

국립국어원 2019-01-50

발 간 등 록 번 호
11-1371028-000767-01

주격 무형 대용어 복원 말뭉치 구축

사업 책임자
곽 용 진

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구 용역 계약에 따라 ‘주격 무형 대용어 복원 말뭉치 구축’에 관한 최종 보고서를 작성하여 제출합니다.

■ 사업기간: 2019년 7월 15일 ~ 2020년 1월 15일

2020년 1월 15일

사업 책임자: 곽용진((주)이르테크)

사업 수행 기관	주식회사 이르테크 충남대학교 산학협력단
사업 책임자	곽용진(이르테크)
사업 참여자	이숙의(충남대), 김진수(충남대), 김정인(충남대), 이정은(충남대), 장지현(충남대), 정민경(충남대), 임보람(충남대), 강동훈(충남대), 구름(충남대), 김인혜(충남대), 문찬국(충남대), 박준형(충남대), 안수빈(충남대), 유민수(충남대), 이예지(충남대), 허정(충남대), 정해영(이르테크), 이순미(이르테크), 장호림(이르테크), 이선덕(이르테크), 홍은기(이르테크), 한문성(이르테크), 김상선(이르테크), 서보원(이르테크), 김영환(이르테크), 박재은(이르테크)

<사업 수행자>

주식회사 이르테크 · 충남대학교 산학협력단

사업 책임자	곽용진((주)이르테크)
사업 참여자	이숙의(충남대학교 인문과학연구소 전임연구원)
	김진수(충남대학교 국어국문학과 교수)
	김정인(충남대학교 국어국문학과 박사)
	이정은(충남대학교 국어국문학과 박사)
	장지현(충남대학교 국어국문학과 박사)
	정민경(충남대학교 국어국문학과 석사)
	임보람(충남대학교 국어국문학과 석사)
	강동훈(충남대학교 국어국문학과 학사)
	구름(충남대학교 국어국문학과 학사)
	김인혜(충남대학교 국어국문학과 학사)
	문찬국(충남대학교 국어국문학과 학사)
	박준형(충남대학교 국어국문학과 학사)
	안수빈(충남대학교 국어국문학과 학사)
	유민수(충남대학교 국어국문학과 학사)
	이예지(충남대학교 국어국문학과 학사)
	허정(충남대학교 국어국문학과 학사)
	정해영((주)이르테크)
	이순미((주)이르테크)
	장호림((주)이르테크)
	이선덕((주)이르테크)
	홍은기((주)이르테크)
	한문성((주)이르테크)
	김상선((주)이르테크)
	서보원((주)이르테크)
	김영환((주)이르테크)
	박재은((주)이르테크)

주격 무형 대용어 복원 말뭉치 구축

본 사업은 한국어 문장에서 용언의 필수격, 특히 주격이 생략된 경우, 생략된 부분을 채워 줄 수 있는 문장 성분을 찾아 복원하는 사업이다. 문장에서 생략된 주어를 복원하면 문서에 등장하는 개체 정보를 명확하게 규명할 수 있게 된다. 따라서 주격 무형 대용어 복원 작업은 문서 내에서 전달하고자 하는 의미를 명확하게 해 주며, 정보의 일관성을 향상시킨다. 생략어의 복원 결과물은 정보의 검색 및 추출, 질의응답, 기계 번역 등의 분야에서 유용하게 활용될 수 있다.

본 사업은 주격 무형 대용어 복원에 대한 기본 지침 수립과 이를 바탕으로 한 주격 무형 대용어 복원 데이터 구축에 중점을 두었다. 이에 따른 주요 과업과 연구의 성과는 다음과 같다.

- 주격 무형 대용어 복원 말뭉치 지침 수립

자연 언어 처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서 크게 벗어나지 않도록 지침을 수립하였다. 생략 술어 탐지 및 선행어 결정에 관한 기본 지침을 수립하고, 다어절 선행어, 절 선행어, 보조 용언 구문, 의존 명사 구문 등에 관한 세부 지침을 추가하였다. 구어 텍스트는 주어가 비실현되는 경우가 빈번하고, 축약형이 등장하는 등 문어 텍스트와는 다른 특수성을 지니므로 구어 지침을 별도로 수립하여 지침을 개선·보완하였다. 지침 개선에 반영될 수 있는 여러 예외 상황에 대한 검토와 시험을 지속적으로 진행하여 작업자 간의 일관성을 높이려고 하였다.

- 주격 무형 대용어 복원 말뭉치 구축

말뭉치 구축 수행 도구를 개발하여 안정적인 클라우드 환경에서 작업 환경을 구성하였다. 작업자는 전처리 단계, 탐지 단계, 복원 작업 단계, 자가 진단 단계를 거쳐 작업을 진행하였으며, 외부 인력의 접근을 차단하여 자료 보안 기준을 준수하였다. 주격 무형 대용어 복원 말뭉치 결과물로 납품된 데이터의 총 어절 수는 3,025,769어절(7,688개 문서)이며, 이 가운데 문어는 2,019,322어절(7,265개 문서), 구어는 1,006,447어절(423개 문서)이다.

- 주격 무형 대용어 복원 말뭉치 검증

주격 무형 대용어 복원 말뭉치는 내부 검증과 외부 검증을 통한 철저한 검수 과정을 거쳐 납품되었다. 내부 검증으로는 작업자가 스스로 자신의 작업 내용 오류를 수정할 수 있는 작업자 검증과 미작업 대상에 대한 자동 검수가 이루어지는 기계적 검증을 함께 실시하였다. 작업자와 검수자 간의 소통을 통한 절차적 검증 및 관리자의 관리적 검증도 수행하였다. 외부 검증으로는 말뭉치 형식 검증과 내용 오류 검증을 실시하였는데, 주관 기관이 정답 세트를 제시하고, 수행 기관이 정답 세트와의 일치도를 분석하였다.

무형 대용어 복원 말뭉치는 기존의 구축 사례가 거의 없고, 구축된 말뭉치에 대한 국어학적 검토나 연구가 많지 않았다. 본 사업에서 구축한 말뭉치는 현재까지 구축된 가장 대량의 무형 대용어 복원 말뭉치라는 점, 4차 산업 및 언어 연구에서 즉각적인 전산 처리가 가능한 말뭉치라는 점에서 의의가 있다.

주요어 : 무형 대용어, 주격 무형 대용어, 무형 대용어 복원, 말뭉치

차 례

제1장 사업 개요

1. 사업 목적	2
2. 사업 범위	2
3. 사업 수행	3

제2장 사업 수행 내용

1. 수행 환경 구성	5
1.1. 원시 데이터	5
1.2. 수행 환경 구성	5
1.3. 초기 지침 및 작업자 교육	7
1.4. 데이터 납품, 검증 정책	8
2. 지침	10
2.1. 한국전자통신연구원 생략어 복원 태깅 가이드라인 분석	10
2.2. 주격 무형 대용어 복원 지침	13
2.3. 문어와 구어 태깅의 실례	20
3. 데이터 구축 수행 도구 활용	50
3.1. 시스템 설치 및 구성	50
3.2. 자료 보안 및 외부 인력 접근 제어	50
3.3. 구축 도구 활용	51
3.4. 구축 절차	53
4. 말뭉치 구축 및 납품	62
4.1. 말뭉치 구축	62
4.2. 말뭉치 납품	64

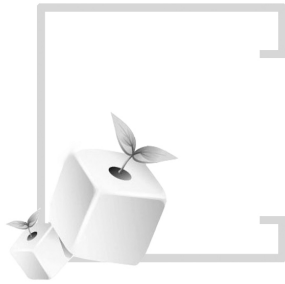
차 례

5. 검증 및 산출물 보고	66
5.1. 내부 검증	66
5.2. 외부 검증	68
5.3. 산출물	71
5.4. 사업 보고	71

제3장 향후 계획

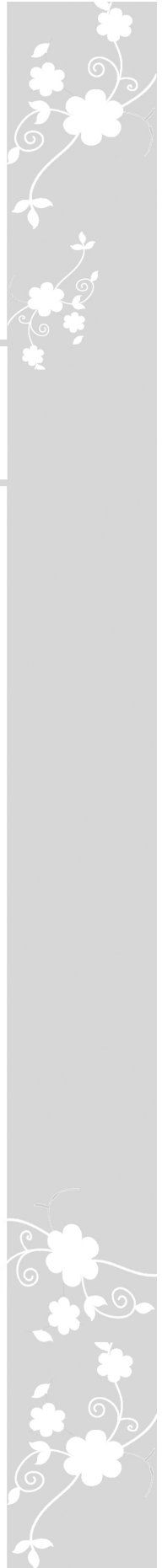
1. 개선 방향	73
2. 기대 효과	75

참고문헌	77
Abstract	79



제 1 장

사업 개요



1. 사업 목적

본 사업의 목적은 문장에서 생략된 주어를 복원함으로써 문서에 등장하는 개체와 그에 대한 정보를 명확하게 규정하는 것이다. 한국어에서 주어는 필수 문장 성분이지만 선행어의 제시 혹은 암묵적인 용인 등으로 인해 문장에서 생략되는 경우가 있다. 생략된 주어의 복원 작업은 문서 내에서의 명확한 의미 전달, 정보의 일관성 향상 등을 위해 필요한 작업이다. 특히 생략어의 복원 결과물은 정보의 검색 및 추출, 문서 요약, 질의응답, 기계 번역 등의 분야에서 유용하게 활용될 수 있다.

