에서 해당 논항이 충족되지 않으면 목적격이 생략된 것으로 간주하고 목적어를 복원한다.

→ 서술어 '퇴원한'의 목적어를 '중환자실'로 복원한다.

(3) 의미 부류가 맞지 않더라도 목적격 조사가 실현되어 격틀이 충족되면 목적어 복원 대상이 아니다.

(가) 김철수는 유명 인사들에게 보낸 편지에서 "편한 시간에 미술관을 방문하기를 원하며 미술관 방문을 **초청한다**" 고 밝혔다.

→ '초청하다'의 격틀은 [누구-을/를 어디-에]를 필요로 한다. 그러나 사전의 의미 부류와 맞지 않더라도 […을] 격틀이 표면적으로 실현되었으므로 목적어 복원 대상이 아니다.

(나) 나는 학교까지 다섯 시간을 **걸었다.**

→ 이 문장에서도 '걷다'의 목적어는 '길¦산¦거리' 등이 적절하지만 '다섯 시간을'이 목적격으로 실현되었으므로 서술어 '걸었다'는 목적어 복원 대상이 아니다.

(다) 나는 학교까지 다섯 시간 **걸었다**.

→ 다만, 목적격 조사 […을]로 실현되지 않을 경우, 시간, 거리, 횟수 등은 부사어로 분석하고, 목적어가 필요하면 복원한다.

→ 서술어 '걸었다'의 목적어를 길, 거리 등과 같은 명사로 복원하고, 다른 선행어 후보가 없을 시 '무엇을'로 복원한다.

(라) 노동자들의 삶을 담기 위해 2년에 걸쳐 5번을 **방문했다**.

→ 위의 예문에서도 격 표지가 목적어 복원의 기준이 된다. 목적격 조사가 실현된 '5번을'이 출현했으므로 서술어 '방문했다'는 목적어 복원 대상이 아니다.

(마) 노동자들의 삶을 담기 위해 2년에 걸쳐 5번 **방문했다**.

→ […을]이 실현되지 않은 문장에서는 서술어 '방문했다'의 목적어를 복원한다.[21)]

→ 서술어 '방문했다'의 목적어를 사무실, 현장 등과 같은 명사로 복원하고, 다른 선행어 후보가 없을 시 '무엇을'로 복원한다.

(4) 동사에서 서술성 명사만 분리하여 선행어 후보로 삼지 않는다.

(가) 신문은 또, 대응에 '모든 옵션'이 고려되고 있다고 여러 차례 언급했다며, 이 옵션에 교육적 정책이 포함된다는 뜻으로 **받아들였다고** 한다.

→ '받아들였다고'의 목적어로 '언급'이 가능하더라도 '언급하다'의 어근만 분리하여 선행어 후보로 삼지 않는다.

→ 서술어 '받아들였다고'의 목적어를 '언급', '의견' 등과 같은 명사로 복원하고, 다른 선행어 후보가 없을 시 '무엇을'로 복원한다.

(5) 기호 및 괄호 제외

(가) 김영희 씨가 **거론한** **김철수씨**(20대), 공개수배 전환

→ 서술어 '거론한'의 생략된 목적어 복원 시 부가 요소인 '(20대)'와 쉼표를 제외한다.

21) 구문 분석 말뭉치 분석 지침에서는 '시간, 거리, 수량'의 경우 목적격 조사 이외의 다른 격 조사가 결합할 수 없는 경우에는 목적격 조사가 생략되어 있어도 목적어로 분석하고 있다. 무형 대용어 복원에서는 목적격 조사의 출현 여부를 우선적으로 복원 기준으로 삼았다는 점이 다르다.

2.2. 세부 유형별 가이드라인

2.2.1. 관형절의 복원

(1) 관형절에 속하는 피수식어(후행하는 NP)로 우선하여 복원하는 것을 원칙으로 한다.

(2) 관형절의 피수식어가 의존 명사 ‘것’일 경우, ‘것’ 앞에 지시 관형사를 넣었을 때 자연스러우면 ‘것’으로 복원한다.[22)]

(가) 이 빵집은 쌀가루로 **만든 것**을 주로 판다.

→ ‘만든 (그) 것’이 자연스러우므로 수식을 받는 의존 명사로 목적어를 복원한다.

→ 지시 관형사를 삽입하였을 때 문맥이 자연스러우면 수식을 받는 의존 명사로 목적어를 복원한다. 이 문장에서의 ‘것’은 ‘빵’을 대체할 수 있다.

→ 서술어 ‘만든’의 목적어를 ‘것’으로 복원한다.

(나) 2000년 개인전에서 **선보였던 것**을 다시 대중에 공개하였다.

→ ‘선보였던 (그) 것’이 자연스러우므로 수식을 받는 의존 명사로 목적어를 복원한다.

→ 지시 관형사를 삽입하였을 때 문맥이 자연스러우면 수식을 받는 의존 명사로 목적어를 복원한다. 이 문장에서의 ‘것’은 ‘작품’을 대체할 수 있다.

→ 서술어 ‘선보였던’의 목적어를 ‘것’으로 복원한다.

---

22) 국립국어원의 2019년 주격 무형 대용어 복원 말뭉치 구축과 일관된 규칙을 적용하였다.

① 세포벽은 식물 세포의 가장 바깥층을 에워싸고 있는 약간 두꺼운 막이다.

구문 분석) 세포벽은 → NP_SBJ 막이다.

에워싸고: 주격 복원 대상 서술어

선행어 후보: ‘세포벽’ 또는 ‘막’

- 답: <막이> 에워싸고

(약간 두꺼운 막이) 식물 세포의 가장 바깥층을 에워싸고 있다.

② 멜라닌은 사람의 피부색을 결정하는 주요 요소이다.

구문 분석) 멜라닌은 → NP_SBJ 요소이다.

결정하는: 주격 복원 대상 서술어

선행어 후보: ‘멜라닌’ 또는 ‘요소’

- 답: <요소가> 결정하는

(주요 요소가) 사람의 피부색을 결정한다.

(다) 아는 것이 없다.

→ ‘아는 (그) 것’이 부자연스러우므로 수식을 받는 의존 명사로 목적어를 복원할 수 없다.

→ 지시 관형사를 삽입하였을 때 문맥이 부자연스러우면 수식을 받는 의존 명사를 선행어로 삼지 않는다. 이 문장에서의 ‘것’은 구체적인 의미를 지니고 있지 않다.

→ 서술어 ‘아는’의 목적어를 ‘것’으로 복원하지 않고, ‘무엇을’로 복원한다.

2.2.2. 관형절에 후행하는 복합 명사 구문의 복원

(1) 복합 명사 구문의 구조가 중의적인 경우, 인접 어절 간의 의존 관계를 기준으로 무형 대용어 목적격을 복원한다. (‘국립국어원 구문 분석 지침’에 따름)

(가) 은행이 제시한 자영업자 대출총액 추산치 480조2천억원

→ 서술어 ‘제시한’의 목적어를 ‘추산치’로 복원한다.

(나) 다빈치가 프랑스로 가져간 작품 3점 중 하나가 모나리자였다.

→ 서술어 ‘가져간’의 목적어를 ‘작품’으로 복원한다.

2.2.3. 정의문의 복원

(1) 본 지침에서 정의문은 어떤 용어의 뜻을 규정하는 문장을 의미한다. 주로 ‘A는 -한 B이다’ 유형이 이에 속한다. ‘국립국어원 구문 분석 지침’에서 ‘모문과 내포문의 주어가 같고, 서술어가 다른 경우’의 분석과 관련된다.[23]

(2) 목적격 무형 대용어 복원에서는 정의문 유형의 목적격 복원이 관형절 복원의 원칙과 관련이 깊다. 따라서 2.2.2와 2.2.3의 관형절 복원의 원칙과 동일하게 복원한다. 따라서 (2)의 예에서 ‘A’가 주어인 경우에는 관형절 복원 원칙에 따라 목적어를 복원한다.[24]

---

23) 정의문에서는 관형절의 피수식어를 우선하여 복원하되, 그 의미가 맞지 않는다면 다른 선행어를 찾는다. 예를 들어, 2019년 주격 무형 대용어 복원 말뭉치에서와 같이 ‘김철수는 금메달을 딴 주인공이다’라는 문장에서 ‘딴’의 주어로 ‘주인공’을 복원할 수 없다. ‘주인공’이 금메달을 딴 것이 아니라, 금메달을 땄기 때문에 ‘주인공’이 된 것이다. 이렇듯, 의미적으로 무관한 경우가 아니라면 관형절 안에서 선행어를 찾는다.

24) 의존 명사의 복원에 관해서는 ‘국립국어원 구문 분석 세부 지침’을 참고한다. <아래: 관련 내용>

(가) 이번 계획은 승인 권한이 중앙 정부에서 지자체로 넘어가고 나서 처음 **수립하는 것**이다.

→ 구문 분석) 계획은 →NP_SBJ 것이다.

→ '것'은 '계획'과 동일하므로, 서술어 '수립하는'의 목적어를 '것'으로 복원한다.

(나) 경찰은 "이 기술은 한 회사에서 1조1000억원의 연구비를 들여 **개발한 것**"이라고 설명했다.

→ 구문 분석) 기술은 →NP_SBJ 것"이라고

→ '것'은 '기술'과 동일하므로, 서술어 '개발한'의 목적어를 '것'으로 복원한다.

(다) 직업교육센터는 발달 장애 학생들을 위해 **설립하는** 직업교육·훈련 **기관**이다.

→ 구문 분석) 직업교육센터는 →NP_SBJ 기관이다.

→ '직업교육·훈련 기관'은 '직업교육센터'와 동일하므로, 서술어 '설립하는'의 목적어를 '기관'으로 복원한다.

표면적으로는 아래에 제시된 의사 보조 용언 구성에 해당되더라도, 의존 명사나 명사가 서법을 나타내지 않는 경우는 의사 보조 용언 구성에 해당되지 않는다.

(가) 이 사료는 가축들이 먹는 것이다. (것 = 사료, 식품)
사료는 → NP_SBJ 것이다.
가축들이 → NP_SBJ 먹는
먹는 → VP_MOD 것이다.
것이다. → VNP ROOT

(나) 모나리자는 레오나르도 다빈치가 그린 것이다. (것 = 그림, 모나리자∈그림)
모나리자는 → NP_SBJ 것이다.

(다) 철수가 천재라는 것이다. (것 ≒ 철수)
철수가 → NP_SBJ 천재라는
천재라는 → VNP_MOD 것이다.
것이다. → VNP ROOT

(라) 내가 놀란 것은 철수가 천재라는 것이다. (것 = 사실 → 사실 = 사실)
것은 → NP_SBJ 것이다.
철수가 → NP_SBJ 천재라는

(마) 변함없는 사실은 철수가 천재라는 것이다. (것 = 사실)
사실은 → NP_SBJ 것이다.

즉, '것'을 주로 주어와 동일하거나 포함 관계인 내용 명사로 대체하지 못할 때 의사 보조 용언 구성을 적용한다.

2.2.4. 자타 양용 동사의 복원

(1) 자타 양용 동사 중 이동 동사에 해당하는 '가다/오다/다니다' 등에 대하여, 문맥에 따라 타동사적 기능을 결정한다.[25)]

(가) 학생이 빨리 <u>간다</u>.
→ '가다'의 격틀은 […에/에게][…으로][…을]이다. 생략된 목적어가 '학교'인 경우, '학교에 가다'와 '학교를 가다' 모두 치환이 가능하면 서술어 '가다'를 자동사로 보고, 목적어 복원 대상으로 삼지 않는다.
→ 생략된 것은 부사격으로 목적격이 아니다. 단, 아래와 같이 맥락에 따라 타동사로 쓰였을 경우에는 목적어를 복원한다.

(나) 조카가 벌써 <u>갈</u> 나이가 되었나?
→ 맥락상 '시집/장가를 가다'로 해석된다면 "조카가 벌써 {시집을|장가를} 갈 나이가 되었나?"로 복원해야 한다. 문서 내에서 복원할 대상이 없다면 목적어를 '무엇을'로 복원한다.

(다) 조카가 벌써 ({시집을|장가를}) <u>갈</u> 나이가 되었나?
→ 서술어 '갈'의 목적어를 '시집' 또는 '장가'로 복원한다.

(라) 조카가 벌써 ({학교에|군대에}) <u>갈</u> 나이가 되었나?
→ 서술어 '갈'은 목적어 복원 대상이 아니다.

(2) 자타 양용 동사 중 '화답하다'의 격틀 […을]은 '시와 노래'에 제한된 쓰임을 보인다. 일반적인 문장에서는 서술어 '화답하다'의 목적어를 찾기 어렵다.

(가) 시민들은 표지석을 세워 줄 것을 요청했고 시청은 당초 요구보다 더 크게 <u>**화답했다**</u>.
→ '화답하다'는 […에/에게][…을] 격틀을 가지나, […을]의 쓰임은 많지 않

25) 자타 양용 동사의 경우 […에], […으로], […와] 등 부사격 조사로 대체가 되는 경우 자동사로 본다. 반대로 대체가 되지 않을 때에는 타동사로 보고 복원 대상 서술어 후보가 될 수 있다.
(가) 조카가 시집을 갈 나이가 되었다.
→ 서술어 '갈' 앞의 '시집을'은 부사격 조사로 치환이 되지 않기 때문에 '가다'를 타동사로 본다.

다. 이와 같이 […을]의 쓰임이 어색하면 목적어 복원 대상 술어가 아닌 것으로 간주한다.

2.2.5. ‘-게 하다’와 ‘-도록 만들다/하다’, ‘-게 만들다’ 구문의 복원

(1) 장형 사동 구문 ‘-게 하다’에서 ‘하다’는 보조 용언이므로 목적격 복원 대상 서술어가 아니다.[26]

(2) ‘-도록 만들다/하다’의 ‘만들다’, ‘-게 만들다’의 ‘만들다’는 보조 용언이 아니므로 목적격 복원 대상 서술어 후보가 될 수 있다.

(가) 김철수가 상대방을 **압박하도록** **만든다는** 설명이다.
→ 구문 분석) 김철수가 → NP_SBJ 압박하도록
상대방을 → NP_OBJ 압박하도록
→ 주어 ‘김철수가’는 선행하는 부사절에 의존하므로 ‘만든다’는 복원 대상 서술어이다.
→ 서술어 ‘만든다는’의 목적어를 ‘김철수’로 복원한다.

(나) 경쟁사가 노면 소음 저감 **기술**이 자동차의 소음을 **차단하게** **만들었다**.
→ 구문 분석) 소음을 → NP_OBJ 차단하게
→ ‘경쟁사가 <기술을> [기술이 자동차의 소음을 차단하게] 만들었다.’로 분석할 수 있다.
→ 서술어 ‘만들었다’의 목적어를 ‘기술’로 복원한다.

(다) **구름**이 눈을 **토해내도록** **하는** 것이다.
→ 구문 분석) 눈을 → NP_OBJ 토해내도록
→ ‘무언가가 <구름을> [구름이 눈을 토해내도록] 하다.’로 분석할 수 있다.
→ 서술어 ‘하는’의 목적어를 ‘구름’으로 복원한다.

---

26) ‘-게 하다’ 구문은 사동의 의미와 방법/방식의 의미로 나뉜다. 문어에서는 보조 용언 ‘하다’가 사동의 의미로 쓰인 ‘-게 하다’ 구문이 출현 빈도가 높은 데 반해, <표준국어대사전>의 ‘하다’ 의미 번호 2, 8, 9와 같이 방식의 의미로서 [-게] 격틀을 요구하는 ‘-게 하다’ 구문은 출현 빈도가 낮다. 이때의 ‘-게 하다’는 방법/방식을 나타내는 ‘어떻게, 이렇게, 그렇게’ 등의 부사어를 많이 취하는데, <표준국어대사전>의 ‘하다’ 의미 번호 8, 9번의 의미로 특별한 선행어가 없거나 목적격 복원을 하면 어색한 문장에서는 목적어를 복원하지 않는다. 자세한 내용은 ‘3.2.6. ‘-게 하다’ 구문의 복원’을 참고할 수 있다.

(라) **상대**를 꼼짝 못 **하게 만들다**.

→ 구문 분석) 꼼짝 → AP 하게

상대를 → NP_OBJ 만들다.

→ 서술어 '만들다'는 목적어 복원 대상이 아니다.

(마) **혈압**을 **올라가게 만들다**.

→ 구문 분석) 혈압을 → NP_OBJ 만들다.

→ 서술어 '만들다'는 목적어 복원 대상이 아니다.

(바) **부하들**을 명령에 **복종하도록 만들다**.

→ 구문 분석) 부하들을 → NP_OBJ 만들다.

→ 서술어 '만들다'는 목적어 복원 대상이 아니다.

2.2.6. 목적격 명사구가 다른 격으로 실현된 경우

(1) 사전의 격틀에 따라 원래 목적격으로 실현되어야 하는 단어가 다른 격으로 표현된 예가 다수 존재한다. 이 경우에 해당하는 문장은 일차적으로 사전 격틀을 기준으로 문서 내 목적격 무형 대용어로 복원 가능한 명사구가 있는지 탐색하여 목적어를 복원한다.

(가) 8번의 연습경기에서 모두 **패한** 가장 큰 이유는 공격 때문이 아니었다. 8판에서 모두 뒷문이 허술했기 때문이다.

→ '패하다' 격틀은 […에/에게][…을]이다. 서술어 '패한'의 생략된 목적어를 '연습경기'로 볼 수 있지만, '연습경기'는 이 문장에서 부사어로 실현되었으므로 다른 명사로 목적어를 복원해야 한다.

→ 이 경우, 문서 내에 '8판'과 같은 다른 표현이 존재하면, 서술어 '패한'의 목적어를 '8판'으로 복원한다.

(2) 목적격 명사구가 다른 격으로 실현된 경우, (1)의 예처럼 문서 내에 대체 가능한 명사구가 없다면 목적어를 비지시적 대명사 '누구를', '무엇을'로 복원한다.

(가) 더 이상 이 문제로 **질의한** 적이 없다.

→ ‘질의하다’ 격틀은 […에/에게 …을][…에/에게 -ㄴ지를][…에/에게 -고] ((‘…을’ 대신에 ‘…에 대하여’가 쓰이기도 한다))이다. 서술어 ‘질의한’의 부사어인 ‘문제로’를 제외한 다른 선행어 후보가 문서 내에 없다면 목적어를 비지시적 대명사로 복원한다.

(나) 아들이 아버지와 서로 **껴안다**.

→ ‘껴안다’ 격틀은 […을]이다. 서술어 ‘껴안다’의 부사어인 ‘아버지와’와 ‘서로’를 제외한 다른 선행어 후보가 문서 내에 없다면 목적어를 비지시적 대명사로 복원한다.

(다) 첫 질문은 “왜 이 영화에 **천착하느냐**”는 것이다.

→ ‘천착하다’ 격틀은 […을]이다. 서술어 ‘천착하느냐’의 부사어인 ‘영화에’를 제외한 다른 선행어 후보가 문서 내에 없다면 목적어를 비지시적 대명사로 복원한다.

2.2.7. 한정 명사구에서의 선행어 선택

(1) 생략된 목적어와 한정 명사구의 의미가 동일한 경우, ‘수식 표현+명사구’, ‘고유 명사+명사구’, ‘지시사+명사구’ 등 한정 명사구의 일부인 후행 명사구를 목적격 무형 대용어로 복원할 수 있다.

(가) 구단은 김 **감독**을 경질하면서 더 이상 **문책하지** 않기로 했다.

→ 위 문장에서 ‘문책하지’의 대상은 ‘김 감독’이다. 문서 내에서 ‘김 감독’, ‘김영희 감독’, ‘그 감독’과 같이 동일한 대상을 가리키는 ‘감독’을 ‘문책하지’의 목적어로 복원할 수 있다.

→ 예시에서는 서술어 ‘문책하지’의 목적어를 선행절의 ‘감독’으로 복원한다.

(2) 생략된 목적어를 복원하는 선행어의 의미는 해당 어절을 선행하는 수식어를 포함한다. 예를 들어, 서술어의 목적어로 특정인의 ‘직위/직책명’이 아닌 일반적 의미의 ‘직위/직책명’을 선택해야 할 경우, ‘고유 명사+직위/직책명’

내의 ‘직위/직책명’만 목적어로 선택할 수 없다.

(가) 대통령은 김철수 전 대변인과 관련해 “전문성이 있는 인물에게 **맡기자**”고 했다.
→ 위 문장에서 ‘맡기자’의 목적어는 ‘김철수 전 대변인’이 아니라 ‘대통령 대변인/업무/직책’ 등에 해당한다.
→ 같은 문장 내 ‘대변인’은 ‘김철수 전 대변인’을 의미하나, ‘맡기자’의 목적어는 일반적인 대변인을 지칭하는 명사로 복원해야 한다. 이에 해당하는 표현이 문서 내에 없다면 ‘무엇을’로 복원한다.

2.2.8. 미등재어의 처리

(1) 사전에 없는 서술어는 <우리말샘>의 문법 및 어휘 정보를 참고한다.
(2) <표준국어대사전>과 <우리말샘>에도 등재되어 있지 않은 어휘는 아래와 같이 복원한다.

(2-1) 두 서술어의 결합형

(가) **선수들**을 보자 차에서 내린 뒤 일일이 **악수·격려하며** 기념사진까지 찍었다.
→ ‘악수·격려하다’, ‘격려·악수하다’ 상관없이 둘 중 하나가 […을] 격틀을 가지는 서술어라면 어절 단위로 분석하여 전체를 대상 서술어로 볼 수 있다. 따라서 하나의 서술어라도 목적어가 필요하면 복원한다.
→ 서술어 ‘악수·격려하며’의 목적어를 ‘선수들’로 복원한다.

(2-2) ‘NP+받다/NP+시키다’ 결합형
- 사전에 등재되어 있지 않아도 목적어가 생략된 것으로 보인다면 목적격 무형대용어 복원 대상이 된다.
- ‘NP+받다/NP+시키다’ 결합형의 목적어는 아래와 같이 복원한다.

(가) **암**에 걸려 **치료받은** 환자 중 완치된 사람들은 주치의를 믿고 긍정적인 생각을 가졌던 사람이 많다.
→ 서술어 ‘치료받은’의 목적어를 ‘암’으로 복원한다.

(나) 아버지가 어떻게 해서든 **설득시키고야** 말았다.

→ 서술어 ‘설득시키고야’의 목적어는 설득시키는 대상(어머니, 나 등)의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘누구를’로 복원한다.

(2-3) 형용사의 동사형 ‘감정 형용사+-아/어하다’

- ‘즐거워하다’, ‘아쉬워하다’와 같이 사전에 등재되어 있으면 그 격틀에 따라 복원 여부를 결정한다.
- 사전에 미등재되어 있는 단어는 복원할 대상이 있다고 판단되면 아래와 같이 최대한 복원하는 방향으로 한다.

(가) 철수는 영희의 질타에 **부끄러워해야** 한다.

→ 서술어 ‘부끄러워해야’의 목적어는 잘못, 실수 등의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘무엇을’로 복원한다.

(나) 축구 팬들은 세계적인 스타들이 펼치는 축구 향연에 **즐거워했다**.[27]

→ 서술어 ‘즐거워했다’의 선행어로 같은 절 내의 부사어인 ‘향연에’ 외에 다른 후보가 없다면 ‘무엇을’로 복원한다.

2.2.9. 괄호의 처리

(1) 원문에서 괄호에 싸인 삽입 어구에 목적격 조사가 부착되어 있으면 구문 분석에서 목적어로 분석한다. 따라서 해당 문장의 술어는 목적어 복원 대상이 아니다.

(가) 앞으로 인사위원회를 통해 좀 더 철저하게 검증하고 제도적으로 (인사 시스템을) **보완하겠다**.

→ 구문 분석)　　시스템을)　　→ NP_OBJ　보완하겠다.

→ 서술어 ‘보완하겠다’는 목적어 복원 대상이 아니다.

---

27) 무형 대용어 목적격 복원에서는 ‘축구 향연을 즐거워하다’가 아닌 ‘축구 향연에 즐거워하다’와 같이 서술어 ‘즐거워하다’의 목적어가 다른 격으로 실현될 때, 생략된 목적어 ‘향연’을 ‘향연에 향연을 즐거워하다’와 같이 중복되게 복원하지 않고 비지시적 대명사로 복원한다. 이와 관련하여 ‘2.2.6. 목적격 명사구가 다른 격으로 실현된 경우’를 참고할 수 있다.

2.2.10. 의미상 목적어로 해석되는 부사구(AP)의 처리

(1) ‘서로’, ‘스스로’, ‘매일’, ‘모두’ 등과 같이 명사와 부사의 품사가 통용되는 단어는 격 조사가 없으면 구문 분석에서 ‘AP(부사구)’로 분석한다. 이러한 부사구가 의미적으로 목적격 격틀을 만족하더라도 문장 내 OBJ가 없으면 목적어를 복원한다.

(가) 이 문제에 대해서 오늘도 서로 **비난하며** 충돌했다.
→ “서로를 비난하다”라고도 해석할 수 있지만, ‘서로’가 AP(부사구)로 분석되는 구문 분석 결과와 일치하도록 다른 선행어를 목적어로 복원해야 한다.
→ 이미 AP로 비난하는 상대를 밝혔으므로, 서술어 ‘비난하며’의 목적어를 비난하는 잘못, 결점 등의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘무엇을’로 복원한다.

(나) 경기를 치르면서 선수들이 서로 **격려하고** **믿는** 마음이 많이 생겼다.
→ “서로를 격려하다”라고도 해석할 수 있지만, ‘서로’가 AP(부사구)로 분석되는 구문 분석 결과와 일치하도록 다른 선행어를 목적어로 복원해야 한다.
→ 서술어 ‘격려하고’, ‘믿는’의 목적어를 격려하는 대상, 상대방 등의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘누구를’로 복원한다.

2.2.11. ‘-ㄴ지/-ㄹ지+서술어’ 구문의 복원

(1) ‘-ㄴ지/-ㄹ지+서술어’(‘-ㄴ지/-ㄹ지’+확인하다/생각하다/고민하다/이해하다 등)의 구문에서는 ‘-ㄴ지/-ㄹ지’에 격 조사가 결합되지 않더라도 목적어(절)로 분석한다. 구문 분석의 결과에 따라 ‘-ㄴ지/-ㄹ지’가 의존하는 서술어는 목적격 복원 대상이 아니다.[28]

(가) 하찮은 재료를 긁어모아 엮을 때 얼마나 아름다운 꽃이 **피어나는지** 관객들

28) ‘국립국어원 구문 분석 지침’에 따르면, 주어, 목적어, 보어의 경우 조사가 생략되거나 보조사로 표지된 명사구 또는 이에 상응하는 용언구와 절도 서술어 구문 틀에 따라 격 조사로 대치가 가능하면 해당 기능 태그로 분석한다. 이에 따라 해당 서술어에 OBJ가 의존되어 있으면 목적격 복원 대상 술어로 보지 않는다.

과 함께 **확인하고** 싶어요.

→ 구문 분석)　　피어나는지　→ VP_OBJ　확인하고

→ 서술어 ‘확인하고’는 목적격 복원 대상이 서술어가 아니다.

2.2.12. ‘비추다/견주다/맞추다’류의 복원

(1) ‘비추다/견주다/맞추다’의 활용형인 ‘-에 비춰/견줘/맞춰’ 등은 관용구로 자주 쓰이나 격틀에 맞추어 목적어를 복원하도록 한다.

(가) “내 경험에 **비추어** 볼 때 이 **사업**은 성공하기가 어렵다.”

→ ‘이 사업을 경험에 비추다’로 해석할 수 있다.

→ 서술어 ‘비추어’의 목적어를 ‘사업’으로 복원한다.

2.2.13. 띄어쓰기 오류

(1) 문어에서는 띄어쓰기 오류에 해당되더라도 어절 단위를 기준으로 복원 대상 서술어를 선택한다. 또한 띄어쓰기 된 서술어의 격틀에 맞게 목적어를 복원한다.

(가) 공장이 들어설 것이라는 소문에 온 마을 사람이 **들고 일어났다**.

→ 위 예문의 ‘들고 일어났다’는 ‘들고일어나다’의 띄어쓰기 오류이며, ‘들고일어나다’는 격이 필요 없는 자동사이다. 그러나 띄어 쓴 원문을 기준으로 ‘들고’와 ‘일어났다’를 각각 서술어로 선택한다.

→ 서술어 ‘들고’는 ‘마을 사람이 반발/반기를 들다’로 해석한다.

→ 서술어 ‘들고’의 목적어를 반기 등과 같은 명사로 복원하고, 다른 선행어 후보가 없을 시 ‘무엇을’로 복원한다.

→ ‘일어나다’는 목적격을 필요로 하지 않으므로 서술어 ‘일어났다’는 목적어 복원 대상이 아니다.

(나) 검색과 지도는 중소기업이 **따라하기** 힘든 **서비스**다.

→ ‘따라하다’가 사전에 등재되어 있지 않으며, 올바른 띄어쓰기는 ‘따라(서) 하기’이다. 그러나 어절 단위를 기준으로 하여 ‘따라하기’를 서술

어로 선택하여 목적어를 복원한다.