최근 4차 산업 혁명으로 인해 대규모, 고품질 우리말 자원의 수요가 증대되었다. ‘주격 무형 대용어 복원 말뭉치’ 구축은 국어 자원의 활용도와 가치를 높일 수 있도록 민간에서 활용 가능한 국가 공공재로서의 말뭉치 구축의 일환으로, 인공지능 기술 수준 향상에 이바지할 것이다.

2. 사업 범위

본 사업은 다른 분석 말뭉치 구축 사업과는 다르게 기존 말뭉치 구축 지침의 표준이 존재하지 않고, 해외에서의 연구는 활발하나 국내 연구 및 구축이 거의 이루어지지 않은 분야의 분석 말뭉치이다. 따라서 단순한 양적, 질적 기준이 아닌 앞으로 한국어 분석 말뭉치 연구, 활용의 시금석이 되는 사업으로 말뭉치의 품질, 활용, 구축 지침의 표준을 제시하는 사업이다.

○ 최초의 표준 지침 수립

- 한국정보통신기술협회(TTA) 등 기타 관련 분야 분석 표지 및 분석 지침을 검토하고, 기존 외국어 연구 사례와 한국어 특성에 대한 지침을 비교한다.
- 기존 지침의 문제점 분석 및 보완책을 제시하여 주격 무형 대용어 복원 지침을 수립하며, 세부 지침, 파일명 부여 방식, 표지 부착 방식, 형식 등은 주관 기관과 협의한다.

○ 지침에 따른 말뭉치 가공

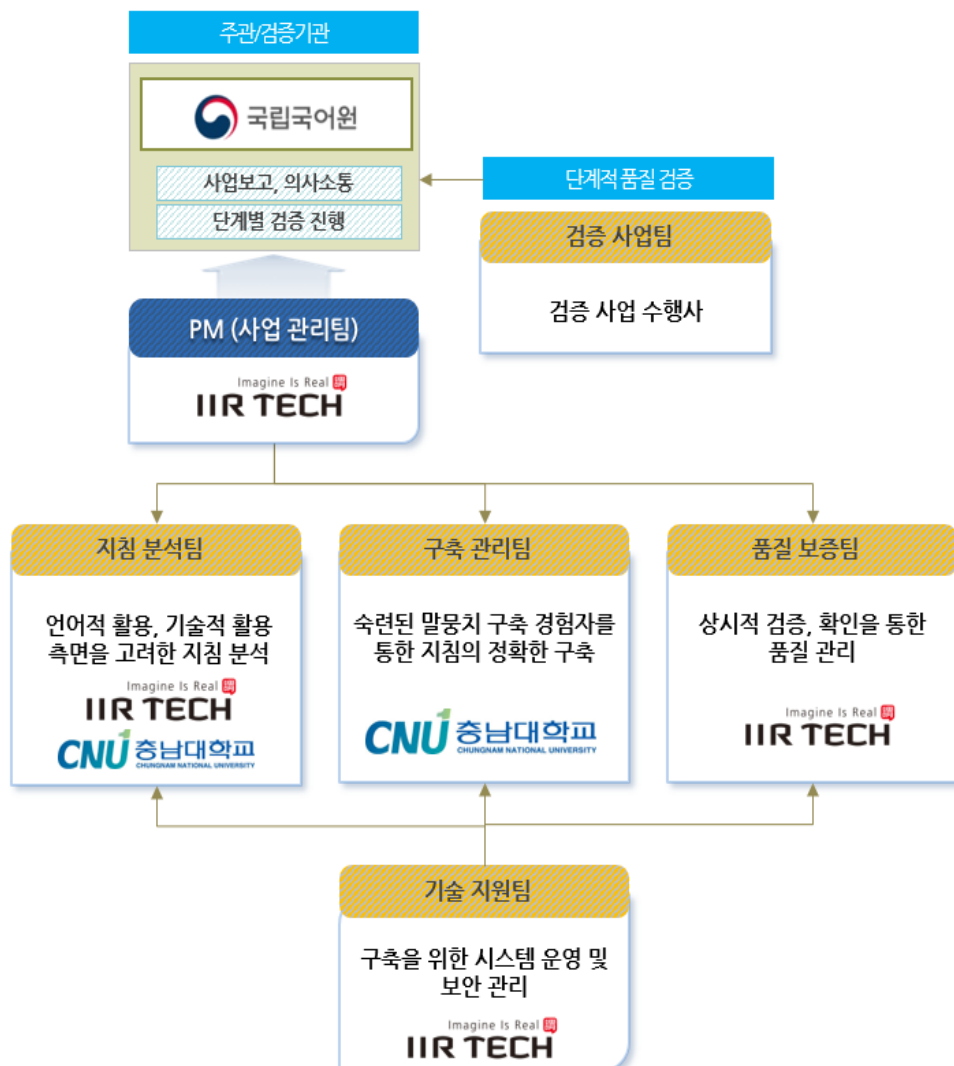
- 원시 말뭉치 300만(문어 200만, 구어 100만) 어절을 대상으로 문서 단위로 주격

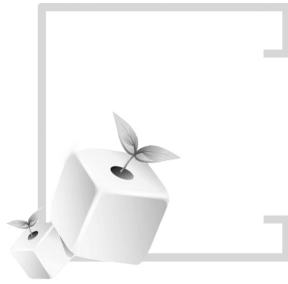
무형 대용어 복원 결과를 구축하며, 문서 내 각 서술어에 대한 주격 무형 대용어 복원 정보를 부착한다.

○ 말뭉치의 단계적 품질 검증

- 오류율 5% 이내를 달성하기 위하여 3차에 걸쳐 납품하고, <말뭉치 통합 검증> 사업단의 검증을 받아 단계적으로 품질을 점검하고 최종 결과물에 반영한다.
- 품질 검증 사업자와 지침 표준화 및 검증 데이터 형식을 표준화한다.

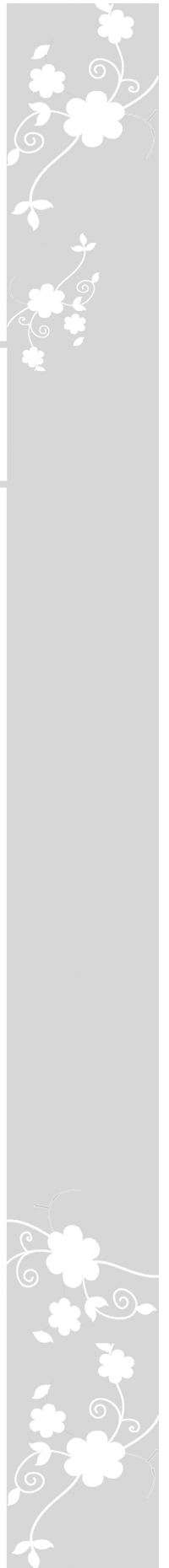
3. 사업 수행





제 2 장

사업 수행 내용



1. 수행 환경 구성

1.1. 원시 데이터

주관 기관인 국립국어원에서 제공한 원시 말뭉치 약 300만 어절(문어 2,019,322어절, 구어 1,006,447어절)을 대상으로 주격 무형 대용어 복원 분석을 실시하였다.

1.2. 수행 환경 구성

1.2.1. 데이터 구축 시스템 준비 현황

분석 작업은 사업 수행자가 개발한 웹 기반 말뭉치 주석 작업 시스템(KRONOTH Annotation System)을 사용하였다. 시스템을 통해 말뭉치의 구축 및 제어 작업이 가능하며 실시간 현황 관리가 이루어진다. 또한 인가자만이 시스템에 접근할 수 있게 설계하여 주관 기관의 보안 요구 사항을 준수하였다.

1.2.2. 수행 조직 및 인력 구성

조 직	이 름	담당 업무
사업 관리팀	곽용진	•프로젝트 관리 및 통제(PM) 및 지침 연구, 설계, 협의
	이순미	•의사 소통 지원 및 행정 지원(대관업무 및 비용 관리 등)
지침 수립팀	곽용진	•데이터 구축 지침, 표지 부착 방식 등 구축 지침 연구, 설계 •국립국어원과 구축 지침 협의 및 수립
	정해영	•지침 연구 자료 조사 및 수집 데이터 및 샘플 데이터 결과 분석
	이숙의	•구축 지침 연구, 설계 및 지침 운용 현황 확인, 개선안 수립
구축 관리팀	김진수	•지침 연구 및 연구
	이숙의	•주격 무형 대용어 말뭉치 구축/검수 작업 관리, 지침 연구
	김정인	•주격 무형 대용어 복원 정보 검수 작업 및 통제
	이정은	•주격 무형 대용어 복원 정보 검수 작업
	장지현	•주격 무형 대용어 복원 정보 검수 작업
	정민경	•주격 무형 대용어 복원 정보 검수 작업
	임보람	•주격 무형 대용어 복원 정보 태깅 작업
	강동훈	•주격 무형 대용어 복원 정보 태깅 작업
	구름	•주격 무형 대용어 복원 정보 태깅 작업
	김다은	•주격 무형 대용어 복원 정보 태깅 작업
	김인혜	•주격 무형 대용어 복원 정보 태깅 작업
	문찬국	•주격 무형 대용어 복원 정보 태깅 작업
	박준형	•주격 무형 대용어 복원 정보 태깅 작업
	안수빈	•주격 무형 대용어 복원 정보 태깅 작업
	유민수	•주격 무형 대용어 복원 정보 태깅 작업
	이예지	•주격 무형 대용어 복원 정보 태깅 작업
	허정	•주격 무형 대용어 복원 정보 태깅 작업
품질 보증팀	장호림	•구축 데이터 형식 오류 검사, 규격 검사, 검증 결과 관리
	이선덕	•구축 데이터 형식 오류 검사, 규격 검사, 검증 결과 관리
	홍은기	•대외, 대내 조직 사업 수행 협력 지원
기술 지원팀	한문성	•시스템 아키텍처 구성 및 분석 말뭉치 기술 자문
	김상선	•본문 주석 등 부가 필요 주석 개발 지원 등 시스템 맞춤 수정
	서보원	•사업 수행, 관리 및 데이터 추적 관리 지원 등 시스템 맞춤 수정
	김영환	•시스템 환경, 사용자 관리 등 시스템 운영 및 유지
	박재은	•납품 데이터 변환 등 검증, 납품용 데이터 구조 맞춤 수정

1.3. 초기 지침 및 작업자 교육

- 한국정보통신기술협회(TTA), 한국전자통신연구원(ETRI) 등에서 발간한 구문 분석 지침과 주격 무형 대용어 복원 지침을 참고하되 본 사업의 목적에 맞도록 지침을 수정·보완하여 교육을 진행하였다.
- 교육 내용은 크게 지침 교육과 시스템 사용 교육으로 나뉜다. 작업에 대한 이해도를 높이기 위해 무형 대용어 복원 및 말뭉치 구축 제반에 대한 기본 교육을 함께 실시하며, 이를 시스템에 적용하는 실습 교육을 동시에 진행하였다.
- 국어국문학 전공 교수와 박사 참여자가 교육을 주도하였고, 이하 국어국문학을 전공하는 석사 및 학사 참여자가 교육에 참여하였다. 국어학 전공자들로 인력을 구성하여 작업의 전문성을 제고하였다.
- 실습 교육은 동일 문서로 작업하였으며, 작업 결과를 기반으로 작업자별 지침 숙지도를 파악하여 결과가 저조한 작업자에 대해 재교육을 실시하는 등 실시간 인력 관리를 지원하였다.
- 구축 작업 결과 기반으로 오류 발생 사항 및 일관성이 다른 사항 등 사례 중심 교육 위주로 교육을 지속하였다.
- 작업자 교육은 사업의 의미, 목적 소개, 작업자 소양 교육 및 보안 교육을 포함하였다.
- 지침 수립 및 변경 후에는 일주일 이내에 변경된 지침에 맞게 작업자 교육 및 시스템 사용자 교육을 지속적으로 진행하는 것을 원칙으로 하였다.
- 작업 결과물은 지침 및 데이터 샘플로 활용하였으며, 발주 기관과 사례 중심 지침을 수립하는 데에 활용하였다.

교육 주제	지침 교육	시스템 교육
기본 교육	사업의 정의, 목적, 필요성 등 사업 개요 전반	시스템 제반에 대한 이해 및 보안 교육
주어가 생략된 서술어 탐지	지배소와 피지배소의 관계 이해 / 서술어의 격틀 이해	서술어 태깅
복원 대상이 아닌 서술어 배제	구문 분석 이해	서술어 삭제
선행어 탐색	선행어 후보에 대한 단계적 파악	선행어 확인 및 태깅
선행어 복원	문맥 확인	선행어 복원 태깅

〈주격 무형 대용어 복원 말뭉치 구축을 위한 교육 내용〉



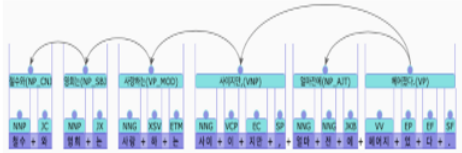
〈초기 지침 수립 및 작업자 교육 순환도〉

1.4. 데이터 납품, 검증 정책

- 데이터 검증을 위해 전문가를 통한 언어학적 검증과 시스템을 통한 기술적 검증을 동시에 진행한다.
- 국어학 분야의 전문가를 통해 데이터 오류 및 지침에 대한 예외 사항 등을 수집하고 분석하여 데이터 품질을 향상시키고, 주관 기관인 국립국어원과 예외 사례를 공유하여 지침을 개선·보완한다.
- 작업 과정 중 지침 적용이 모호하거나, 다수 작업자 간 지침 해석이 불일치하는 대상에 대해 작업자는 상위 전문 검수자에게 시스템을 이용한 검토 요청을 하여, 해당 데이터에 대한 처리 방침을 최종 결정하도록 하는 절차적 검증 단계를 거친다. 이 과정에서 데이터의 변경, 작업 이력에 대한 통계 관리가 이루어져 말뭉치의 품질이 개선되며, 데이터의 단계별 격리 저장과 각 단계별 역할 부여로 철저한 검증이 이루어진다.
- 작업한 대상의 오류 위치를 쉽게 추적하고, 작업 완료 시 텍스트 내 미작업 태깅 및 분석 대상 누락에 대한 자동 검수가 가능하도록 하는 기계적 검증을 적용하여 검증의 효율성과 작업 품질을 동시에 높인다.
- 개선된 지침을 바탕으로 작업자들에게 반복 실습 교육을 실시하여 데이터의 오류율을 최소화한다.
- 의존 구문, 의미역 분석기를 활용하여 복원 적절성을 검토한다.
- 데이터 활용성을 고려한 지침의 설계 및 보완을 원칙으로 하며, 발주 기관과의 협의 결과에 따른 말뭉치 재구축 작업으로 품질을 향상한다.

의존 구문, 의미역 분석기를 활용한 복원 적절성 검증

복원 전 분석 결과



Predicate	시도 1	재미 1
황우석(AMOD)	황우석	
배아(AMOD)	배아	
사용하다(V)		
인간(AMOD)		
배아(AMOD)		
만들었다(V)		

AI-HUB 의존구문 분석, 의미역 분석 공개API 활용

복원 후 분석 결과



Predicate	시도 1	재미 1
황우석(AMOD)	황우석	황우석
배아(AMOD)	배아	
사용하다(V)		
인간(AMOD)		
배아(AMOD)		
만들었다(V)		

<구문 분석기 결과를 통한 검증의 예>

2. 지침

2.1. 한국전자통신연구원 생략어 복원 태깅 가이드라인 분석

본 사업은 생략어 복원 지침 수립을 위해 한국전자통신연구원(ETRI)의 생략어 복원 태깅 가이드라인(3.3버전)을 참고하였다. 한국전자통신연구원 가이드라인은 주어 외에도 목적어의 생략 명사구 복원까지 포함한다. 본 사업은 생략된 주어만 복원하는 것을 목표로 하므로 주어 복원에 관한 참조 내용만 제시한다. 한국전자통신연구원 가이드라인의 주요 내용을 요약하면 다음과 같다.¹⁾

2.1.1. 개요

생략어 복원(Zero Anaphora Resolution)이란 문장에서 주어, 목적어 등의 필수 성분이 생략되어 있을 때 해당 문장 성분을 찾아내 복원해 주는 자연어 처리 문제이다. 이때 생략된 문장 성분은 ‘생략어(zero anaphora, zero pronoun, dropped pronoun)’ 또는 ‘무형 대용어’라고 하며, 생략된 문장 성분을 논항으로 갖는 동사를 ‘지배소(head)’라고 한다. 그리고 생략어가 복원되어야 할 표현을 ‘선행어(antecedent)’라고 한다. 선행어가 동일한 문서 내에 존재하는 경우에는 보통 생략 지점보다 앞서 나오며, 암묵적으로 알고 있는 대상일 경우 문장에 드러나지 않기도 한다.

2.1.2. 목표 및 범위

한국전자통신연구원(ETRI) 가이드라인의 1차적 활용은 생략된 문장 성분으로 인해 의존 구문 분석 기술에서 누락된 결과를 보완하기 위한 것이다. 생략어 복원 대상 범위는 의존 구문 관점에서 서술어에 의존 관계로 연결되지 않은 주어와 목적어에 해당된다.

2.1.3. 생략어 복원 대상

1) 한국전자통신연구원(ETRI)의 ‘생략어 복원 태깅 가이드라인’은 비공개 문서이므로, 2.1.에 등장하는 용례는 ‘류지희 외(2017), 한국어 생략어 복원 가이드라인, 『한글 및 한국어 정보처리 학술대회 논문집』 29, 213-219.’에서 인용하였다.

1) 지배소 후보

의존 구문에서는 서술어가 지배소가 되고, 주어, 목적어, 부사어 등이 서술어에 의존하는 피지배소가 되는 것이 일반적이다. 따라서 구문 분석에서 VP로 분석된 문장 성분이 지배소의 후보가 된다. VP로 표현되는 동사 표현 어구는 개체와 개체 간의 사건, 행동 및 상태와 같은 정보를 서술하는 형태이다. 긍정지시사구인 VNP도 지배소의 후보가 된다. 단, VP_MOD로 분석되는 용언의 관형형 가운데 ‘관한, 대한, 의한, 향한, 인한, 통한, 따른, 아닌, 같은’ 등과 같이 서술어에 해당하지 않는 형태는 지배소로 분류하지 않는다.

케냐 공화국 또는 케냐는 동아프리카의 공화국이다*[VNP]. 인도양에 면해*[VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 국경을 맞닿고*[VP] 있다.

<지배소의 예시>

본 사업에서는 한국전자통신연구원(ETRI)의 생략어 복원 태깅 가이드라인 v3.3에 따라 구문 분석 결과 구문 태그가 VP인 어절을 복원 대상 서술어로 삼으며, ‘와 같은/잇달아/비슷한/따라/따르면/이어/대해/통해/걸쳐/관련한/관련된/비해/어떠한/와 더불어/에 뒤이어/불구하고/위한/지난 00일/최근 들어/인 듯한’과 같은 표현은 분석에서 제외한다. 또한 구문 분석은 한국정보통신기술협회(TTA)의 ‘의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계 설정 방법’을 따른다.