→ 서술어 ‘따라하기’의 목적어를 ‘서비스’로 복원한다.

(다) 지난해 200% **생산량**을 올린 만큼 다시 큰 폭으로 올리 **는** 것은 무리라고 생각한다.

→ 구문 분석) 폭으로 →NP_AJT 는

→ ‘는’을 복원 대상 서술어로 보고 목적어를 복원한다.

→ 서술어 ‘는’의 목적어를 ‘생산량’으로 복원한다.[29)]

(2) 단, 두 개 이상의 품사가 결합되어 있는 경우 서술어에 해당하는 부분만 선택하여 목적어를 복원한다.

(가) 달걀, **설탕**을 중탕으로 섞다가 거품이 올라오면 얼음물에**식히면서** 휘핑한다.

→ 부사어와 서술어의 결합은 품사가 다르므로 서술어 부분만 선택한다.

→ 서술어 ‘식히면서’의 목적어를 ‘설탕’으로 복원한다.

(나) 삭제된 두 연에는 “말탄섬나라 사람이,/ **길**을뭇고**지남**이 이상한일이다.”와 같은 대목이 들어 있다.

→ 목적어와 서술어의 결합 중 복원 대상이 필요한 서술어인 ‘지남’만 선택한다.

→ 서술어 ‘지남’의 목적어를 ‘길’로 복원한다.

(다) 정말 **할말** 없어

→ 서술어와 주어의 결합은 품사가 다르므로 서술어 부분만 선택한다.

→ 서술어 ‘할’의 목적어를 ‘말’로 복원한다.

2.2.14. 전문 용어의 복원

(1) <표준국어대사전>, <우리말샘> 등에 등재된 전문 용어로 사용된 경우 사전의

29) 구어 말뭉치에서는 이 경우 ‘올리 는’을 모두 서술어로 선택한다. 단, ‘올리’ 또는 ‘는’ 중 하나라도 <trunc></trunc>인 경우에는 모두 복원 대상 서술어에서 제외한다. 관련 내용은 ‘3.2.3. 문어와 다른 복원 대상 서술어의 처리’를 참고한다.

정보를 따라 목적어 복원 대상 여부를 결정한다.

(가) **옮김**

→ [물리]의 전문 용어 '옮김'은 <우리말샘>에서 명사로 등재되어 있다. 물리 용어에서 명백히 명사로 사용되므로 대상 술어에서 제외한다. 반면에, 역저를 표기하는 'OOO 옮김'의 경우에는 [물리] 용어로 분석할 수 없다. 이때 '-ㅁ'은 명사형 전성어미로 구문 분석되므로 동사 '옮기다'의 사전 격틀에 따라 '옮김'을 목적어 복원 대상 서술어로 삼는다.

(나) **알** 권리

→ '알 권리'의 경우 <우리말샘>에서 명사구로 등재되어 있다.

→ 구문 분석) 알 → VP_MOD 권리

→ 따라서 목적어 복원 대상 서술어로 삼는다.

2.2.15. 발화자에 따른 인칭 및 지칭 표현의 선택

(1) 인용문에서는 발화자와 복원 대상 명사구의 관계를 고려하여야 한다.

(2) 동일 대상을 지시하는 명사구가 존재하더라도 인용문에서는 인칭을 고려하여 생략된 목적어를 복원한다.

(가) 김철수는 "팬들이 **나**를 많이 좋아해준다. **좋아해주는** 팬들에게 감사하다."라고 말했다.

→ '김철수'와 인용문 내의 '나'는 동일한 인물이다. 인용문 내에서는 목적어의 인칭을 고려하여 '좋아해주는'의 목적어를 '나'로 복원한다.

3. 목적격 무형 대용어 복원을 위한 구어 지침

3.1. 기본 원칙

(1) 문어 말뭉치와 다른 구어 말뭉치의 특성을 중심으로 작업 지침을 기술한다.
(2) 구어 특성상 끼어들기, 동시 발화 등이 발생한다는 점에서 문어와 다른 문장 단위를 설정하며, 하나의 발화 단위(<u></u>) 안에서는 종결 부호로 문장이 나뉘더라도 한 문장으로 처리한다.
(3) 단, 인용문은 분할하여 처리한다.
(4) 선행어 후보가 이중 목적어인 경우, 의미상 실질적인 명사구로 목적어를 복원한다.

(가) P1-1: **떡볶이**는 **양념 먼저** 만드나요?
P2-1: 어 네 **만들어야** 해요.
P2-2: 양념 먼저 만들어야 해요.
→ P2-1의 '만들어야'의 선행어 후보로 '떡볶이'와 '양념'이 있다. 이 중 의미상 실직적인 목적어인 '양념'으로 복원한다.
→ P2-1의 서술어 '만들어야'의 목적어를 '양념'으로 복원한다.

3.1.1. 대상 서술어

- 목적격 무형 대용어 복원을 위한 문어 지침인 2.1.1.에서 (8)과 (9)를 제외한 나머지 지침을 동일하게 적용한다. 띄어쓰기에 관한 구어 지침은 '3.2. 세부 유형별 가이드라인'에 명시한다.

3.1.2. 목적격 무형 대용어 복원

- '2.1.2.의 목적격 복원을 위한 문어 지침'과 동일하게 적용한다.

3.2. 세부 유형별 가이드라인

3.2.1. 복원 단위의 기준

(1) 마침표/느낌표/물음표 등 종결 부호로 끝나는 범위까지를 하나의 문장 단위로 보지 않고 발화 단위(<u></u>)를 기준으로 복원한다.[30][31]

(가) P1-1: 한 주의 **이슈**를
P1-2: 시원하게 **풀어보겠습니다.**
→ 위의 P1-1과 P1-2를 하나의 문장으로 합쳐서 분석하지 않고, 각각의 문장 단위로 처리한다.
→ P1-2의 서술어 '풀어보겠습니다'의 목적어를 '이슈'로 복원한다.

(2) 단, 아래와 같이 간투사는 마침표가 있더라도 하나의 분석 단위로 간주한다.

(가) 송편을. 어. 어. 어. **먹었어**.
→ 위의 문장은 간투사가 삽입된 하나의 문장으로, 서술어 '먹었어'는 목적어 복원 대상이 아니다.

(3) 휴지에 의해 마침표가 삽입되어 있더라도 하나의 문장으로 간주하여 의존 관계를 파악하고 생략된 목적어를 복원한다.

(가) 커피는 빵을 먹은 다음에. **마시긴** 했지.
→ 서술어 '마시긴'은 '커피는'을 목적어로 가지므로 목적어 복원 대상이 아니다.

---

30) 문어 말뭉치에는 종결 부호 기준보다 더 작은 단위로 문장이 나뉘어 있는 경우가 없다. (1)은 구어 말뭉치에 나타나는 문장 분할 특징이다. 구어에서는 종결 부호가 아니라 원시 말뭉치에서 주어진 발화 단위(<u></u>)를 분석 단위로 삼아 목적어 복원을 실시한다.

31) 무형 대용어 복원은 '국립국어원 구문 분석 지침'과의 일관성을 위해 구어의 경우 여러 문장이 하나의 주석 단위로 구분된 경우, 별개의 문장 단위로 분석하지 않는다. 이는 '국립국어원 구문 분석 지침' 가운데 '여러 문장이 한 주석 단위에 포함되는 경우' 분석 사례를 참고할 수 있으며, 이에 관한 내용은 다음과 같다.
- 구어 자료에서는 하나의 주석 단위를 크게 한 문장과 동등하게 한다. 즉, 한 주석 단위 안에 마침표가 있더라도 별개의 문장으로 보지 않는다. 다만, 별개의 문장임이 확실한 경우에는 문어와 마찬가지로 한 문장의 최상위 지배소를 다음 문장의 최상위 지배소에 연결한다.

3.2.2. 인용 표지가 없는 인용문의 처리

(1) 구어 특성상 인용 표지 ‘이라고, 라고, 다고’ 등이 없지만 인용문으로 해석되면, ‘이라고/라고/다고+하다’ 형으로 분석한다.[32)]

(2) 문어의 목적격 무형 대용어 복원 지침과 일치하게 복원한다. ‘이라고/라고/다고+하다’ 형은 사전에서 인용격 조사와 결합하는 격틀 구조를 구성하여 목적격 복원 대상에서 제외한다.[33)]

(가) 이 사람이 갑자기 돌발 발언을 하면 어떡하지 **하면서**.
→ “‘~어떡하지’라고 하면서”로 해석한다.
→ 구문 분석) 어떡하지 → VP_CMP 하면서
→ 서술어 ‘하면서’는 목적어 복원 대상이 아니다.

(나) 아. 또 너희들 싸우고 그래 **하면서** 가기도 하고.
→ “‘~그래’라고 하면서”로 해석한다.
→ 구문 분석) 그래 → VP_CMP 하면서
→ 서술어 ‘하면서’는 목적어 복원 대상이 아니다.

(3) 단, 발화 단위 내 인용문이 인용 표지 없이 완전한 문장으로 구성된 경우, 선행하는 인용문은 인용절로 처리하지 않고, 독립된 문장 단위로 처리한다. 후행 문장을 독립된 문장으로 보고 목적어가 생략된 경우, 목적어를 복원한다.[34)]

(가) P1-1: 가장 큰 **문제**는 공인이라는 자리는
P1-2: 대중에게 평가받는 자리잖아요?
P2-1: 그렇죠.
P1-3: 그러니까 여러 상황 가능성을 고려하되

32) 무형대용어 목적격 복원에서 문어 텍스트는 신문 말뭉치를 대상으로 하였다. 신문 기사의 특성상 인용격 조사의 생략이 생기지 않기 때문에 인용격 조사 생략은 구어의 세부 유형별 가이드라인에서 다룬다.
33) 단, 해당 인용 부분에 문장 부호가 없는 경우에 해당된다.
34) ‘국립국어원 구문 분석 지침’의 ‘인용 표지가 없는 인용의 주석’에서는 인용절과 후행 문장이 완전한 경우에만 각각을 독립된 문장으로 보고 있다. 2019년 주격 무형 대용어 복원 말뭉치 구축에서는 인용절의 독립성을 기준으로 후행 문장의 생략어 복원 여부를 결정하여 말뭉치를 구축하였다. 본 사업의 목적격 무형 대용어 복원 방향은 본 보고서 각주 5) “목적격 무형 대용어 복원 분석 시 구문 분석 결과를 기준으로 생략어 복원 대상 서술어 여부를 결정하지만, 이 경우 주격 무형 대용어 복원 지침과의 일관성이 유지되어야 한다는 전제가 우선한다.”에 제시한 바에 따라 2019년 주격 무형 대용어 복원 말뭉치 구축 원칙과 일관성을 유지하여 말뭉치를 복원하였다.

P1-4: 결단력 있는 리더십

P1-5: 이게 가장 중요한 덕목이라고 생각합니다.

P2-2: 김철수는 그런 자질을 보여주지 못하고 있다.

P2-3: 그래서 인기가 떨어졌다. (그렇게) **보시는** 거죠?

→ P2-3을 "그래서 인기가 떨어졌다고 (그렇게) 보시는 거죠?" 와 같이 인용문으로 분석하지 않고, 별도의 문장으로 처리하여 [떨어졌다. → VP_CMP 거죠?]로 분석한다.

→ 서술어 '보시는'의 목적어를 '문제'로 복원한다.

3.2.3. 문어와 다른 복원 대상 서술어의 처리

(1) 구어 원시 말뭉치에 마크업 기호 <trunc></trunc>가 부착되어 있는 경우 복원 대상 서술어에서 제외한다.

(2) 그러나 마크업 기호 <unclear></unclear>가 부착되어 있는 경우는 복원 대상 서술어로 분석하여, 분절된 경우에도 하나의 서술어로 간주하여 목적어를 복원한다.

(가) 나중에 <trunc>**죄송함**</trunc> 막 제가 다 <trunc>**잘못했어**</trunc> 제발 <trunc>도</trunc> 좀 알려주세요 막 이러는데도

→ (1)에 따라 위의 구어 예문에서 <trunc></trunc>로 처리된 '죄송함', '잘못했어'는 복원 대상 서술어에서 제외한다.

(나) 어~ <trunc>**만드**</trunc> **는** 회사의 대표로 있습니다.

→ 위와 유사한 예로, 위처럼 '만드는'의 일부인 '만드'만 <trunc></trunc>로 표기되어 분절된 경우에도 '는'을 서술어로 선택하지 않고, 복원 대상 서술어에서 제외한다.

(다) 피해자가 사건을 공개하길 바라면 제가 <unclear>**xx했어요**.</unclear>

→ <unclear></unclear>로 처리된 'xx했어요'는 복원 대상 서술어이나, 타동사로 판명이 어려운 경우에 해당되어 목적어 복원 대상이 아니다.

3.2.4. 미등재어의 처리

(1) 미등재어 처리의 기본 원칙은 문어 복원 지침과 같다. <표준국어대사전>에 없는 동사는 <우리말샘>의 문법 및 어휘 정보를 참고하며, 가능한 한 생략된 목적어를 최대한 복원하고자 한다.

(2) 구어 원시 말뭉치는 발화자의 발음을 전사하는 과정에서 소리 나는 대로 적은 표현, 오탈자 등으로 올바른 표기법과 다른 경우가 존재한다. 문맥상 올바른 표기의 추측이 가능한 경우, 문맥에 따라 문장의 의존 관계를 고려한다.

(가) 그거 진짜 거기 **늏면** 컴퓨터가 다라라라락해서
→ '늏면'은 '넣으면'으로 해석되며, [그거 →NP_OBJ 늏면]으로 분석된다.

(나) **이걸르** 생각해보면 여성할당제는 반드시 필요하다
→ 문맥상 '이걸르'는 [이걸르 → NP_OBJ 생각해보면]으로 분석이 가능하며, 이에 따라 서술어 '생각해보면'은 목적어 복원 대상이 아니다.

3.2.5. 목적어 없이 조사와 연결 어미만 남아 있는 문장의 복원

(1) 구어 말뭉치의 특성상 명사구 어절이 쪼개지거나 명사가 불분명하여 격 조사로 시작하는 문장도 있다.

(2) 목적격 조사나 복원 대상 서술어의 격틀을 충족하는 조사가 존재하면 목적어 복원 대상이 아닌 것으로 간주한다.

(가) P1-1: 사실 이게 합법화 되면은
P1-2: <unclear></unclear> 을 **막을** 이유가 없다고 생각을 하거든요.
→ '막다'의 격틀 구조에 따라 명사의 결합 없이 목적격 조사 '을'만 있더라도 구문 분석에서 [을 → X_OBJ 막을]로 분석된다.
→ 서술어 '막을'은 목적어 복원 대상이 아니다.