2) 생략어 후보

한국전자통신연구원(ETRI) 가이드라인의 생략어 복원 대상은 ‘지배소 서술어의 생략된 주어’와 ‘지배소 서술어의 생략된 목적어’이다. 본 사업에서는 서술어의 생략된 주어만을 생략어 복원 대상으로 한다.

케냐 공화국 또는 케냐는 동아프리카의 공화국이다. [?는]¹ 인도양에 면해*[VP] 있으며 북동쪽으로 소말리아, 북쪽으로 에티오피아와 남수단, 서쪽으로 우간다, 남쪽으로 탄자니아와 [?는]¹ 국경을 맞닿고*[VP] 있다.

<생략어의 예시>

3) 선행어 후보

선행어 후보가 같은 문장 내에 있다면 그것을 우선적으로 선행어로 여긴다. 문장 내에 선행어 후보가 없는 경우, 생략어와 가까운 위치의 것을 선행어로 결정한다. 필자가 생각하고 있는 주된 포커스가 명시적으로 또는 묵시적으로 존재하는 경우, 이러한 포커스가 선행어가 될 수 있으며 상호 참조 해결의 대상이 되는 개체(entity) 또한 선행어로 나타날 수 있다.

비텐베르크 대학교의 요한 스타우피츠 교수[†]는 루터가 성서에 대해 진지하게 공부하면 평안을 찾을 것이라고 생각하였다. 그래서 [요한 스타우피츠 교수[†]는][‡] 그를 성서학 교수사제로 임명하였는데^{*} [VP], 스타우피츠 교수의 결정은 루터가 신앙적인 고민을 해결하는데 도움이 되었다.

<선행어 후보>

한편, 용언 관형어(VP_MOD)의 수식 대상은 서술어의 주어가 될 수 있으므로 이러한 대상을 선행어로 보는 것이 가능하다면 선행어로 태깅한다. 한국전자통신연구원(ETRI)에서 제시한 용언 관형어의 주어는 후행하는 명사구로, 해당 작업에서는 구문 분석에 의존한 복원 경향이 강하다. 그러나 본 사업에서는 구문 분석 결과뿐 아니라 의미적 연관성도 고려하여 선행어를 복원하였다. 후행하는 명사구가 용언 관형어와 의미적 관련성이 없는 경우, 문맥상 관련성이 있는 명사²⁾를 찾아 선행어로 지정하였다.

선행어 후보를 문서 내에서 찾을 수 없다면 문서 내에 존재하지 않는 알 수 없는 대상이 선행어가 된다. 이때 해당 생략어는 비지시적(non-anaphoric) 대용어³⁾이며, ‘누군가(somebody)’ 또는 ‘무언가(something)’라는 대명사가 선행어가 된다.

2.1.4. 생략어 복원 태깅 사례

생략어 태깅의 전형적인 순서는 첫째, 문장 내에서 지배소 후보들을 찾은 뒤, 둘째, 각 지배소 후보에 따른 생략어 후보가 존재하는지 탐색하는 것이다. 그리고 셋째, 선행어 후보들을 검토한 뒤, 해당 선행어를 복원하였을 때 필자의 의도가 명확해지는지 확인한다.⁴⁾

2) 복원 대상 선행어는 일반적으로 구문 분석상 체언에 해당하는 구문 태그 명사구(NP)에 해당된다. 그러나 ‘명사+이다’형의 긍정 지정사구(VNP)의 명사도 선행어 대상에 포함되기 때문에 명사구 대신 명사를 선행어로 지정한다고 기술하였다.

3) 비지시적 대명사란, 뭔가를 지시하는 것이 주목적이 아닌 대명사를 일컫는다. 이러한 비지시적 대명사로 의문대명사, 비한정/부정 대명사를 분류하여 제시하였는데, 본 지침에서는 의문사가 아무런 형태 변화 없이 그대로 쓰인 ‘누군가’ 또는 ‘무언가’에 해당한다.(박진호(2007), 「유형론적 관점에서 본 한국어 대명사 체계의 특징」, 『국어학』 50, 국어학회, 115-147쪽.)

문재인 대통령은 21일 '경제 사령탑'인 경제부총리 겸 기획재정부 장관 후보자에 김동연(60) 아주대 총장, 외교부 장관 후보자에 여성인 강경화(62) 유엔 사무총장 정책특보를 각각 내정했다. ... 김동연 부총리 후보자[†]는 충북 음성 출신으로 '고졸신화의 인간승리' 드라마로 불린다. [김동연 부총리 후보자[†]는] 덕수상고 졸업 뒤 은행에 취직해^{*}[VP] 직장생활을 하며^{*}[VP] 행정고시와 입법고시에 합격한^{*}[VP_MOD] 입지전적의 인물로 평가 받는다^{*}[VP].

```
"ZA" : [
  {
    "id" : 0,
    "type" : "s", // 주어
    "head_wid" : 4, // 5 번째 어절(취직해)
    "ant_text": "후보자",
    "ant_sid" : 6, // 7 번째 문장
    "ant_wid" : 2, // 3 번째 어절(후보자는)
    "ant_is_title" : 0 // 표제어 아님
  }, ...
]
```

<생략어 복원 태깅 사례>

주격 무형 대용어 복원 작업의 주된 사항은 술어가 요구하는 생략된 논항을 찾는 것이다. 따라서 구문 분석과 의미역 인식 등 타 층위의 복합적 지식이 요구되므로 지침에 대한 분석이 선행되어야 한다.

2.2. 주격 무형 대용어 복원 지침

주관 기관 국립국어원과 수행 기관 (주)이르테크 공동수급체는 2.1.에서 기술한 한국전자통신연구원(ETRI)의 지침을 바탕으로 다음과 같은 분석 지침을 수립하였다.

2.2.1. 기본 방향

- 1) 자연 언어 처리를 위한 일관성 유지와 효율성 제고에 초점을 두되, 일반 언어학적 관점에서도 크게 벗어나지 않도록 한다.
- 2) 서술어의 필수격 중 주격에 해당하는 명사구가 생략된 경우에만 복원한다.
- 3) 한국정보통신기술협회(TTA)에서 제정한 구문 분석에 기초하되, 주어와 서술어 간의 의미적 연관성을 최대한 고려하여 선행어를 탐색한다.

4) 생략어 복원 태깅 용례는 '류지희 외(2017), 한국어 생략어 복원 가이드라인, 『한글 및 한국어 정보처리 학술대회 논문집』 29, 217쪽'에서 인용하였다.

2.2.2. 기본 지침

1) 생략 술어 탐지

가. 주어가 생략된 술어 탐지

구문 분석 결과, 지배하는 주어가 없는 VP를 대상으로 생략된 주어를 문서 내에서 복원한다.

나. 주어 복원 대상이 아닌 술어 배제

(1) 보조 용언

『표준국어대사전』에 등재된 표제어 가운데 품사가 보조 용언인 술어에 대해서는 주어를 복원하지 않는다.

예) 가다¹, 보다¹, 있다¹, 주다¹, 하다¹, ...

가다¹

발음 [가다] 

활용 가[가] , 가니[가니] 

【II】「보조 동사」

((주로 동사 뒤에서 ‘-어 가다’ 구성으로 쓰여))

말하는 이, 또는 말하는 이가 경하는 어떤 기준점에서 멀어지면서 앞말이 뜻하는 행동이나 상태가 계속 진행됨을 나타내는 말.

- * 책을 다 읽어 **간다**.
- * 밥이 식어 **가는데** 물 좀 올려라.
- * 하는 일은 **갈수록** **가다**?

[더 보기 >](#)

<표준국어대사전의 보조 용언 예시>

(2) 의사 보조 용언 구성

한국정보통신기술협회(TTA)의 지침을 참조하여, 서술어 다음에 서법을 나타내는 언어 단위들이 오는 경우, 의사 보조 용언 구성으로 간주한다. 이때 주어는 주 서술어와 의존 관계를 연결하고, 뒤따르는 언어 단위들에 대해서는 주어를 복원하지 않는다.

예) -르 수/리(가) 있다/없다

-ㄴ/ㄹ {것!터!뿐!따름!모양!지경!참!중!노릇!예정!길}이다

-르 {만!뻔!듯}하다, -는 말이다, -ㄴ/ㄹ 듯(도) 하다

-ㄴ/ㄹ 것 같다, -르 것을(걸) 그랬다, -어서는 안 된다, -고 해서, -든지 하다

(3) 서술어에 해당하지 않는 동사구

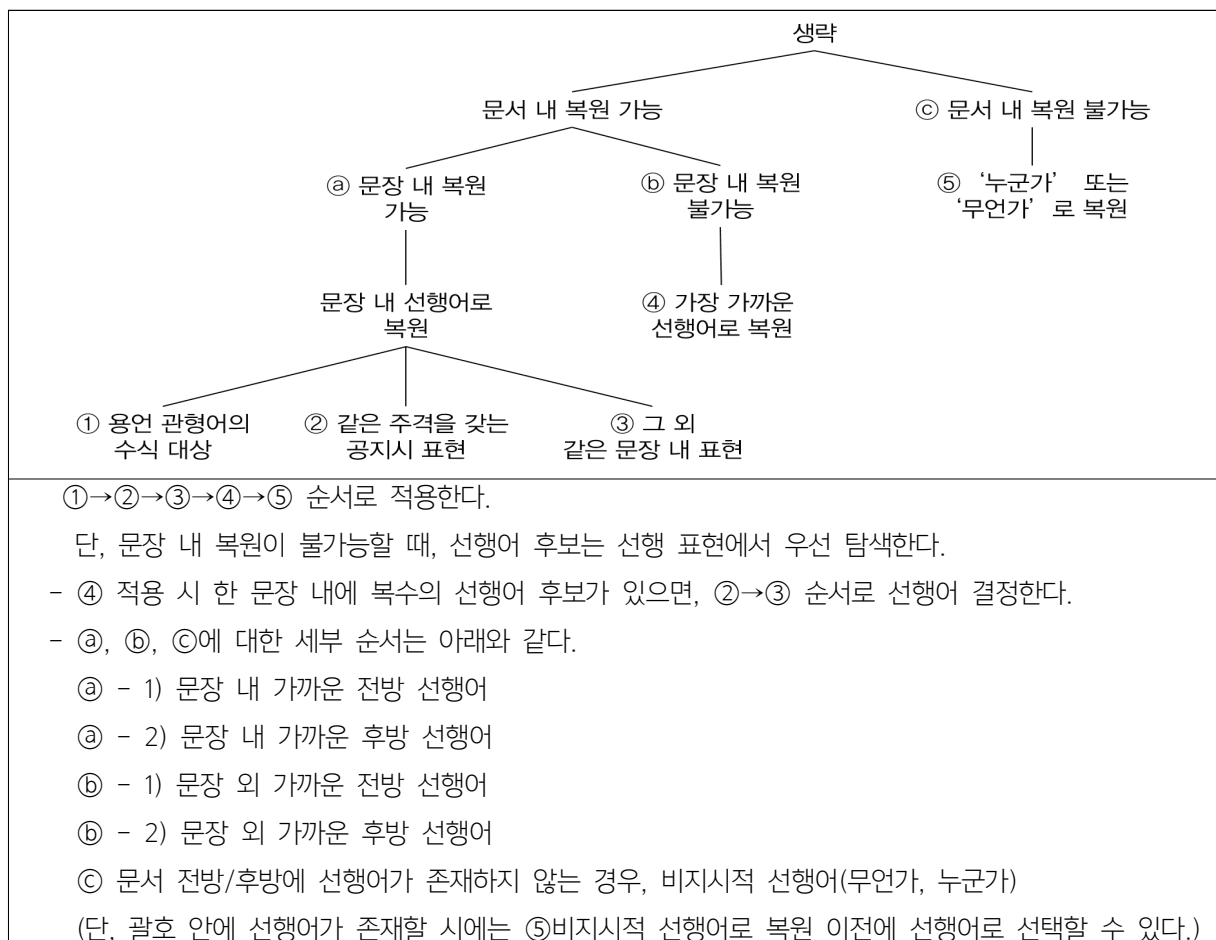
구문 분석 결과 VP로 분석되었으나 서술어에 해당하지 않는 동사구는 주어 복원 대상 술어에서 제외한다. ‘관한, 대한, 의한, 향한, 인한, 통한, 따른, 아닌, 같은’ 등의 VP_MOD와 ‘관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고, 그러면서’ 등의 VP_AJT가 이에 속한다.

예) 전기 요금이 오른 데에 이어 수도 요금이 올랐다.

→ *전기 요금이 오른 데에 {잇는다}이었다.

2) 선행어 결정

본 분석의 선행어는 전방조응사뿐 아니라 후방조응사도 포괄한다. 생략어 복원의 분석 대상은 문서(document)로, 앞뒤 문맥을 확인하여 생략된 주어를 복원한다.



〈생략어 복원 절차〉

가. 문서 내 복원 가능

(1) 작년에는 완패했던 서울이 매서운 공격력으로 선전했다.

	작년에는	완패했던	서울이	매서운	공격력	으로	선전했다.
술어		VP_MOD		VP_MOD			VP
주어							서울이
생략		서울		공격력			

※ '주어'는 술어가 구문 분석상 지배하는 주어를 뜻함.

※ '생략'은 술어가 구문 분석상 지배하는 주어가 생략되어 있고 이것이 기입된 선행어로 복원됨을 뜻함.

(2) 그가 교수라고 불리기를 원했다.

	그가	교수라고	불리기를	원했다.
술어			VP_OBJ	VP
주어		그가		
생략			그	그

(3) 철수의 경우에 취업하자마자 휴학계를 제출했다.

	철수의	경우에	취업하자마자	휴학계를	제출했다.
술어			VP		VP
주어					
생략			철수		철수

나. 문서 내 복원 불가능

(4) 이탈리아어로 '안녕'을 'ciao'라고 한다.(생략된 주어: 일반적인 이탈리아 사람들)

	이탈리아어로	'안녕'을	'ciao'라고	한다.
술어				VP
주어				
생략				누군가

(5) 김영희 의원이 같은 민주당인 이철수 의원을 불렀다.(생략된 주어: 정당, 소속, 소속 정당 등)

	김영희	의원이	같은	민주당인	이철수	의원을	불렀다.
술어			VP_MOD				VP
주어							의원이
생략			무언가				

2.2.3. 세부 지침

1) 다어절 선행어

: 중심어(해당 구의 핵)를 선행어로 태깅한다.

(1) 돌아온 테니스 여제가 결승에 진출했다.

	돌아온	테니스	여제가	결승에	진출했다.
술어	VP_MOD				VP
주어					여제가
생략	여제				

2) 의존 명사

: ‘①용언 관형형의 수식 대상’으로 간주하여 의존 명사를 복원한다.

(2) 버스 뒤에 온 것은 택시였다.

(3) 버스 뒤에 온 차량은 택시였다.

(4) 가위바위보에 이긴 쪽이 부전승으로 결승전에 올라갔다.

(5) 가위바위보에 이긴 청군이 부전승으로 결승전에 올라갔다.

3) 절 선행어(clausal antecedent)

: 다어절 선행어로 간주하여 중심어(즉, 술어)를 선행어로 태깅한다.

(6) 대질조사에서 김민아가 이철수에 말리면 결정적 패착이 될 수 있다.

	[대질조사에서	김민아가	이철수에	말리면]	결정적	패착이	될	수	있다.
술어							VP_MOD		
주어									
생략							말리면		

단, 명사 선행어 후보가 없을 때에 한하여 제한적으로 절을 선행어로 삼는다.

(7) 줄거리는 비슷비슷하지만 나의 경험을 담으면 나만의 독서감상문이 된다.

(8) (독서감상문은) 줄거리는 비슷비슷하지만 (그것에) 나의 경험을 담으면 (그것이) 나만의 독서감상문이 된다.

(7)의 문장을 절을 선행어로 취하는 용례로 볼 수도 있으나, (8)과 같이 문서 내에서 ‘독서감상문’을 복원할 수 있으므로 이 경우에는 ‘독서감상문’을 생략어로 복원한다. (6)과 (7)에서 ‘되다’의 생략된 주어를 ‘그것’이라 할 때, ‘그것’이 명제를 지시하는지 개체를 지시하는지 구분하여, 명제로 해석되는 경우에 한하여 절을 선행어로 취하는 것으로 분석한다.

4) ‘-게/로’

: 논항이나 부가어(부사 제외)를 취하지 않으면 구로 간주하여 복원 대상에서 제외한다.

(9) 민아가 철수를 나쁘게 말했다.

민아가 → NP_SBJ 말했다.

철수를 → NP_OBJ 말했다.

나쁘게 → VP_AJT 말했다.

『표준국어대사전』에서 제시하는 ‘말하다’의 구문 정보는 [...을 -게]이다. ‘나쁘게’는 ‘말했다’를 수식하는 부사구로 취급하여 주격 무형 복원 대상에서 제외한다.

5) 보조 용언 여부 주의

가. [-게 하다, -게 되다] 부류

(10) 민아가 철수를 집에 가게 했다.

(11) 김민아 선생님께서 오늘부터 여러분에게 한국어를 가르치게 되셨어요.

(10)의 ‘-게 하다’의 ‘하다’는 보조 용언이므로 주격 무형 복원 대상이 아니지만, (11)의 ‘-게 되다’의 ‘되다’는 본용언이므로 주어 복원 대상이 된다. (11)의 구문 분석 결과, ‘되셨어요’의 생략된 주어는 ‘선생님’으로 복원이 가능하다.

나. 보다, 보이다

(12) 옷을 입어 봤다.

(13) 공원에 가 봤다.

(12), (13)의 ‘보다’는 문장에서 ‘-어 보다’의 꼴로 쓰여 ‘어떤 행동을 시험 삼아 함’, ‘어떤 일을 경험함’ 등을 나타내는 보조 동사이므로 주격 무형 복원 대상이 아니다.

(14) 그는 단순한 후배로만 보였다.

그는 → NP_SBJ 보였다.

후배로만 → NP_AJT 보였다.

(15) 철수는 30대처럼 보인다.

철수는 → NP_SBJ 보인다.

30대처럼 → NP_AJT 보인다.

(16) 오늘따라 과장님께서 불쌍해 보여요.

오늘따라 → NP_AJT 불쌍해

과장님께서 → NP_SBJ 불쌍해

불쌍해 → VP 보여요.

(17) 오늘따라 과장님께서 즐거워 보여요.

오늘따라 → NP_AJT 즐거워

과장님께서 → NP_SBJ 즐거워

즐거워 → VP 보여요.