(나) P1-1: 그동안 쏟아졌던 악플을 보면
P1-2: 사람들이 언제 철들 거야?

P1-3: 음.

P1-4: 라고 계속 **물어본다고**.

→ ‘물어보다’의 격틀은 […에/에게 …을][…에/에게 -ㄴ지를][…에/에게 -고]이다.

→ P1-4의 서술어 ‘물어본다고’에 인용격 조사가 있으므로 목적어 복원 대상 술어가 아니다.

(3) 문두 이외의 위치에 나타나는 목적격 조사 단독 어절이나 연결 어미만 남은 경우도 위와 마찬가지로 처리한다.

(가) P2-1: 사회적 약자에 대한 편견 어쩌라고?

P2-2: 네 에 대해서 **이야기해** 봤는데

→ ‘이야기하다’의 격틀은 […에/에게 …을][…에/에게 -ㄴ지를][…에/에게 -고] ((‘…을’ 대신에 ‘…에 대하여’가 쓰이기도 한다))이다.

→ P2-2의 서술어 ‘이야기해’는 ‘…에 대하여’가 있으므로 목적어 복원 대상이 아니다.

3.2.6. ‘-게 하다’ 구문의 복원

(1) ‘하다’가 보조 용언으로 쓰이지 않고 본용언으로 쓰이면 격틀 구조에 따라 목적어를 복원한다.

(2) 본용언 ‘하다’ 중 ‘-게 하다’의 구문으로 쓰이는 예가 구어 문서에서 자주 나타나는데, 다음과 같이 <표준국어대사전> ‘하다’의 의미 번호 2, 8, 9[35]에 해당된다.[36]

35) 이하 <표준국어대사전> 동일하므로 생략한다.

36) 3.2.6.은 구어에서 나타나는 ‘방법, 방식’의 의미로 쓰인 ‘-게 하다’ 구문의 문제로, 2.2.5.의 문어에서 나타나는 사동의 ‘-게 하다’ 구문과 같은 구조로 인식하여 복원 오류가 생기는 일을 방지하기 위해 삽입한 것이다.

〈‘하다’ 의미 번호 2〉
【…을 -게】
((**주로 ‘어떻게’와 함께 쓰이며** ‘-게’ 대신에 ‘-듯’이나 ‘…대로’ 따위가 쓰이기도 한다)) 사건이나 문제 따위를 처리하다.

· 상금으로 받은 돈을 어떻게 하는 것이 좋을까요?
· 이 사기범을 어떻게 할까요?
· 그는 그 많은 돈을 떡 주무르듯 한다.

〈‘하다’ 의미 번호 8〉
【…에/에게 -게】
((‘-게’ 대신에 ‘-노라고, -ㄴ다고’ 따위나 ‘…대로, …처럼’ 따위의 부사어가 쓰이기도 한다)) 다른 사람에게 **특별한 방식으로 어떤 영향을 주거나 대하다.**

· 아이에게 어떻게 했기에 아이가 저렇게 기가 죽어 있냐?
· 다른 부서에 부당하게 하면 회사의 조직력에 해를 입힐 수 있다.
· 그녀는 자신의 남자친구에게 하노라고 했는데도 좋은 소리를 듣지 못하고 있다.

〈‘하다’ 의미 번호 9〉
【-게】
((‘-게’ 대신에 ‘…대로, …처럼’ 따위의 부사어가 쓰이기도 한다)) **어떤 방식으로 행위를 이루다**.

· 앞으로 어떻게 할 생각이냐?
· 네가 하는 대로 하면 나도 잘할 수 있을까?
· 앞으로는 네가 한 것처럼 할 작정이다.

(3) ‘하다’의 의미 번호 8, 9는 ‘어떻게’, ‘방법 및 방식’을 나타낸다는 점에서 의미 번호 2와 의미적으로 유사하다. 하지만 의미 번호 2는 목적격을 요구하고, 의미 번호 8, 9는 그렇지 않다는 점에서 차이를 보인다. 따라서 문맥상 반드시 목적격이 요구되면 2의 의미로 보아 목적어를 복원하고, 특별한 선행어가 없거나 목적격 복원을 하면 어색한 문장에서는 8, 9의 의미로 보아 목적어를 복원하지 않는다.

(가) 눈만 보이게 그렇게 **하니까**
→ ‘그렇게’는 ‘방법, 방식’의 의미로 ‘하다’ 의미 번호 9에 따르면 목적어가 반드시 필요하지 않다.

→ 서술어 ‘하니까’는 목적어 복원 대상이 아니다.

(나) 오늘도 오면 **어떡해**

→ ‘어떡하다’는 ‘어떠하게 하다’의 줄임말이므로 ‘하다’의 의미 번호 9의 의미로 분석한다.

→ 서술어 ‘어떡해’는 목적어 복원 대상이 아니다.

3.2.7. 호칭어의 처리

(1) 구어 말뭉치에서는 호칭어 다음에 쉼표, 말줄임표와 같은 문장 부호가 자주 생략된다.

(2) 호칭어는 일반적으로 모문의 서술어와 의존 관계에 있는 것으로 분석된다. 따라서 호칭어와 목적어가 동일하게 해석될 때, 구문 분석상 IP로 주석된 어절에 대해 목적어로 오인하지 않도록 유의해야 한다.

(가) **선생님** 저 **만나고** 싶었어요 예.

→ ‘선생님’은 호칭어로 [선생님 → IP 싶었어요]로 분석된다.

→ ‘만나고’는 지배하는 NP_OBJ가 없으므로 목적격 복원 대상 술어이다.

→ ‘만나고’의 목적어를 ‘선생님’으로 복원한다.

(나) 김 **선생님** **사랑해요**.

→ 문맥에서 ‘선생님’이 호칭어로 판단되면 구문 분석에서 [선생님 → IP 사랑해요]로 분석한다.[37)]

→ ‘사랑해요’의 목적어를 ‘선생님’으로 복원한다.

3.2.8. 감탄사의 처리

(1) 구어에서 출현 빈도가 높은 감탄사 ‘뭐’를 대명사 ‘무엇’과 혼동하지 않

37) 맥락에 따라 2인칭 또는 3인칭을 판단해야 한다. ‘선생님’이 청자와 동일한 인물이라면 2인칭으로 판단하여 위와 같이 복원하지만, ‘선생님’이 청자와 동일한 인물이 아닌 3인칭으로서 지칭하는 대상이라면, 아래와 같이 복원하지 않는다.
- P2-1: 누구를 사랑한다고요?
P1-1: 김 선생님 **사랑해요**.
→ P2가 P1이 사랑하는 선생님이 아닐 경우에는 NP_OBJ로 분석되므로 ‘사랑해요’는 목적격 복원 대상 술어가 아니다.

도록 유의한다.

(2) 본 사업의 구문 분석 지침에 따르면 감탄사는 바로 다음 어절에 의존한다. 따라서 의미적으로 목적어로 해석되더라도 IP로 주석된 어절을 목적어로 분석하지 않도록 주의한다.

(가) 뭐 아까 박앵커님이 정확히 **말씀하셨는데**

→ 이 문장에서 '뭐'는 감탄사로 분석된다. 따라서 지배하는 목적어가 없는 서술어 '말씀하셨는데'의 목적어는 '사건', '사항' 등의 명사로 복원한다. 해당 명사가 문서 내에 없다면 '무엇을'로 복원한다.

3.2.9. 발화자에 따른 인칭 및 지칭 표현의 선택

(1) 인칭 및 지칭 표현의 선택은 문어 말뭉치와 동일하게 처리한다.

(2) 단, 선행어 후보가 복수일 때 동일 발화자의 표현이 우선한다.

(가) P1-1: **지수**랑 선후배 사이라고 그러던데
P2-1: 네 한 학번 선배님이세요
P1-2: 대학 때 **본** 적 있어요?

→ P2-1의 '선배님'과 P1-1의 '지수'가 동일한 사람을 지칭하지만, P1-2의 서술어 '본'의 목적어를 P1의 발화 속 지칭 표현인 '지수'로 복원한다.

→ P1-2의 서술어 '본'의 목적어를 '지수'로 복원한다.

(나) P3-1: **선생님**께서 나오신다 해서 기대하고 있었습니다.
P3-2: 시청자분들께 인사 부탁드립니다.
P4-1: 예. 안녕하세요. 글 쓰는 직장인 OOO사 김철수 대리입니다.
P3-3: 제가 **소개하길** 직장인에게 가장 인기 있는 작가라고 말씀드렸는데.

→ P3-1의 '선생님'과 P4-1의 '대리'가 동일한 사람을 지칭하지만, P4의 '대리'는 발화자 김철수 자신을 지시하므로, P3-3의 서술어 '소개하길'의 목적어를 P3이 지칭하는 '선생님'으로 복원한다.

→ P3-3의 서술어 '소개하길'의 목적어를 '선생님'으로 복원한다.

3.2.10. 방송 텍스트의 ‘함께 보시죠/함께 만나 보시죠’ 구문의 복원

(1) 구어 문서에는 방송 프로그램의 진행자와 게스트, 진행자와 청중 간의 대화가 다수 포함되어 있다.

(2) ‘함께 보시죠/함께 만나 보시죠’의 출현 빈도가 높다.

(3) ‘보시죠’의 목적어는 프로그램 내용, 화면 등의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘무엇을’로 복원한다.

(4) ‘만나’의 목적어는 프로그램의 인물, 사람, 이름 등의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘누구를’로 복원한다.

(가) 안내견 훈련사를 만나봅니다. 함께 **보시죠**.

→ ‘보시죠’의 목적어로 생각되는 ‘영상’과 같은 명사가 문서 내에 없다면 ‘무엇을’로 복원한다.

(나) P1-1: 안내견 **훈련사**를 소개해 드립니다.

P1-2: 함께 **만나** 보시죠.

→ P1-2의 서술어 ‘만나’의 목적어를 ‘훈련사’로 복원한다.

3.2.11. 띄어쓰기 오류 및 비문의 처리

(1) 띄어쓰기 오류를 포함한 서술어와 선행어의 선택에서는 어절 단위 태깅 원칙을 우선 적용한다. 따라서, 아래와 같이 두 어절로 분리되어야 하는 띄어쓰기 오류는 어절 단위 태깅 원칙을 적용하여 그대로 복원 술어의 단위가 된다.[38)]

(가) P1-1: 딸이 나한테 **해줘갖고**

P1-2: 나 너무 좋았잖아

→ 서술어 ‘해줘갖고’의 올바른 띄어쓰기는 ‘해 줘 갖고’ 또는 ‘해줘 갖고’이나 어절 단위 태깅을 원칙으로 적용하여 ‘해줘갖고’를 하나의 복원 술어 단위로 삼는다.

→ P1-1의 서술어 ‘해줘갖고’의 목적어는 ‘물건’, ‘효도’ 등의 명사로 복원하고 해당 명사가 문서 내에 없다면 ‘무엇을’로 복원한다.

38) 구어 말뭉치와 달리 문어 말뭉치에서는 띄어쓰기 오류가 극히 드물어 어절 단위 태깅 원칙만 적용하도록 하였다.

(2) 단, 다음과 같은 경우 띄어쓰기 원칙을 어절 단위 태깅 원칙보다 우선 적용한다.

(2-1) 띄어 쓰지 않아야 하나 잘못 표기되어 한 단어가 두 개 이상 어절로 분리된 경우가 있다. 이때 각각의 어절 단위로 해석 시 전체 문장 의미 해석이 달라지면 한 어절이 아니라도 복원 대상 서술어 또는 선행어로 선택할 수 있다.

(가) 이 일로 **먹고 살면서**

→ '먹고 살면서'는 띄어쓰기 오류로, 올바른 형태는 한 어절인 '먹고살다'이다. 이 경우, '이 일로 먹다'와 '이 일로 살다' 등 두 술어 각각의 의미로 해석되지 않으므로 '먹고 살면서'를 하나의 서술어로 분석한다.

→ '먹고살다'는 목적격을 요구하지 않으므로 서술어 '먹고 살면서'는 복원 대상 서술어가 아니다.

(2-2) 본용언+보조 용언 결합에서 일부 띄어쓰기가 잘못된 경우, (2-1)처럼 각각의 어절 단위로 해석 시 전체 문장 의미 해석이 달라지면 한 어절이 아니라도 복원 대상 술어 또는 선행어로 선택할 수 있다.

(가) P1-1: 제 **영상**을 보고 **재밌어 해주고**

P1-2: 공감해주는 게 즐겁고

→ '재밌어 해주고'의 올바른 띄어쓰기는 '재밌어해 주고' 이다.

→ '(시청자들이) 재밌다'와 '(시청자들이) 해주다' 등 두 술어 각각의 의미로 해석되지 않으므로 (2)와 일관되게 전체 '재밌어 해주고'를 술어로 선택한다.

→ P3-1의 서술어 '재밌어 해주고'의 목적어를 '영상'으로 복원한다.

(2-3) 띄어쓰기가 되지 않은 구문의 서술어 또는 명사구의 분리가 다음과 같이 필요한 경우, 하나의 어절에서 일부를 분리하여 복원 대상 술어 또는 선행어로 선택할 수 있다.

(가) 주로 동생이 **아끼는**걸 숨기는 장난을 했죠.

→ 품사가 다른 '서술어+(의존)명사'가 붙여쓰기 되어 있는 경우, 띄어쓰기 오류를 인정하여 후행 명사를 제외한 서술어만 선택한다.

→ 서술어 '아끼는'의 목적어를 '걸'에서 조사를 삭제한 '거'로 복원한다.

(3) 올바른 문맥을 찾기 어려운 비문은 그대로 해석하여 서술어의 목적어를 복원한다.

(가) 어디서 나는 오래된 **음식**이라고 **먹을** 수 있다 그랬나?

→ '음식이라고 먹을 수 있다'는 '음식이라고,(휴지) 먹을 수 있다'와 '음식이라도 먹을 수 있다'와 같이 두 가지 경우로 해석된다. 이와 같이 '올바른 표기형'을 가정하는 것이 어렵다면 문장에서 나타난 표면형 그대로 '음식이라고'로 문장을 해석한다.

→ 서술어 '먹을'의 목적어를 '음식'으로 복원한다.

3.2.12. 문장 내 휴지

(1) 구어 말뭉치에서는 휴지를 표시하는 쉼표, 말줄임표 등이 자주 생략된다.

(2) 발화에 휴지가 존재한다고 생각되면 휴지로 나뉘는 범위 내에서 선행어를 우선 선택한다. 이는 구문 분석 결과에 일치하게 처리한다.

(가) 이런 사건 지금 **말씀하신** **사건**에서는

→ 예시 문장은 '말씀하신'이 '이런 사건'을 목적어로 취하는 것이 아니라 '이런 사건'과 '지금 말씀하신 사건에서는'이 동격 관계를 이룬다.

→ 즉, '이런 사건, //휴지// 지금 말씀하신 사건에서는'의 구조이다.

→ 관형절 복원 원칙에 따라 서술어 '말씀하신'의 목적어를 관형절 피수식어인 두 번째 '사건'으로 복원한다.

4. 말뭉치 구축 및 납품

4.1. 말뭉치 구축

4.1.1. 말뭉치 구축 절차

(1) 구어 구문 분석 말뭉치

작업자들은 원시 문장에 대한 자동 구문 분석 결과를 확인하고 오류를 발견하여 수정하는 작업을 수행하였다. 작업자는 문장의 의존 관계를 트리 형식으로 확인할 수 있으며, 각 어절이 올바른 지배소 및 의존소에 연결되었는지, 구문 태그 및 기능 태그가 올바르게 부여되었는지 검수하였다.

(2) 목적격 무형 대용어 복원 말뭉치

말뭉치 구축을 위해 관리 기관은 원본 데이터 형태 분석 및 본문 주석 등을 자동으로 처리하고, 목적격 무형 대용어 복원을 위한 서술어를 추출하였다.

목적격 복원 대상 서술어 후보군이 자동으로 제공되면 작업자들은 이들을 바탕으로 서술어 추가 및 삭제 작업을 진행하였다. 작업자는 목적격 복원 대상 서술어를 선택한 뒤 적절한 선행어를 연결하였으며, 문서 내에 적절한 선행어가 없는 경우, 비지시적 대명사인 ‘누구를’ 또는 ‘무엇을’로 선행어를 복원하였다.

4.1.2. 말뭉치 구축 결과

주관 기관에서 제공한 원시 말뭉치 약 300만 어절을 대상으로 구어 구문 분석 말뭉치 및 목적격 무형 대용어 복원 말뭉치를 구축하였다.

(1) 구어 구문 분석 말뭉치

구문 분석은 ‘2019년 국립국어원 구어 형태 분석 말뭉치’를 대상으로 하였으며, 7월 27일부터 11월 20일까지 말뭉치 구축 작업을 수행하였다. 구축이 완료된 구어 구문 분석 말뭉치의 총 어절 수는 1,006,448어절이며, 문장 수는 221,489개, 문서 수는 423개이다.

구어 구문 분석 말뭉치는 총 4차에 걸쳐 구축되었는데 차수별로 약 25만 어절씩 고르게 구축되었다. 구어 구문 분석 말뭉치에 부착된 주석을 확인한 결과, 구문 태그 개수는 총 1,006,448개로 모든 어절에 구문 태그가 주석된 것을 확인하였다. 기능 태그 개수는 404,047개, 지배소를 갖는 어절 수는 1,006,448개, 의존소를 갖는 어절 수는 531,136개이다. 지배소 값에는 ROOT값(최상위 지배소)이 포함되었으므로 모든 어절에 빠짐없이 지배소 주석이 부착되었다.

<표 7> 구어 구문 분석 말뭉치 주석 결과

총 어절 수	총 문장 수	구문 태그 수	기능 태그 수	지배소를 갖는 어절 수	의존소를 갖는 어절 수
1,006,448	221,489	1,006,448	404,047	1,006,448	531,136

(※지배소 값에는 ROOT값 포함)

(2) 목적격 무형 대용어 복원 말뭉치

목적격 무형 대용어 복원 분석은 '2019년 국립국어원 주격 무형 대용어 복원 말뭉치'를 대상으로 하였으며, 7월 28일부터 11월 26일까지 말뭉치 구축 작업을 수행하였다. 구축이 완료된 목적격 무형 대용어 복원 말뭉치의 총 어절 수는 3,006,661어절로 문어가 2,000,213어절, 구어가 1,006,448어절이다. 문어의 구축 문서 수는 7,265개, 구어의 구축 문서 수는 423개로 총 7,688개의 문서가 구축되었다.

목적격 무형 대용어 복원 말뭉치는 총 5차에 걸쳐 구축되었는데 1-3차 기간에는 문어 말뭉치를, 4-5차 기간에는 구어 말뭉치를 구축하였다. 목적격 무형 대용어 복원 말뭉치에 부착된 주석을 확인한 결과, 문어의 복원 대상 서술어 개수는 40,446개, 구어의 복원 대상 서술어 개수는 43,149개로 나타났다. 복원 대상 서술어와 선행어는 하나의 쌍으로 주석되므로, 총 주석 수는 문어가 80,892개, 구어가 86,298개로 총 167,190개의 주석이 부착되었다.

<표 8> 목적격 무형 대용어 복원 분석 말뭉치 주석 결과

	총 어절 수	총 문장 수	복원 대상 서술어 수	선행어 수
문어	2,000,213	150,082	40,446	40,446
구어	1,006,448	221,489	43,149	43,149
총합	**3,006,661**	**371,571**	**83,595**	**83,595**

문어 말뭉치가 구어 말뭉치에 비해 총 어절 수가 두 배가량 많음에도 불구하고, 구어 말뭉치의 복원 대상 서술어 수가 더 많다는 사실을 통해 구어에서의 목적격 생략 비율이 문어에서의 목적격 생략 비율보다 월등히 높음을 알 수 있다.

4.2. 말뭉치 납품

본 사업은 말뭉치의 품질 확인을 위해 3차에 걸쳐 말뭉치 납품을 진행하였다. 납품 후에는 주관 기간에서 주석 내용을 확인하여 피드백을 송부하였으며, 구축 기관은 해당 내용을 작업자에게 교육하고, 이후 작업 및 지침에 해당 사항을 반영하였다.

(1) 구어 구문 분석 말뭉치

<표 9> 구어 구문 분석 말뭉치 납품 결과

	자동 분석 결과 납품	1차 납품	2차 납품	3차 납품
납품 일시	2020-07-17	2020-09-11	2020-10-30	2020-12-21
납품 대상	자동 분석 결과	1차 구축분	1~3차 구축분	1~4차 구축분
납품 문서 수	423	137	377	423
납품 어절 수	1,006,448	245,658	741,706	1,006,448
납품 문장 수	**221,489**	**41,433**	**157,761**	**223,962**
구축 문장 수	-	40,876	155,972	221,489
빈 문장 수 (언어 외적 표지)	-	557	1,789	2,473

구어 구문 말뭉치 분석을 위해 작업 초기 주관 기관에서 제공한 구어 형태 분석 말뭉치를 자동 분석하여 그 결과를 납품하였다. 결과는 JSON 파일로 납품하였으며, 총 납품 어절 수는 1,006,448어절(221,489문장)이다. 이후 1차 납품에서는 245,658어절(41,433문장), 2차 납품에서는 741,706어절(157,761문장)을 납품하였다. 마지막 3차 납품에서는 1-4차 구축분을 모두 납품하였으며, 총 1,006,448어절(223,962문장)을 납품하였다. 납품 수량에는 언어 외적 표지(구어 마크업 기호)만 있는 문장도 포함되었다.[39)]

39) 자동 분석 결과 납품본과 3차 납품본과의 주석 차이 수는 총 357,196개로 집계되었다. 이는 총 어절 수

(2) 목적격 무형 대용어 복원 말뭉치

<표 10> 목적격 무형 대용어 복원 분석 말뭉치 납품 결과

	1차 납품	2차 납품	3차 납품	
납품 일시	2020-09-11	2020-10-30	2020-12-21	
납품 대상	문어 일부	문어 전체	문어, 구어 전체	
납품 문서 수	3,195	7,265	문어 7,265	구어 423
			7,668	
납품 문장 수	73,864	150,082	문어 150,082	구어 223,962
			374,044	
납품 어절 수	980,717	2,000,213	문어 2,000,213	구어 1,006,448
			3,006,661	

목적격 무형 대용어 분석 말뭉치는 1차 납품 시에는 문어 일부를, 2차 납품 시에는 문어 전체를 납품하였다, 마지막 3차 납품 시에는 문어와 구어를 전체 납품하였다. 1차 납품 어절 수는 980,717어절(3,195개 문서), 2차 납품 어절 수는 2,000,213어절(7,265개 문서)이다. 3차 납품에서는 문어 2,000,213어절(7,265개 문서), 구어 1,006,448어절(423개 문서), 총 3,006,661어절(7,668개 문서)을 납품하였다. 문어의 경우, 바이라인(byline) 등을 후처리에서 삭제하고 납품하였으므로 실제 구축 수와는 차이를 보인다.

1,006,448어절 가운데 약 35%에 해당하는 수치이다. 지배소의 수정이 92,325건, 의존소의 수정이 152,753건, 구문 태그 및 기능 태그의 수정이 112,118건으로 나타났다.

5. 검증 및 산출물 보고

본 사업은 납품 자료 전체에 대한 일관성 및 정확성을 검증하여 국내 표준화 작업에 기여하고 참조 기반 자료가 될 수 있는 고품질 말뭉치를 구축하였다.

5.1. 검증 절차

5.1.1. 작업자 검증

말뭉치 주석 작업자가 자신의 작업을 검증하는 최초의 검증 단계이다. 작업 목록에서 오류 항목을 발견, 선택하면 해당 문장의 위치로 자동 이동되어 작업자 스스로 직관적인 검증을 하도록 지원하였다. 고민이 필요한 항목은 작업 보류 목록으로 지정하여 추후에 해당 항목만 모아서 확인하였다. 작업자 스스로 판단이 어려운 항목은 검수자에게 검토를 요청해 검증 절차를 강화하였다.

5.1.2. 기계적 검증

작업자가 문서 작업을 완료하면 문서 내 미작업 대상에 대해 자동으로 검수를 시행하였다. 분석 후보에서 작업이 누락된 어절이 있을 경우 이를 작업자에게 알려 문서를 재확인하게끔 하였다.

5.1.3. 절차적 검증

말뭉치 구축 인력을 작업자와 검수자로 구분하여 작업자의 작업 결과를 검수자가 검토할 수 있도록 단계적 역할을 부여하였다. 검수자는 주석 결과의 오류 여부를 확인하는 동시에 지침 준수 여부를 검토하여 작업자에게 피드백을 제공하고 교육을 실시하였다. 작업자 간 지침 해석이 불일치하는 경우, 원인을 찾아 지침을 수정 및 세분화하여 말뭉치의 일관성을 확보하고, 품질을 개선하였다.

5.1.4. 관리적 검증

관리자는 도구에서 작업의 전반적인 진행 현황을 파악할 수 있다. 작업자의 전체 작업 현황 및 동일 기간 내 작업자별 수행량을 측정하여 말뭉치 구축 일정이 차질이 없도록 관리하였다. 내용적인 측면에서는 작업 보류 및 검토 요청 이력을 격리 저장하여, 작업의 품질과 작업자의 수행 성과를 관리하였다.

5.1.5. 통계적 검증

앞선 검증을 통해 구축 작업이 완료된 문서를 대상으로 통계적 검증을 실시하였다. 다수의 작업자가 대량의 작업을 수행하므로 전체 자료의 품질을 보증하려면 통계적 접근이 필수적이다.

통계적 검증은 구축 자료 가운데 샘플링 자료를 검증하는 방식으로 진행되었으며, 2019년 검증 방법에 활용된 정답 세트 기반 검증 방법을 적용하였다. 정답 세트가 되는 샘플링 자료는 각 구축 기관에서 선별한 문서 및 문장을 대상으로 전체 구축 자료의 10%로 구성하였고, 관리 기관에서 작업자에게 매주 정답 세트 문서를 비공개로 배정하였다. 정답의 기준이 되는 문서는 각 구축 기관의 박사 과정 이상 전문가 집단이 주석 작업하였으며, 정기 회의를 통해 정답 세트 간 일관성을 확보하였다.

통계적 검증을 위해 반복적인 검증 주기를 부여하여, 주 단위 작업 결과에 대한 정답 세트 비교 검증을 수행하였다. 구문 분석 말뭉치의 경우, 1주일 단위로 총 9회 사전 구축한 정답 세트(총 100,958어절)로 검증하였다. 목적격 무형 대용어 복원 말뭉치의 경우, 문어 6회(총 201,158어절), 구어 5회(총 101,425)로 총 11회 사전 구축한 정답 세트로 검증하였다.[40]

<표 11> 구어 구문 분석 말뭉치 정답 세트 작업량

구축 차수	1차	2차	3차	4차
구축 기간	7.27-9.4	9.9-9.25	10.5-10.23	10.26-11.20
작업 어절 수	24,998	24,997	25,965	24,998
누적 어절 수	24,998	49,995	75,960	**100,958**
총 어절 수	**100,958**			

40) 차수별 검증 점수 추이를 '5.2. 검증 결과'에 제시하였다.

<표 12> 목적격 무형 대용어 복원 분석 말뭉치 정답 세트 작업량

구축 차수	1차	2차	3차	4차	5차
구축 기간	7.28-8.13	8.14-9.10	9.11-9.24	10.16-11.5	11.6-11.26
사용역	문어			구어	
작업 어절 수	80,931	69,769	50,458	51,886	49,539
누적 어절 수	80,931	150,700	**201,158**	51,886	**101,425**
총 어절 수	**302,583**				

주 단위 품질 측정을 통해 품질 수준에 대한 신뢰성을 높였으며, 품질 수준 변동률을 산출해 진척 향상률을 확인하였다.