‘보이다’는 보조 용언 용법이 없으므로 일관적으로 본용언으로 분석한다. 따라서 (16), (17)의 문장에서 ‘보여요’는 모두 생략어 복원 대상 술어가 된다. 이들 문장에서는 ‘과장님’이 각각 주어로 복원된다.

6) ‘에서’ 주격 조사, 부사격 조사 여부 유의

: ‘에서’는 주격 조사로서의 쓰임과 부사격 조사로서의 쓰임을 두루 갖는다. 아래 (18)의 ‘학교에서’는 주어로 사용되었으며, (19)의 ‘학교에서’는 주어와 부사어로 모두 분석할 수 있어 부사어로 취급하기로 하였다. 따라서 (19)의 ‘열었다’는 주격 복원 대상 술어이며, 문맥에 따라 ‘학교’ 또는 ‘누군가’로 주어를 복원할 수 있다.

(18) 이번 대회는 우리 학교에서 우승을 차지했다.

학교에서 → NP_SBJ 차지했다.

(19) 우리 학교에서 이번 대회를 열었다.

학교에서 → NP_AJT 열었다.

7) 인칭

: 생략어를 복원할 때에는 선행어의 인칭을 그대로 유지한다. 따라서 (20)의 ‘외쳤다’와 의존 관계를 이루는 주어로 ‘철수’를 복원한다. ‘철수’와 ‘나’가 동일 지시라는 것은 상호 참조 분석 결과로 확인할 수 있다.

(20) 철수는 밥을 먹으면서 울컥 짜증이 치밀었다. 그래서 “내가 미쳐!”라고 외쳤다.

2.3. 문어와 구어 태깅의 실례

2.2.의 지침 내용을 토대로 문어와 구어 말뭉치에서 주어 복원의 실례를 제시하면 다음과 같다.⁵⁾ 주격 무형 대용어 복원 세부 분석 지침은 사업 추진 일정에 따라 문어를 중심으로 수립되었으며, 이후 구어 말뭉치의 특성을 고려하여 주의 사항이 보강되었다.

2.3.1. 문어

주관 기관에서는 2019년도 국어 빅데이터 구축 사업 분석 말뭉치 구축팀을 대상으로 과제 초반 전체 구축 분량의 1%에 대해 시범 구축 및 검증을 실시하였다. 시범 구축 결과 다음과 관련하여 분석 오류가 빈번히 발생하는 것으로 나타났다. (1) 인용절 내포문의 주어 판단, (2) 내포문의 인칭, (3) 관형형의 후행 NP가 복수로 나타나는 경우, (4) 서술어에 해당하지 않는 동사구를 포함한 복문, (5) 보조 용언의 주어 복원 여부, (6)

5) 태깅 실례를 제시할 때, 1) 말뭉치상에서 나타난 띄어쓰기 오류 등의 정서법 오류는 수정하지 않고, 해당 용례를 그대로 제시하였다. 2) 용례로 제시된 문장에서 복원된 주어를 모두 표기하면 혼란을 초래할 수 있으므로 각각의 설명에 해당하는 주어만 복원시켜 제시하였다. 3) 국립국어원 주격 무형 대용어 복원 결과물에는 복원되는 주어에 주격 조사 ‘가’를 붙이지 않지만, 본 문서에서는 가독성을 위해 ‘가’를 붙여서 제시하였다. 4) 문장 내 선행어는 굵은 글씨로 표시하고, 복원 대상 술어에는 밑줄을 그어 구분하였다. 복원 선행어는 < > 안에 제시하였다.

의사 보조 용언 처리 등이다. 각각에 해당되는 사례와 주어 복원 방법에 대해 제시하면 아래와 같다.

1) 인용절에서의 주어 복원

인용절에서는 구문 분석 결과에 따라 주어 복원 대상 술어가 결정된다. 구문 분석에서는 문장 부호에 의하여 구분된 단위를 준수하여, 간접 인용과 직접 인용에 대하여 다른 지침을 적용한다.⁶⁾ 특히, 모문과 내포문의 주어가 같은 경우, 간접 인용에서는 모문의 주어가 내포문의 서술어를 지배소로 갖지만, 직접 인용에서는 모문의 주어가 모문의 서술어를 지배소로 갖는다.

(1) 철수는 집에 가겠다고 <철수가> 말했다.

(2) 철수는 “집에 가겠다”고 말했다.

(1)에서 ‘철수는 → NP_SBJ 가겠다고’와 같은 의존 관계를 보이므로 ‘말했다’의 주어로 ‘철수’를 복원한다. 하지만 (2)에서는 ‘철수는 → NP_SBJ 말했다’와 같이 구문 분석되므로 ‘말했다’가 복원 대상 술어에서 제외된다.

또한, 직접 인용에서 ‘-다’며’로 인용된 절은 VP의 연결 구성으로 보고, 모문의 주어가 ‘VP-며’에 의존하는 것으로 분석한다.

(3) 철수는 “밥을 먹고 있다”며 “곧 집에 갈 예정”이라고 <철수가> 말했다.

(3)은 ‘철수는 “밥을 먹고 있다”라고 말하며 “곧 집에 갈 예정”이라고 말했다’와 동일한 구성으로 분석되므로, ‘철수는 → NP_SBJ 있다’며’와 같은 의존 관계를 갖는다. 따라서 (3)의 ‘말했다’는 주어 복원 대상 술어이다.

[실제 태깅 예시]

6) 간접 인용과 직접 인용의 구별은 문장 부호(따옴표 등)를 기준으로 삼는다. 예를 들어, 아래 문장은 전형적인 직접 인용은 아니며 신문 기사 특유의 문체이다. 그러나 본 사업에서는 큰따옴표를 기준으로 직접 인용으로 간주하였다. 또한 직접 인용된 내포문이 원시 말뭉치에서 여러 개 문장 단위로 분리되어 있을 경우, 쌍을 이루는 문장 부호를 기준으로 직접 인용 여부를 판단하였다.

예) 고려 불화 권위자인 정우택 동국대박물관장은 “지난 3월 동국대 개교 111주년 기념사업회 후원으로 일본 지역 한국 불교 미술품을 조사하는 과정에서 고려 불화 ‘수월관음도’를 발견했다”고 18일 밝혔다.

◎ 간접 인용절

모문과 내포문의 주어 및 서술어가 각각 다른 경우, 모문의 주어는 모문의 서술어에 의존하고, 내포문의 주어는 내포문의 서술어에 의존한다. 따라서 이들 의존 관계를 기준으로 주어를 복원한다.

(4) 나는 현대차가 잘했다고 생각한다.

(5) 그만큼 상황이 급박하다고 <본부장이> 판단했다.

(4)에서 모문의 서술어 ‘생각한다’는 모문의 주어인 ‘나’를 의존소로 가지고, 내포문의 서술어 ‘잘했다’는 내포문의 주어인 ‘현대차’를 의존소로 가지므로 (4)에서는 복원 대상 술어가 존재하지 않는다. 반면, (5)에서는 내포문의 주어 ‘상황’과 서술어 ‘급박하다’가 각각 의존 관계를 맺는 것에 반해, 모문의 서술어인 ‘판단했다’와 연결되는 모문의 주어가 생략되어 있다. 따라서 ‘판단했다’의 주어로 문서 내 선행어인 ‘본부장’을 복원하였다.

한편, 간접 인용에서 모문과 내포문의 주어가 동일하고, 서술어가 다른 경우에 선행하는 주어는 내포문의 서술어를 지배소로 갖는다. 따라서 모문의 서술어에 대한 주어를 복원해야 한다.

(6) 신씨는 항소하겠다고 <신씨가> 했다.

(7) 정작 그는 그런 명성이 달갑지 않다고 <그가> 한다.

예문 (6)의 주어 ‘신씨’는 내포문 서술어인 ‘항소하겠다고’에, (7)의 주어 ‘그’는 ‘달갑지’에 의존한다. 따라서 (6)에서 ‘했다’의 주어로 ‘신씨’가, (7)에서 ‘한다’의 주어로 ‘그’가 각각 복원되었다.

◎ 직접 인용절

모문과 내포문의 주어가 동일하고, 서술어가 다른 경우, 간접 인용에서는 모문의 주어가 내포문의 서술어에 의존했던 것과 달리, 직접 인용에서는 모문의 주어가 모문의 서술어에 의존한다. 즉, 따옴표 밖에 있는 주어는 따옴표 안의 주어와 동일하더라도 모문의 서술어에 연결된다.

(8) 플라티니 회장은 “메시가 내 기록을 깎 것으로 <나는> 확신한다”고 밝혔다.

(9) 그는 “너무나 큰 책임감을 <누군가> 느낀다”고 했다.

(8)에서 인용 부호 바깥에 있는 주어 ‘회장’은 ‘밝혔다’에 의존하므로, 내포문의 서술어 ‘확신한다’의 주어로 ‘나’를 복원하였다. 그리고 (9)에서 ‘그’는 ‘했다’에 의존하므로 ‘느낀다’의 주어로 ‘누군가’를 복원하였다.⁷⁾

모문과 내포문의 주어 및 서술어가 각각 다른 경우에는 간접 인용과 마찬가지로 모문의 주어는 모문의 서술어에 의존하고, 내포문의 주어는 내포문의 서술어에 의존한다.

(10) 한 승객이 “경찰이 오고 있다”고 외쳤다는 얘기도 있다.

위 문장에서 승객은 인용 부호 바깥에 있으므로 “외쳤다”에 의존한다. 따라서 ‘외쳤다’는 주어 복원 대상 술어가 아니며, 내포문의 ‘오고’ 역시 ‘경찰’을 주어로 취하므로 주어 복원 대상 술어가 아니다.