관리 기관 ㈜이르테크는 정답 세트 간 불일치 여부를 전수 대조·검증하여 검증 보고서를 작성하였다. 검증 보고서에는 작업자별 오류 상세 내용 및 검증 점수가 모두 포함되었으며, 구축 기관에 해당 보고서를 전달하여 동일 오류 재발을 방지하고, 기준 점수 이하의 작업자는 재교육을 실시하였다. 정답 세트 검증 보고서의 누적을 통해 통계적인 오류 분포와 양상의 검증이 가능하였으며, 누적 데이터를 활용해 품질 개선의 집중도를 높였다.

```
작업자:::::qaqzq

5. 검증문장:::::G_SBRW1800000139.9 이제는 안티안티에이징이 대세가 됐다고 해요.

{'word_id': 1, 'word_form': '이제는', 'head': 4, 'label': 'AP -> NP_AJT', 'dependent': []}
{'word_id': 2, 'word_form': '안티안티에이징이', 'head': 4, 'label': 'NP_SBJ', 'dependent': []}
{'word_id': 3, 'word_form': '대세가', 'head': 4, 'label': 'NP_CMP', 'dependent': []}
{'word_id': 4, 'word_form': '됐다고', 'head': 5, 'label': 'VP_CMP', 'dependent': [1, 2, 3]}
{'word_id': 5, 'word_form': '해요.', 'head': 0, 'label': 'VP', 'dependent': [4]}


6. 검증문장:::::G_SBRW1800000139.5 연령대도 이십대부터 사십대까지 굉장히 다양했다라고 해서

{'word_id': 1, 'word_form': '연령대도', 'head': 5, 'label': 'NP_SBJ', 'dependent': []}
{'word_id': 2, 'word_form': '이십대부터', 'head': '3 -> 5', 'label': 'NP_AJT', 'dependent': []}
{'word_id': 3, 'word_form': '사십대까지', 'head': 5, 'label': 'NP_AJT', 'dependent': '[2] -> []'}
{'word_id': 4, 'word_form': '굉장히', 'head': 5, 'label': 'AP', 'dependent': []}
{'word_id': 5, 'word_form': '다양했다라고', 'head': 6, 'label': 'VP_CMP', 'dependent': '[1,3,4] -> [1,2,3,4]'}
{'word_id': 6, 'word_form': '해서', 'head': 0, 'label': 'VP', 'dependent': [5]}


7. 검증문장:::::G_SBRW1800000139.4 네~

{'word_id': 1, 'word_form': '네~', 'head': 0, 'label': 'IP', 'dependent': []}


8. 검증문장:::::G_SBRW1800000139.8 예전에는 안티에이징이라고 해서 노화에는 무조건 대항해야 된다라고 했다면

{'word_id': 1, 'word_form': '예전에는', 'head': 8, 'label': 'NP_AJT', 'dependent': []}
{'word_id': 2, 'word_form': '안티에이징이라고', 'head': 3, 'label': 'VNP_CMP', 'dependent': []}
{'word_id': 3, 'word_form': '해서', 'head': '8 -> 6', 'label': 'VP', 'dependent': [2]}
{'word_id': 4, 'word_form': '노화에는', 'head': 6, 'label': 'NP_AJT', 'dependent': []}
{'word_id': 5, 'word_form': '무조건', 'head': 6, 'label': 'AP', 'dependent': []}
{'word_id': 6, 'word_form': '대항해야', 'head': 7, 'label': 'VP', 'dependent': '[4,5] -> [3,4,5]'}
{'word_id': 7, 'word_form': '된다라고', 'head': 8, 'label': 'VP_CMP', 'dependent': [6]}
{'word_id': 8, 'word_form': '했다면', 'head': 0, 'label': 'VP', 'dependent': '[1,3,7] -> [1,7]'}


<<<<<<<<<<정답 스코어>>>>>>>>>>
recall : 90.9090909090909
precision : 92.10526315789474
F1 Score : 91.50326797385621
```

<그림 17> 구어 구문 분석 말뭉치의 검증 보고서 예시

<그림 18> 목적격 무형 대용어 복원 분석 말뭉치의 검증 보고서 예시

5.1.6. 정제

작업 결과물의 최종 납품 전, 일정 정제 기간을 마련하여 집약적으로 오류를 수정할 수 있도록 하였다. 정답 세트 검증을 통해 구축된 데이터를 바탕으로 오류가 빈번히 일어난 항목을 목록화하여 이들을 집중적으로 수정하였다.

구문 분석 말뭉치 구축에 대한 정제는 오류 유형에 해당하는 문장을 전수 추출하여 검증하는 방식으로 진행되었다. 분석 오류는 지침 수정에 따른 오류와 빈발 오류로 구분되며, 구축 기관이 각각의 항목을 유형화하여 전달하면 관리 기관이 이에 해당하는 문장을 추출하여 검수할 수 있도록 지원하였다. 분석 오류 가운데 빈발 오류에 대한 예시를 보이면 다음과 같다.

[구문 분석 말뭉치 구축에서 나타난 오류 예시]

1. [NP 중 NP] 연쇄

: ‘NP 중에 NP’ 구성의 ‘중에’는 서술어에 의존하지만, ‘NP 중 NP’ 구성의 ‘중’은 두 번째 NP에 의존하는 것이 원칙. 이때, ‘중’이 NP가 아닌 서술어에 의존하는 오류

<오류 예시> 제가 제일 어려웠던 얘기 중(→ NP 얘깁니다.) 하나가 그런 얘깁니다.

(정답) 중 → NP 하나가

2. 명사-부사 통용어

: 명사-부사 통용어의 품사가 잘못 주석되어 있는 오류

<오류 예시> 오늘(→ NP_AJT) 사용한 건 들기름이에요.

(정답) 오늘 → AP

3. ‘적’ 파생어

: 관형사 역할을 하는 ‘적’ 파생어를 NP로 주석한 오류

<오류 예시> 자존감에 있어서 부정적(→NP) 영향이 나타났기 때문에

(정답) 부정적 → DP

4. 수관형사 분석

: 수관형사를 NP로 태깅한 오류

<오류 예시> 육(→NP) 년 연속 최고의 공항이 된

(정답) 육 → DP

5. ‘라도/이라도’ 분석

: ‘라도/이라도’를 VNP로 주석한 오류

<오류 예시> 단 몇 명이라도(→VNP_SBJ)

(정답) 명이라도(→NP_SBJ)

무형 대용어 복원 말뭉치 구축에 대한 정제는 문서를 전수 검증하는 방식으로 진행되었다. 작업자들이 문서를 하나씩 확인하여 오류가 있는 어절을 발견하면 이를 즉시 수정하였다. 검수자는 정제 전 작업자 전체 교육을 실시하여 작업자들에게 지침을 다시 한번 주지시켰다. 정제의 대상이 된 오류 예시를 보이면 다음과 같다.

[무형 대용어 복원 말뭉치 구축에서 나타난 오류 예시]

1. 의존 명사 구성

: 관형절로 묶이는 의존 명사를 우선적으로 복원하지 않은 오류 (지시 관형사 삽입 가능 시)

<오류 예시> 서로 **(무엇을) 원하던** 것을 얻었다는 관측이 나온다.

(정답) '원하던'의 선행어 : 것

2. '하다'의 의미 번호 8번, 9번으로 실현된 경우

: <표준국어대사전> '하다'의 8번, 9번 의미 번호로 쓰여, 방법 및 방식을 나타내고 있으나 목적격을 복원한 오류

<오류 예시> 문제는 어떻게 **(무엇을) 하면** 원두커피 같은 맛을 내는 인스턴트 커피를 만드느냐였습니다.

(정답) 목적어 복원 대상 술어가 아님.

3. 발화자에 따른 지칭 표현

: 선행어 후보가 복수일 때 동일 발화자의 표현으로 선행어를 복원하지 않은 오류

<오류 예시> P1-1 안녕하세요 대표님?

P2-1 아. 네. 안녕하세요? 김영희입니다.

P1-2 예전에 한 번 **(김영희를) 만났었는데** 기억나시나요?

(정답) '만났었는데'의 선행어: 대표님

5.2. 검증 결과

작업자가 할당된 정답 세트 문서를 작업 완료하면 관리 기관은 해당 결과물을 미리 작업해 둔 정답 문서와 비교하여 검증 점수를 산출하고 검증 보고서를 작성하였다.

검증 점수는 기본적으로 정밀도(precision), 재현율(recall), F1 점수(F1 score)를 모두 산출하고, 최종적으로 F1 점수를 사용하였다. F1 점수는 정밀도와 재현율의 조화 평균으로 검증의 정확도에 대한 척도로 사용된다. 정답 세트의 점수 산출 방식은 아래와 같다.

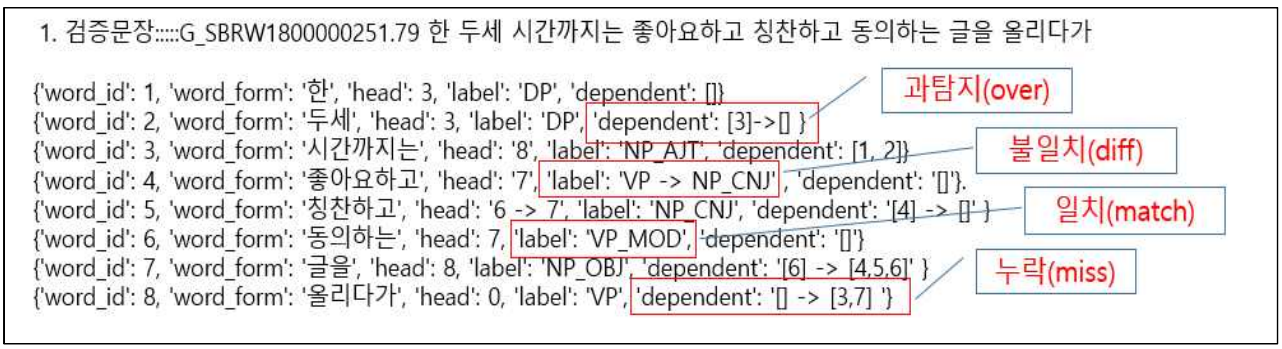
정밀도(precision)
작업자가 맞힌 개수/작업자가 작업한 주석 개수*100

재현율(recall)
작업자가 맞힌 개수/정답 주석 개수*100

F1점수(F1 score)
((precision * recall) / (precision + recall)) * 2

<수식 1> 정답 세트 점수 산출 방식

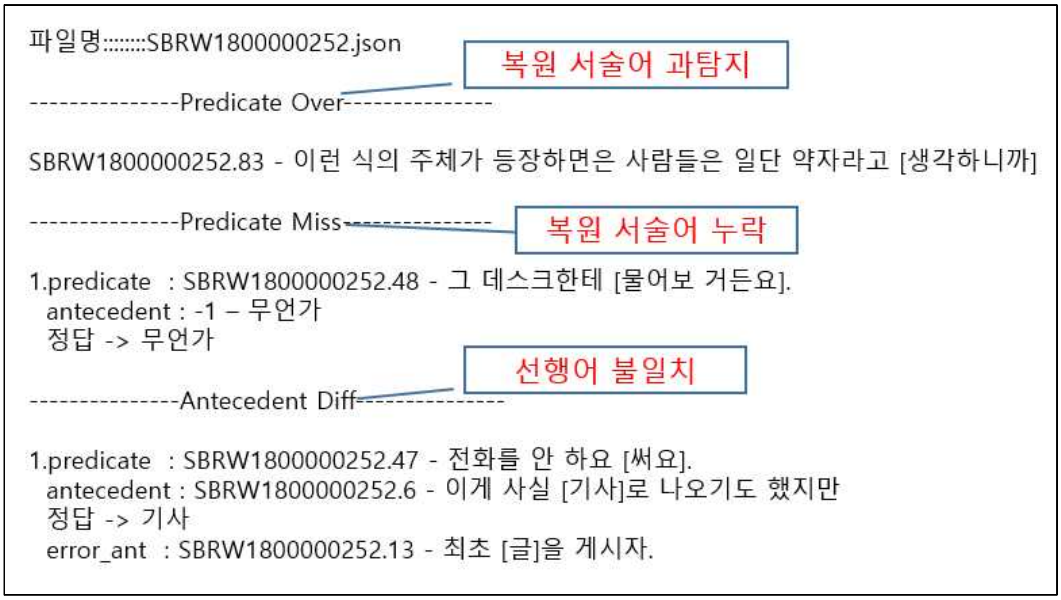
구문 분석 말뭉치의 경우, 작업자의 주석 결과를 누락(miss), 과탐지(over), 불일치(diff), 일치(match)의 네 가지 값으로 분류하여 점수를 산출하였다, 작업자별 검증 점수는 검증용 문장 전체 주석을 대상으로 합산 산출하였다.



<그림 19> 구어 구문 분석 말뭉치의 정답 세트 주석 대조 결과 예시

목적격 무형 대용어 복원 말뭉치의 경우, 작업자의 주석 결과를 복원 서술어 누락

(predicate miss), 복원 서술어 과탐지(predicate over), 선행어 불일치(antecedent diff)의 세 가지 값으로 분류하여 점수를 산출하였다. 작업자별 검증 점수는 검증용 문서 전체 주석을 대상으로 합산 산출하였다.



<그림 20> 목적격 무형 대용어 복원 분석 말뭉치의 정답 세트 주석 대조 결과 예시

정답 세트 검증이 완료되면 주석 대조 결과 및 검증 점수가 기재된 검증 보고서를 구축 기관에 전달하였다. 작업자는 말뭉치 구축 도구를 통해서도 자신의 검증 평균 점수를 확인할 수 있다.

<그림 21> 주석 도구를 통한 정답 세트 검증 점수 확인

구문 분석 말뭉치의 정답 세트 검증은 총 9회에 걸쳐 이루어졌으며 검증 결과는 아

래 표와 같다.

<표 13> 구어 구문 분석 말뭉치 정답 세트 검증 결과

검증일 / 작업자	8월24일	8월31일	**9월18일**	9월28일	**10월5일**	10월12일	10월19일	**10월26일**	**11월16일**
A	86.52	83.68	83.60	89.06	92.06	-	83.74	86.68	93.50
B	88.54	88.70	88.70	82.67	84.62	82.76	81.58	82.12	91.59
C	89.47	88.44	88.44	85.31	85.04	100.00	79.63	87.31	90.11
D	89.60	90.21	89.80	84.67	88.87	-	-	82.39	93.16
E	90.18	90.21	89.90	86.45	90.29	-	100.00	87.96	88.83
F	81.87	82.51	82.51	86.68	87.08	84.21	84.21	84.77	86.15
G	82.56	82.87	82.87	86.38	86.38	100.00	100.00	82.65	91.34
H	94.06	93.36	93.36	98.07	98.07	100.00	91.50	79.04	86.32
I	91.24	90.96	90.96	92.92	87.92	-	-	89.40	87.84
J	92.38	89.88	89.88	89.87	88.87	95.35	79.04	85.20	82.98
K	85.71	87.02	87.02	89.69	83.69	-	-	-	-
평균	88.38	87.99	**87.91**	88.34	**88.43**	93.72	87.46	**84.75**	**89.18**
총합 평균	88.46								

구문 분석 말뭉치 100만 어절은 총 4차에 걸쳐 약 25만 어절씩 구축되었다. 각 차수별로 구축이 완수된 9월 18일, 10월 5일, 10월 26일에는 1차, 2차, 3차 전체 구축물에 대한 검증을 각각 진행하였다. 4차 프로젝트 구축물에 대한 전체 검증은 11월 16일에 진행되었는데, 4차 프로젝트에서는 정답 세트 할당이 한 번만 이루어졌기 때문에 11월 16일 건 외의 검증 점수는 산출되지 않았다.