(11) 장 대표는 “그보다는 <판권료가> 낮다”고 말했다.

(11)에서 ‘대표’는 ‘말했다’에 의존하므로, ‘낮다’의 주어로 문서 내 선행어인 ‘판권료’를 복원하였다.

(12) 그린피스도 성명을 내 “원주민들의 기념비적 승리이자, 수자원을 보호하기 위한 환경운동가들의 승리”라고 <그린피스가> 평가했다.

(13) 청와대 관계자는 7일 오전 기자들과 만나 “지난해 5월 10일 외교안보수석 주재 한국형 전투기 개발 사업 전문가 오찬 간담회가 있었고, 이 회의는 다양한 의견을 청취하기 위한 성격의 회의였다”고 <관계자가> 해명했다.

(12)와 (13)은 모문에 두 개 이상의 술어가 존재하는 문장으로, 이 경우 주어는 선행하는 술어에 의존한다. (12)에서 주어인 ‘그린피스’는 부사절 술어인 ‘내’에 의존하며, ‘내’는 ‘평가했다’에 의존한다. 따라서 ‘평가했다’에 의존하는 주어

7) (9)의 경우, 문서 내에서 선행어 ‘나’를 찾을 수 없어, ‘누군가’로 복원하였다. 인용절의 인칭 제약의 대해서는 ‘2)인용절 모문과 내포문의 인칭’에서 상세히 서술하고 있다.

가 없으므로 ‘그린피스’를 복원하였다. 마찬가지로, (13)의 모문의 주어 ‘관계자’는 부사절 술어 ‘만나’에 의존하므로 동일 명사구를 주어로 취하는 모문의 서술어 ‘해명했다’는 복원 대상 술어가 된다.

인용 부호 내 문장이 분리된 경우에는 원시 말뭉치에서 주어진 문장 단위에 따라 주어 복원 분석 결과가 달라진다.

(14)

<s>양동근은 “홀쭉해졌다.</s>
<s>비결이 뭐냐“고 <양동근이> 물었다.</s>

(15)

<s>엘지전자 관계자 는 “스마트폰 사업은 버릴수 있는 사업이 아니다.</s>
<s>계속 유지해야 하는 사업이다“고 <관계자가> 말했다.</s>

(14), (15)는 인용 부호 내 문장이 분리되어 모문의 주어와 서술어가 각각 다른 문장에 위치하게 된 경우이다. (14)에서 후행 문장의 술어인 ‘물었다’에 대한 가시적인 주어가 없으므로 선행 문장의 ‘양동근’을 선행어로 삼아 복원하였다. 만약 (14) 전체가 하나의 문장으로 묶여 있다면 주어 ‘양동근은’이 ‘물었다’에 의존하므로 ‘물었다’는 복원 대상 술어가 아니다. (15) 역시 후행 문장의 ‘말했다’에 대한 가시적인 주어가 없으므로 ‘관계자’를 선행어로 복원하였다. (15) 전체가 하나의 문장으로 묶여 있다면 ‘관계자는’이 ‘말했다’에 의존하므로 ‘말했다’는 복원 대상 술어에서 제외된다.

한편, 인용격조사 ‘(라)고’가 아닌, 어미 ‘-며’가 결합하여 ‘-다”며’로 인용된 절은 VP의 연결 구성으로 보고 선행하는 주어가 ‘-며’에 의존하는 것으로 분석한다.

(16) 지난 18일 서울 충무로 한국·중앙아시아 친선협회 사무실에서 만난 살로히딘 대사는 “국토의 93%가 산지인 우리나라는 잠재력이 무궁무진하다”며 ‘국가 세일즈’에 <대사가> 나섰다.

(16)의 ‘무궁무진하다’며’는 ‘-다고 하면서’의 축약형이 사용된 부사절 술어이다. 이때 주어 ‘대사’는 생략된 ‘하면서’와 관계하므로 서술어 ‘무궁무진하다’며’에 의존하는 것으로 분석한다. 따라서 후행하는 서술어 ‘나섰다’는 구문 분석상 주어를 취할 수 없으므로 선행어 ‘대사’를 주어로 복원하였다.

(17) 박지원 의원은 “김대중 전 대통령께서 성실성과 노력, 실력을 높이 평가했던 분”이라며 이회호 여사의 조전(弔電)을 <의원이> 전달했다.

(17)에서 역시 주어 ‘의원’은 서술어 ‘분’이라며’에 의존하므로 ‘전달했다’의 주어로 ‘의원’을 복원하였다.

(18) 영국 가디언은 “전형적인 미국 4대 스포츠 대신 새로운 것을 찾던 젊은이들에게 MLS가 먹혀들어가기 시작했다”며 “특히 축구가 민족과 인종을 넘어 사랑받는 세계 스포츠라는 점이 젊은 층의 감성을 자극했다”고 <가디언이> 분석했다.

모문의 주어 ‘가디언’은 서술어 ‘시작했다’며’에 의존하기 때문에 모문의 서술어인 ‘분석했다’를 복원 대상 술어로 보고, ‘가디언’을 주어로 복원하였다.

2) 인용절 모문과 내포문의 인칭

직접 인용 구문에서 주어를 복원할 경우, 주어의 인칭에 주의해야 한다. 모문과 내포문의 주어가 동일하다 하더라도, 문장에 따라 문장 부호 밖의 주어와 인용절 내부의 주어의 인칭이 다를 수 있기 때문이다. 가령, 화자가 본인을 주어로 이야기하는 경우, 문장 부호 밖의 주어가 2인칭 혹은 3인칭이라 하더라도 인용절은 화자의 시점에서 말하는 것이므로 주어 복원 시 1인칭을 나타내는 ‘나’, ‘저’ 등의 1인칭 대명사로 복원한다. 만일 문서 내 선행어로 ‘나’를 선택할 수 없는 경우에는 ‘누군가’로 복원한다.

(1)

1	<s>연주 전 만난 슈텐츠는 “생각의 자유로움이 담겨 있고 그러면서 다양한 음악 색채와 명암을 갖고 있다“고 슈만 교향곡을 소개했다.</s>
2	<s>“오케스트라의 각 파트를 제 색깔 내도록 훈련시킨 다음 그걸 조화롭게 아울러서 나아가게 하고 싶어요.</s>
3	<s>2015년 12월 서울시향과 말러 교향곡 1번을 해보니 목관은 말러의 특성을 잘 이해한 소리를 냈고, 브라스 섹션은 에너지로 넘쳤죠.</s>
4	<s>기본이 탄탄하니 그걸 유연하게 <u>살려내는 건 제</u> 몫입니다.”</s>
5	<s>리허설 땀 지휘봉이 세 동강 날 정도로 열정을 쏟아부었다.</s>

5번 ‘쏟아부었다’의 주어로 복원될 수 있는 선행어 후보로 1번 ‘슈텐츠’와 4번 ‘제’를 생각할 수 있다. ‘슈텐츠’와 ‘제’가 가리키는 개체는 동일하기 때문이다. 그러나 ‘슈텐츠’는 3인칭 명사구이고 ‘제’는 1인칭 명사구이다. 각 명사구로 5번 문장을 복원하면 “리허설 땀 지휘봉이 세 동강 날 정도로 <슈텐츠가/*제가> 열정을 쏟아부었다.”이므로, 따옴표 바깥 문장인 5의 ‘쏟아부었다’의 주어로는 ‘슈텐츠’가 복원된다. 반면, 인용절 내부 문장인 4에서는 서술어 ‘살려내는’의 주어로 3인칭 ‘슈텐츠’가 아닌 1인칭 ‘저’가 복원된다.⁸⁾ 같은 개체를 지칭하더라도 인칭에 따라 선행어가 달라지는 것이다.

[실제 태깅 예시]

(2)

최효진은 “내 생애 최고의 날이었다.
<나는> 잊을 수 없다“고 <최효진이> 했다.

(2)의 두 번째 문장에서 모문과 내포문의 주어는 동일 개체이지만, 인용절 내부 서술어인 ‘잊을’의 주어로는 1인칭 ‘나’가, 인용절 바깥 서술어인 ‘했다’의 주어로는 3인칭 ‘최효진’이 각각 복원되었다.

8) 본 사업 결과물은 선행어에 조사를 포함하지 않으므로 4의 ‘제’를 ‘저+의’로 형태 분석하여, ‘저’를 주어로 복원한다.

(3)

마르키는 “공백기 동안 뉴욕에서 아들의 유소년 야구팀을 <나는> <u>지도했다</u> .
그러면서 내 어깨 관리도 꾸준히 <나는> <u>했다</u> ”고 <마르키가> <u>설명했다</u> .

(3)의 모문과 내포문의 주어는 ‘마르키’로 동일하지만, 인용 부호 바깥 서술어인 ‘설명했다’에 대한 주어로는 3인칭 ‘마르키’가, 인용절 내부 서술어인 ‘지도했다’, ‘했다’의 주어로는 1인칭 ‘나’가 각각 복원되었다.

내포문의 주어가 생략되었지만 문서 내에 1인칭 선행어가 존재하지 않는 경우에는 주어를 ‘누군가’로 복원한다.

(4) 박 씨는 무전기로 다른 승무원들에게 “어떻게 <누군가> 대응해야 하느냐”라고 물었지만 답신이 오지 않았다.

(4)에서 내포문 서술어인 ‘대응해야’는 주어 복원 대상 술어이지만, 문서 내에 1인칭 주어를 가리키는 단어가 등장하지 않아 ‘누군가’로 복원하였다. 모문의 서술어인 ‘물었지만’은 ‘박 씨’를 지배소로 가지므로 복원 대상 술어가 아니다.