검증 점수는 85-95%에 분포하여 있고, 총합 평균은 88.46%로 안정적인 점수를 유지하였다. 10월 중간 보고 이후 지침이 개정되면서 점수가 일부 하락하는 양상이 보이기도 하였으나 정답 세트의 기준이 되는 정답 문서가 지침 개정 전 미리 구축된 문서라는 점을 감안할 때 후반부의 정답 세트 점수는 중의성을 지닌다. 정답 문서가 아닌 작업자의 작업 문서가 지침에 부합하다는 해석 또한 가능하기 때문이다. 정답 세트 검증은 구축 기간 동안 작업자의 작업 내용 및 향상률을 측정하기 위해 실시한 검증으로, 정답 세트 점수가 곧 말뭉치의 최종 품질 점수는 아니다. 말뭉치의 최종 품질 점수는 정제가 모두 이루어진 이후에 다른 방식을 적용하여 산출하였고, 이에 대해서는 ‘5.3. 품질 수준’에서 자세히 기술하였다.

목적격 무형 대용어 복원 말뭉치의 정답 세트 검증은 문어 6회, 구어 5회 총 11회에 걸쳐 이루어졌다. 우선 문어 말뭉치의 검증 결과는 아래 표와 같다.

<표 14> 문어 목적격 무형 대용어 복원 분석 말뭉치 정답 세트 검증 결과

검증일 / 작업자	8월10일	8월14일	8월21일	8월28일	9월4일	9월18일
A	75.00	79.19	84.42	80.70	80.39	90.18
B	100.00	87.01	72.60	83.84	80.00	83.33
C	76.92	72.84	80.39	76.36	85.71	82.72
D	85.71	75.49	64.10	80.82	73.33	82.86
E	100.00	81.21	94.34	93.62	78.48	84.11
F	80.00	82.61	88.68	80.00	73.85	83.33
G	50.00	78.02	83.67	79.71	73.21	78.79
H	90.00	80.77	73.33	81.82	84.00	83.70
I	62.96	73.71	81.58	88.24	74.36	88.89
J	80.00	76.56	88.89	73.85	81.08	82.14
K	87.50	82.29	78.18	75.93	86.36	85.92
L	80.00	76.65	83.02	89.53	78.69	83.72
M	66.67	83.06	87.27	78.57	85.71	82.80
N	68.29	78.91	92.59	80.88	82.61	78.95
O	100.00	84.78	86.67	81.82	83.61	83.33
평균	**80.20**	**79.54**	**82.65**	**81.71**	**80.09**	**83.65**
총합 평균	**81.31**					

평균 검증 점수는 80-84%에 분포하여 있고, 총합 평균은 81.31%로 비교적 고른 점수를 보였다. 최종 검증인 9월 18일 검증에서는 83.65%로 최초 검증보다 3% 이상의 향상률을 보였다.

<표 15> 구어 목적격 무형 대용어 복원 분석 말뭉치 정답 세트 검증 결과

검증일 / 작업자	10월23일	10월30일	11월6일	11월13일	11월27일
A	87.93	85.71	76.15	66.67	67.70
B	92.13	80.56	85.94	78.82	83.55
C	82.58	76.87	-	69.35	72.65
D	76.60	-	63.46	74.24	74.29
E	84.62	79.37	66.67	71.43	73.68
F	92.00	88.24	81.36	63.16	81.17
G	79.17	77.61	79.56	68.57	75.60
H	93.65	77.17	75.00	66.67	74.45
I	82.81	72.22	70.00	70.97	70.34
J	83.93	100.00	79.09	81.05	78.90
K	100.00	93.68	79.09	69.01	81.40
L	78.26	87.20	67.82	84.15	75.68
M	79.49	90.91	75.38	75.86	69.94
N	76.92	91.03	92.23	94.44	89.36
O	75.76	87.50	69.53	76.19	77.19
평균	**84.39**	**84.86**	**75.81**	**74.04**	**76.39**
총합 평균	**79.10**				

구어 목적격 무형 대용어 복원 말뭉치의 평균 검증 점수는 74-85%에 분포하여 있고, 총합 평균은 79.10%이다. 구어 말뭉치는 공적 독백, 공적 대화, 사적 대화 등으로 구성되어 있다. 작업자들은 작업 초반 공적 독백(뉴스) 문서를 대상으로 목적격 무형 대용어 복원 분석을 진행하였는데, 공적 독백 문서는 문어와 유사한 특성을 지니므로 작업 초반의 정답 세트 점수는 문어의 정답 세트 점수와 유사하게 나타났다. 이후 공적 대화 문서가 할당되면서 점수 변동 폭이 증가하였으나 집중 정제 기간 동안 해당 문서에 대한 정제를 실시함으로써 최종 검증 점수를 향상시키고자 하였다.

목적격 무형 대용어 복원 분석의 검증 점수에는 대상 술어 탐지 점수와 선행어 일치 점수가 결합되어 있어 구문 분석의 검증 점수에 비해 점수가 낮은 것으로 추정된다. 대상 술어 탐지는 정답 세트와 비교해 목적격 복원 대상인 술어를 얼마나 정확하게 찾았는지를 평가하는 것이다. 한편 선행어 일치는 생략된 목적어에 의미를 부여하는 선행어를 정답 세트와 비교해 얼마나 정확하게 찾았는지를 평가한다. 서술어 일치율은 선행어 일치율에 비해 점수가 높다. 11월 27일에 최종적으로 검증한 결과에서 서술어

일치율은 91%, 선행어 일치율은 57%로 나타났다.

낮은 선행어 일치율은 맥락 의존성에 기인한다. 목적격 무형 대용어 복원 시 선행어가 동일 문장이나 인접한 문장에 있을 수도 있지만, 선행어가 멀리 떨어진 문장에 위치하는 경우도 많다. 특히, 구어에서는 문장 성분이 빈번하게 생략되며,[41] 선행어 복원의 맥락 의존도가 높다. ‘본 프로그램에서 파헤쳤습니다’라는 발화가 있을 때, ‘파헤쳤습니다’의 생략된 목적어로 ‘사건’이나 ‘사실’ 등을 생각할 수 있다. 둘 중 하나로 정답을 특정하기는 어렵지만 본 사업의 검증에서는 복수 정답을 허용하지 않았다. 따라서 정답 세트는 ‘사건’으로 분석하고 작업자는 ‘사실’로 분석하면 이 역시 오류로 간주된다. 선행어 불일치에는 오답이 아닌 오류도 포함되어 있는 것이다. 단일 정답 방식에서는 선행어 일치 점수가 필연적으로 낮고, 이것이 목적격 무형 대용어 복원 전체 검증 점수에도 부정적인 영향을 미쳤다.

41) ‘4.1. 말뭉치 구축’에서 언급한 바와 같이, 문어 말뭉치가 구어 말뭉치에 비해 총 어절 수가 두 배가량 많음에도 불구하고, 목적격 복원 주석 개수가 구어 말뭉치에서 더 많이 집계되었다는 것을 통해 구어 말뭉치에서 문장 성분의 생략이 빈번함을 알 수 있다.

5.3. 품질 수준

말뭉치 최종 납품 전 집중 정제 기간 동안 각 구축 기관은 말뭉치 주석 결과에 대한 정제를 실시하였다. 구문 분석 말뭉치 정제는 오류 유형에 부합하는 문장을 추출하여 검수하는 방식으로 진행되었고, 무형 대용어 복원 말뭉치 정제는 구축된 말뭉치 전수를 검수하는 방식으로 진행되었다.

(1) 구어 구문 분석 말뭉치

구문 분석 말뭉치는 총 1-4차에 걸쳐 구축되었다. 이 가운데 1-2차 말뭉치는 7월 27일부터 9월 25일까지 구축된 말뭉치로, 작업 초반에 구축되었기 때문에 작업 중반 개정된 지침 내용이 반영되지 않았다. 따라서 1-2차 구축 말뭉치 492,474어절을 대상으로 정제 작업을 진행하였다. 정제를 위한 선행 작업으로 데이터를 CoNLL 형식으로 추출하고, 해당 데이터 분석을 통해 오류 유형을 정형화한 뒤,[42] 오류 유형 조건에 부합하는 검수 대상을 추출하였다. 오류 유형에는 지침이 수정되면서 발생한 오류와 작업자들이 고빈도로 생성한 오류를 모두 포함하였다. 관리 기관은 구축 기관에게 전달받은 오류 유형을 바탕으로 해당 조건을 포함하고 있는 오류 후보 문장을 전수 추출하여 구축 기관이 해당 문장을 검토, 정제할 수 있도록 지원하였다. 이러한 과정을 거쳐 추출된 검수 대상 어절은 총 67,707어절이다.

<표 16> 구어 구문 분석 말뭉치 정제 대상 및 추정 오류

정제 대상	정제 기간	총 어절 수	정제 대상 정답 세트 점수	추정 오류율	추정 오류 어절 수	검수 어절 수
1-2차 구축 말뭉치	11/23-12/17	492,474	88.17	11.83	58,260	67,707

구축 기관은 67,707어절을 전수 검수하였고, 그 가운데 오류를 포함하고 있는 19,146어절을 수정하였다. 따라서 전체 어절 수 대비 정제 어절 수의 비율은 3.89%이다. 관리 기관은 정제가 이루어진 19,146어절과 정제가 이루어지지 않은 473,328어절의 샘플 검수를 통해 각각의 품질을 산출하였고, 그 결과 정제 어절의 품질은 98.57, 비정제 어절의 품질은 98.20임을 확인하였다.

42) 오류 유형의 예시는 '5.1.6. 정제'에서 기술하였다.

<표 17> 구어 구문 분석 말뭉치 정제 결과

총 어절 수	정제 어절 수	정제율	정제 어절 품질	비정제율	비정제 어절 품질
492,474	19,146	3.89	98.57	96.11	98.20

관리 기관은 이러한 정제 결과를 바탕으로, 구어 구문 분석 말뭉치에 대한 품질을 검증하였다. 품질 점수는 두 가지 방법을 적용하여 추정하고, 두 점수의 조화 평균값으로 최종 품질을 산출하였다.

첫 번째 방법은 정제량을 기반으로 오류율을 추정하는 것으로, 정제 대상이 모든 오류를 포함한다고 가정하는 경우이다. 말뭉치에서 나타난 오류는 모두 정제했다는 가정이므로 해당 오류율을 기반으로 산출한 품질 점수는 추정 품질의 최댓값이 된다. 관리 기관은 실제로 오류 정제가 잘 이루어졌는지, 정제하지 않은 어절에는 오류가 포함되어 있지 않은지 여부를 검증하고자 정제 어절 및 비정제 어절에 대한 샘플링 검수를 실시하여 품질을 실측하였다. 그 결과 정제 어절의 품질은 98.57, 비정제 어절의 품질은 98.20으로 나타났다. 따라서 아래와 같은 수식을 통해 1-2차 구축 말뭉치에 대한 오류율 산출이 가능하며, 전체 어절에 남아 있는 오류율은 1.79%가 된다.

$$\text{정제량 기반 추정 오류율} = \frac{(\text{정제율} \times \text{정제 어절 오류율}) + (\text{비정제율} \times \text{비정제 어절 오류율})}{100}$$

<수식 2> 구어 구문 분석 말뭉치의 정제량 기반 오류율 추정 수식

두 번째 방법은 정답 세트 검증을 기반으로 오류율을 추정하는 것으로, 정답 세트에 의한 품질 수준이 정제 대상 외에 여전히 균질하다고 가정하는 경우이다. 따라서 정답 세트 점수를 기반으로 오류 어절 수를 예측하고, 추정 오류 어절 수와 실제 정제된 어절 수의 차를 구하는 방법으로 오류율 산출이 가능하다. 비정제 어절에도 오류가 포함되어 있다는 가정이므로 해당 오류율을 기반으로 산출한 품질 점수는 추정 품질의 최솟값이 된다. 아래와 같은 수식을 통해 1-2차 구축 말뭉치 대한 오류율 산출이 가능하며, 전체 어절에 남아 있는 오류율은 8%이다.

$$\text{정답 세트 검증 기반 추정 오류율} = \frac{\text{추정 오류 어절 수} - (\text{정제 어절 수} \times (\frac{\text{정제 어절 품질}}{100}))}{\text{전체 어절 수}} \times 100$$

<수식 3> 구어 구문 분석 말뭉치의 정답 세트 검증 기반 오류율 추정 수식

말뭉치의 최종 품질은 최대 추정 품질 값과 최소 추정 품질 값의 조화 평균으로 산출하였다.

최대 추정 품질 (max) = 100-정제량 기반 추정 오류율 (%)
최소 추정 품질 (min) = 100-정답 세트 검증 기반 추정 오류율 (%)
$\text{추정 품질 평균} = \frac{2 \times (\text{최솟값} \times \text{최댓값})}{\text{최솟값} + \text{최댓값}} (\%)$

<수식 4> 구어 구문 분석 말뭉치의 품질 수준 추정 수식

최대 추정 품질은 98.21, 최소 추정 품질은 92.00이므로 조화 평균값 95.01이 구어 구문 분석 1-2차 말뭉치의 최종 품질 값이 된다.

<표 18> 구어 구문 분석 말뭉치의 품질 수준 산출 결과

정제 대상 말뭉치	총 어절 수	최소 추정 품질	최대 추정 품질	**추정 품질 평균**
1-2차 말뭉치	492,474	92.00	98.21	**95.01**

구어 구문 분석 말뭉치는 1-4차에 걸쳐 총 1,006,448어절로 구축되었다. 그 가운데 정제 작업이 이루어진 1-2차 말뭉치(492,474어절)에 대해서는 위와 같은 과정을 거쳐 품질 추정을 실시하였다.

한편, 정제 대상에서 제외한 3-4차 말뭉치(513,974어절)에 대해서는 전체 어절에 대한 샘플링 검수를 통해 품질 점수를 실측하였다. 3-4차 말뭉치는 10월 5일부터 11월 20일까지 구축한 말뭉치로, 지침 개정이 이루어진 이후에 구축되었으며, 해당 기간에는 작업자들의 숙련도가 일정 수준 이상으로 향상되었으므로 오류가 적을 것이라 판단하였다. 해당 기간 정답 세트 검증 점수는 86.97이지만, 작업 결과물의 품질이 하락한 것은 아니며, 정답 문서의 신뢰도가 낮아진 것으로 해석해야 할 것이다. 정답 세트의 정답 문서는 검수자가 작업 초반에 모두 선구축해 놓은 것이므로 개정된 지침 내용이 반영

되어 있지 않다. 따라서 작업자들이 개정된 지침을 기준으로 작업한 어절은 오류 처리되어 검증 점수가 산출되었다. 실제로 정답 문서와 작업자 문서 간의 불일치 항목을 확인한 결과, 작업자의 작업 결과가 지침에 더 부합함을 확인하였다. 이에 따라 3-4차 구축 말뭉치에 대한 정답 세트 검증 점수는 효용이 낮다고 판단하여, 전체 어절에 대한 샘플링 검수를 통하여 품질 점수를 추정하였다. 3-4차 구축 말뭉치 513,974어절의 1% 샘플 검수를 진행한 결과, 오류 어절은 199어절로, 품질 점수 96.12가 산출되었다. 이러한 결과에 대하여 구축 기관은 3-4차 구축 말뭉치의 품질을 96.12%로 추정하였다.

(2) 목적격 무형 대용어 복원 말뭉치

목적격 무형 대용어 복원 말뭉치 정제는 문서를 전수 검수하는 방식으로 진행되었다. 따라서 총 어절 수와 검수 어절 수가 동일하며, 문어 말뭉치와 구어 말뭉치를 모두 정제 대상으로 하였다. 무형 대용어 복원 말뭉치는 말뭉치의 특성상 어절 수가 아닌 주석 수를 기준으로 오류율 및 오류 주석 수를 추정하였다.

<표 19> 목적격 무형 대용어 복원 분석 말뭉치 정제 대상 및 추정 오류

	정제 기간	총 어절 수	전체 주석 수	정제 대상 정답 세트 점수	추정 오류율	추정 오류 주석 수
문어 말뭉치	9/25-10/9	2,000,213	79,364	81.31	18.69	14,833
구어 말뭉치	11/27-12/16	1,006,448	85,422	79.10	20.90	17,853

문어 말뭉치 정제 결과, 전체 79,364개 주석 가운데 8,595개 주석이 수정되어 정제율은 10.83%였다. 관리 기간은 정제가 이루어진 주석과 그렇지 않은 주석에 대한 샘플 검수를 통해 각각의 품질을 산출하였고, 그 결과 정제 주석의 품질은 96.31, 비정제 주석의 품질은 95.38임을 확인하였다. 구어 말뭉치는 전체 85,422개 주석 가운데 5,155개 주석이 수정되어 6.03%의 정제율을 보였다. 정제 주석의 품질은 97.20, 비정제 주석의 품질은 95.73이었다.

<표 20> 목적격 무형 대용어 복원 분석 말뭉치 정제 결과

	전체 주석 수	정제 주석 수	정제율	정제 주석 품질	비정제율	비정제 주석 품질
문어 말뭉치	79,364	8,595	10.83	96.31	89.17	95.38
구어 말뭉치	85,422	5,155	6.03	97.20	93.97	95.73

목적격 무형 대용어 복원 말뭉치의 정제 내용은 서술어/선행어 추가, 서술어/선행어 삭제, 선행어 수정으로 유형화된다. 각각의 세부 정제 내용을 표로 정리하면 다음과 같다.

<표 21> 목적격 무형 대용어 복원 분석 말뭉치 정제 내용

	추가된 서술어 수	삭제된 서술어 수	추가된 선행어 수	삭제된 선행어 수	수정된 선행어 수	총합
문어 정제	1,721	957	1,721	957	3,239	**8,595**
구어 정제	1,060	622	1,060	622	1,791	**5,155**

목적격 무형 대용어 복원 말뭉치에 대한 품질 점수 역시 최댓값과 최솟값의 두 가지 방법으로 품질 수준을 추정하였다.

첫 번째 방법은 정제량을 기반으로 오류율을 추정하는 것으로, 정제 대상이 모든 오류를 포함한다고 가정하는 경우이다. 해당 오류율을 기반으로 산출한 품질 점수는 추정 품질의 최댓값이 된다. 아래와 같은 수식을 통해 오류율 산출이 가능하며, 그 결과 문어는 4.52%, 구어는 4.18%의 오류율을 보였다.

$$\text{정제량 기반 추정 오류율} = \frac{(\text{정제율} \times \text{정제 주석 오류율}) + (\text{비정제율} \times \text{비정제 주석 오류율})}{100}$$

<수식 5> 목적격 무형 대용어 복원 분석 말뭉치의 정제량 기반 오류율 추정 수식

두 번째 방법은, 최소 품질 값을 추정하는 것으로, 정답 세트에 의한 품질 수준이 정제 대상 외에 여전히 균질하다고 가정하는 경우이다. 아래와 같은 수식을 통해 오류율 산출이 가능하며, 문어 말뭉치의 오류율은 8.26%, 구어 말뭉치의 오류율은 15.03%로 각각 나타났다.

$$\text{정답 세트 검증 기반 추정 오류율} = \frac{\text{추정 오류 주석 수} - (\text{정제 주석 수} \times \frac{\text{정제 주석 품질}}{100})}{\text{정제 전 전체 주석 수}} \times 100$$

<수식 6> 목적격 무형 대용어 복원 분석 말뭉치의 정답 세트 검증 기반 오류율 추정 수식

말뭉치의 최종 품질은 최대 추정 품질 값과 최소 추정 품질 값의 조화 평균으로 산출하였다.

최대 추정 품질 (max) = 100-정제량 기반 추정 오류율 (%)
최소 추정 품질 (min) = 100-정답 세트 검증 기반 추정 오류율 (%)
$\text{추정 품질 평균} = \frac{2 \times (\text{최솟값} \times \text{최댓값})}{\text{최솟값} + \text{최댓값}}$ (%)

<수식 7> 목적격 무형 대용어 복원 분석 말뭉치의 품질 수준 추정 수식

위와 같은 방식으로 문어 말뭉치와 구어 말뭉치에 대한 최대 품질 값, 최소 품질 값, 품질 평균을 각각 추정한 결과 아래 표와 같은 결과가 산출되었다. 문어 말뭉치의 최종 추정 품질은 93.57, 구어 말뭉치의 최종 추정 품질은 90.07이다.

<표 22> 목적격 무형 대용어 복원 분석 말뭉치의 품질 수준 산출 결과

	총 어절 수	최소 추정 품질	최대 추정 품질	추정 품질 평균
문어 말뭉치	2,000,213	91.74	95.48	**93.57**
구어 말뭉치	1,006,448	84.97	95.82	**90.07**

5.4. 산출물

약 100만 어절 규모의 구어 구문 분석 말뭉치와 약 300만 어절 규모(문어 200만 어절, 구어 100만 어절)의 목적격 무형 대용어 복원 말뭉치를 제이슨(JSON) 형식의 파일로 납품하였다. 이 외에도 사업 결과 보고서 30부를 산출물로 납품하였다.

5.5. 사업 보고

ㅇ 착수 보고
- 사업 세부 항목, 사업 수행 방법, 추진 일정 계획, 각 분야별 참여 인력 등을 포함한 세부 과업 수행 계획을 작성하여 보고하였다.

ㅇ 월별 보고
- 사업의 추진 현황을 매달 기록하여 주관 기관에 제출하였다.

ㅇ 중간 보고
- 10월 14일에 온-나라 화상 회의로 진행하였다.

ㅇ 최종 보고
- 12월 15일에 온-나라 화상 회의로 진행하였다.

제 3 장

결론

1. 사업 요약

본 사업은 구문 분석 말뭉치 및 무형 대용어 복원 말뭉치 구축을 목적으로 하였다.

구문 분석 말뭉치의 경우, 구어 약 100만 어절을 대상으로 구어 구문 분석 말뭉치를 구축하였다. 말뭉치 구축을 위해 '2019년 국립국어원 구문 분석 말뭉치 지침'에 구어 지침을 추가하여 지침을 정교화하고 지침의 완성도를 제고하였다.

무형 대용어 복원 말뭉치의 경우, 문·구어 통합 약 300만 어절을 대상으로 목적격 무형 대용어 복원 말뭉치를 구축하였다. 목적격 무형 대용어 복원 말뭉치 구축 지침은 주격 무형 대용어 복원 말뭉치 구축 지침과의 일관성을 유지하되 목적격 복원의 특성을 고려하여 새롭게 수립하였다.

구문 분석 말뭉치와 목적격 무형 대용어 복원 말뭉치는 말뭉치의 품질 확인을 위해 각각 3차에 걸쳐 납품하였다. 구축된 말뭉치의 총 어절 수, 총 문장 수, 주석 수는 아래 표와 같다.

<표 23> 구어 구문 분석 말뭉치 주석 결과

총 어절 수	총 문장 수	구문 태그 수	기능 태그 수	지배소를 갖는 어절 수	의존소를 갖는 어절 수
1,006,448	221,489	1,006,448	404,047	1,006,448	531,136

(지배소 값에는 ROOT값 포함)

<표 24> 목적격 무형 대용어 복원 분석 말뭉치 주석 결과

	총 어절 수	총 문장 수	복원 대상 서술어 수	선행어 수
문어	2,000,213	150,082	40,446	40,446
구어	1,006,448	221,489	43,149	43,149
총합	**3,006,661**	**371,571**	**83,595**	**83,595**

2. 의의 및 기대 효과

인공 지능 산업 발전을 위한 대규모 고품질 우리말 자원 수요가 증대되고 있다. ‘구문 및 무형 대용어 복원 말뭉치 연구 분석’은 이에 대한 이해를 바탕으로 국어 자원의 활용도와 가치 제고를 위해 수행되었다. 본 사업의 결과물은 국내 표준화 작업에 기여하고 참조 기반 자료가 될 수 있는 고품질 언어 정보 부가 말뭉치로서 의의를 갖는다.

대규모 국어 빅데이터 구축은 국가 언어 자원의 활용을 가능하게 하며, 언어 자원 통합 체계 구축을 위한 발판을 마련해 준다. 본 사업을 통해 구축된 구어 구문 분석 말뭉치 및 목적격 무형 대용어 복원 말뭉치는 4차 산업 및 언어적 연구에서 즉각적인 전산 처리가 가능한 말뭉치로 국가 언어 자원 활용성 확산에 기여하였다.

3. 향후 연구

구어 구문 분석 말뭉치 연구 분석 및 무형 대용어 복원 말뭉치 연구 분석은 타 분석 층위 말뭉치와 연계되는 연구라는 점에서 통합 말뭉치의 활용 가능성을 확인시켜 주었다. 구어 구문 분석 말뭉치는 기구축된 문어 구문 분석 말뭉치와의 통합이 가능하며, 이를 통해 구문 분석 말뭉치의 규격화를 수행할 수 있게 되었다. 목적격 무형 대용어 복원 말뭉치는 기구축된 주격 무형 대용어 복원 말뭉치와 연계되며 동일한 말뭉치를 대상으로 주석 작업이 진행되었으므로 결과의 활용도가 높다.

향후에는 현재 연구가 이루어지지 않은 분석 층위를 다각적으로 검토하여 통합 말뭉치 구축 범위를 확장하고, 전산 처리 분야의 기술과 경험을 바탕으로 국어 빅데이터 구축 사업의 전문성을 더욱 제고해야 할 것이다. 또한 문어 자료에 비해 비정형화된 특성을 갖는 구어 자료 구축 및 지침 수립에 대한 세밀한 연구가 수반된다면 말뭉치 품질 향상에 기여할 것이다.

<Abstract>

Research and Analysis of Parsed Corpus and Zero Anaphora Resolution Corpus

This project aims to construct a parsed corpus and an object zero anaphora resolution corpus. The corpus parsing focuses on spoken language while the zero anaphora resolution annotates a deleted object in both spoken and written language. The major tasks and goals of the project are as follows:

To establish the guidelines for the parsed corpus of spoken Korean

: The guidelines were built on the written annotation scheme of 2019 NIKL parsed corpus. In the following version of the guidelines in 2020, colloquial characteristics were fully considered.

To construct a parsed corpus of spoken Korean

: Based on the guidelines, a parsed corpus with one million words was constructed. A parsed corpus was conducted using the 2019 NIKL part-of-speech tagged corpus.

To establish the guidelines for object zero anaphora resolution corpus

: The guidelines were built on the annotation scheme of 2019 NIKL subject zero anaphora resolution corpus. The guidelines for both written and spoken language corpus are described in 2020.

To construct an object zero anaphora resolution corpus

: Based on the guidelines, an object zero anaphora resolution corpus with a scale of three million words was constructed: two million words of written language and one million spoken language.

To verify the analysis of the constructed corpus

: We made a verification method and system for consistency and accuracy of the entire data. In order to minimize the error rate, the corpus was submitted three times, and the quality was improved by reflecting the feedback. In addition, the annotation was evaluated based on the answer set constructed by the verifiers. The results were reported weekly, and individual training was provided to workers with low scores to investigate the quality of the corpus.

The result of this project is significant as a corpus with high-quality language information that can be standardized and reference-based data.

Keywords: parsing, parsing of spoken language, zero anaphora resolution, object zero anaphora resolution, corpus

Project Director: Kwak Yongjin(IIR TECH Inc.)

사업 책임자	곽용진 ((주)이르테크 대표이사)
사업 참여자	김한샘 (연세대학교 언어정보연구원 교수)
	이숙의 (충남대학교 인문과학연구소 전임연구원)
	유현경 (연세대학교 국어국문학과 교수)
	봉미경 (연세대학교 언어정보연구원 교수)
	김진수 (충남대학교 국어국문학과 교수)
	김재훈 (한국해양대학교 제어자동화공학부 교수)
	이공주 (충남대학교 전파정보통신공학과 교수)
	김유섭 (한림대학교 소프트웨어융합대학 교수)
	김학수 (건국대학교 컴퓨터공학부 교수)
	나승훈 (전북대학교 컴퓨터공학부 교수)
	류법모 (부산외국어대학교 경찰정보보호학부 교수)
	이찬영 (연세대학교 국어국문학과 박사 과정)
	박혜진 (연세대학교 국어국문학과 박사 과정)
	신아영 (연세대학교 국어국문학과 박사 과정)
	정주연 (연세대학교 국어국문학과 박사 과정)
	천성호 (연세대학교 국어국문학과 박사 과정)
	이정은 (충남대학교 국어국문학과 박사 과정)
	장지현 (충남대학교 국어국문학과 박사 과정)
	정민경 (층남대학교 국어국문학과 석사 과정)
	정해영 ((주)이르테크 선임연구원)
	이지연 ((주)이르테크 선임연구원)
	박재은 ((주)이르테크 대리)
	최지선 ((주)이르테크 대리) 외 53명

담당 연구원 이승재(국립국어원 언어정보과장)
서샛별(국립국어원 언어정보과 학예연구사)
김명주(국립국어원 언어정보과 연구원)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2020년 12월 21일
발행일: 2020년 12월 21일
인 쇄: 세종기획

※ 이 책은 국립국어원의 용역비로 수행한 '구문 및 무형 대용어 복원 말뭉치 연구 분석' 사업의 결과물을 발간한 것입니다.