(5) 이 사무총장도 “오래전부터 교육부 공무원들이 많이 이동하던 자리로, 공직자 윤리위원회에 <누군가> 알아봤더니 ‘기관’이 아닌 ‘협의회’이기 때문에 대상자가 아니라 문제없다는 판단을 받았다”고 설명했다.

‘알아봤더니’의 주어는 ‘이 사무총장’이지만, ‘나’, ‘저’ 등의 1인칭 명사구가 문서에 나타나지 않아 ‘누군가’로 복원하였다.

(6) 그는 “마치 피난 행렬을 <누군가> 보는 것 같았다”고 말했다.

내포문의 서술어 ‘보는’은 화자 자신을 주어로 취하지만, 1인칭 명사구가 문서에 등장하지 않으므로 ‘누군가’로 주어를 복원하였다.

3) 명사구 수식

한국정보통신기술협회(TTA) 구문 분석은 지배소 후위 원칙에 따른다. 명사구가 여러 어절로 구성된 경우에도 기본적으로 명사구 내에서 가장 후행하는 어절을 지배소로 삼는다. 국립국어원 구문 분석 말뭉치 구축 사업 역시 기본적으로는 명사구 내에서 가장 후행하는 어절을 지배소로 삼지만, 해석이 중의적일 경우에는 인접한 어절과의 관계를 우선시한다.

- (1) 서울에 살고 있는 김철수 씨 > 서울에 <씨> 살고 있는 김철수 씨
서울에 살고 있는 김철수 시인 > 서울에 <시인> 살고 있는 김철수 시인
서울에 살고 있는 철수와 영희 > 서울에 <영희> 살고 있는 철수와 영희
(2) 서울에 살고 있는 시인 김철수 > 서울에 <시인> 살고 있는 시인 김철수

(1)의 ‘서울에서 살고 있는’이 수식하는 후행 명사구 가운데 중심어는 ‘씨, 시인, 영희’이므로, 이에 따라 ‘살고’의 선행어도 ‘씨, 시인, 영희’가 된다. 반면, (2)의 구조는 [[서울에 살고 있는 시인] 김철수]와 [[서울에 살고 있는 [시인 김철수]]로 모두 해석할 수 있다. 인접성 우선에 따라 (2)는 [[서울에 살고 있는 시인] 김철수]로 구문 분석하고, ‘살고’의 생략된 주어는 ‘시인’으로 분석한다. 이러한 수식구 해석의 중의성은 ‘시인 김철수’와 같은 동격 명사구를 수식할 때 주로 발생한다.⁹⁾

하지만 모든 (유사) 동격 구문에 대해 선행 어절을 관형절 술어의 선행어로 삼는 것은 아니며, 주어로 복원했을 때 의미적으로 문제가 없어야 한다.

- (3) 출판계 뒷이야기를 종합하면, 국내 ‘하루키 돌풍’에 <독자가> 일조하는 남성 독자 상당수가 군복무 때 그의 소식을 처음 접한다고 합니다.

(3)의 ‘남성 독자’는 (2)의 동격 구문과 비슷한 구조로 보인다. 그러나 서술어 ‘일조하는’의 의미적인 주어는 ‘독자’이므로 생략된 주어를 ‘독자’로 복원한다.

[실제 태깅 예시]

9) 본 사업에서는 ‘선생 조영권’과 같은 어순은 동격으로, ‘조영권 선생’과 같은 어순은 동격이 아닌 것으로 구문 분석한다. 따라서 ‘김철수 시인’이라는 표현이 말뭉치에 나온다면 이는 동격이 아닌 것으로 분석할 것이다.

(4) 서울에 <동생이> 살고 있는 **동생** 상희(35)씨는 언니와 조카의 참변 소식을 듣자마자 충격을 받고 몸져 누운 것으로 알려졌다.

‘서울에 살고 있는 동생 상희(35)씨’는 [[서울에 살고 있는 동생] 상희(35)씨]와 [[서울에 살고 있는 [동생 상희(35)씨]]로 중의적 해석이 가능하다, 따라서 인접성 우선에 따라 [[서울에 살고 있는 동생] 상희(35)씨]로 구문 분석하고, ‘살고’의 주어로 ‘동생’을 복원하였다.

(5) ‘무관의 여왕’ 디나라 사피나(러시아·1위)와 <뷰티가> 돌아온 ‘러시안 뷰티’ 마리아 샤라포바(31위)는 2라운드에 진출했다.

(5)에서 ‘러시안 뷰티’와 ‘마리아 샤라포바(31위)’는 동격으로 간주하여 ‘돌아온’의 주어로 선행 명사구의 중심어인 ‘뷰티’를 복원하였다.

(6) 런리버노스는 기타와 보컬을 <리더가> 맡고 있는 **리더** 알렉스 황, 기타 다니엘 채, 바이올린 제니퍼 임, 베이스 조 전, 키보드 셸리 강, 그리고 드럼 존 정 등 한인2세 6명으로 구성된 록밴드다.

(6)에서 ‘리더’와 ‘알렉스 황’은 동격 구성으로, ‘맡고’는 서술어와 더 인접한 ‘리더’와 의존 관계를 맺는다. 이에 따라 ‘맡고’의 생략된 주어를 ‘리더’로 복원한다.

(7) 미국 중북부를 <액세스가> 가로지르는 ‘다코타 액세스’ 송유관 건설 사업 중단 결정이 발표된 4일, 시위대와 경찰의 충돌로 매캐한 최루탄 냄새가 진동했던 스탠딩 락 천막시위 농성장은 북 소리 울리는 축제의 장으로 변했다.

‘다코타 액세스’와 ‘송유관’은 동격으로, 선행하는 명사구인 ‘다코타 액세스’의 마지막 어절 ‘액세스’를 ‘가로지르는’의 주어로 복원한다.

(8) 경찰은 또 가짜 유명 의류가 판매되는 줄 <팀장이> 알면서도 이를 <팀장이> 모른 채한 인터파크의 패션사업 담당 **팀장** 기아무개(33)씨도 불구속 입건했다.

‘팀장’과 ‘ㄱ아무개(33)씨’는 동격이므로 ‘알면서도’와 ‘모른’의 주어로 선행 명사구인 ‘팀장’을 복원하였다.

(9) 18세기 말 청나라를 <연암이> 여행한 연암 박지원의 <열하일기> 21세기 버전이랄까.

서술어 ‘여행한’의 후행 명사구 중, ‘연암’과 ‘박지원’은 동격 관계에 있으므로 ‘여행한’의 주어를 ‘연암’으로 복원하였다.

(10) ‘서부전선’은 <국군이> 농사짓다 <국군이> 끌려온 국군 남북(설경구)이 잃어버린 일급작전비밀문서(비문)가 인민군 탱크병 영광(여진구)의 손에 들어가면서 꼬이는 이야기다.

(10)에서도 ‘국군’과 ‘남북(설경구)’를 동격 구성으로 보아 선행하는 명사구 ‘국군’을 주어로 복원하였다.

(11) 현재 중국 내부에는 400년간의 차이가 공존하고 있다고 <작가가> 지적한 <작가가> 저명한 중국 작가 위화의 말처럼 중국의 도농간, 빈부간 격차는 상상을 초월할 정도입니다.

‘중국 작가’는 ‘위화’와 동격으로 분석하여, ‘지적한’과 ‘저명한’의 주어로 ‘작가’를 복원하였다.

(12) 경기도 안산에 <제조업체가> 있는 친환경 부품세척기 **제조업체** 일성리사이클링의 김성일 대표는 지방자치단체의 도움으로 재기에 성공한 경우다.

(12)의 문장은 [[경기도 안산에 있는 친환경 부품세척기 제조업체] 일성리사이클링]으로 구문 분석하여 서술어 ‘있는’의 주어로 ‘제조업체’를 복원하였다.

4) 대해, 관해, 통해 등 복원 대상에서 제외한 술어의 복원

‘에 대해, 에 관해, 에 따라, 에 의해, 를 통해’ 등은 선행하는 주어에 있으면 의존 관계에 있다고 판단하지만, 주어 생략된 경우에는 복원 대상으로 삼지 않는다.

(1) **재판부**는 국가보안법 위반 혐의에 대해 “피고인들이 검찰에서 조사받을 때 ‘고문과 가혹행위를 받은 적이 없다’고 했지만 피고인들의 불법 감금 기간이 상당히 오래이고 검찰이 증거로 제시한 진술서도 상당한 기간이 지난 뒤 작성된 점 등으로 미룰 때 검찰의 조서는 증거가 될 수 없다”고 <재판부가> 밝혔다.

한국정보통신기술협회(TTA) 구문 분석 지침에는 ‘대해, 관해, 통해’ 등의 술어에 대한 별다른 지침이 없으므로 사전 격틀에 따라 다음과 같이 분석하였다.

※ 구문 분석) 재판부는 → NP_SBJ 대해
 대해 → VP 밝혔다.

따라서 ‘밝혔다’는 구문 분석상 지배하는 주어가 없으므로 주어 복원 대상 술어로 삼고, ‘재판부’를 주어로 복원하였다.

한편, ‘대해, 관해, 통해’ 등의 서술어에 대한 구문 분석상 주어가 없을 경우에는
는 이들을 주어 복원 대상 술어로 보지 않는다.

(2) 이번을 계기로 지뢰 사용 자체에 대해 다시 한번 생각해야 합니다.

(2)의 문장에서는 ‘대해’에 대한 가시적 주어가 등장하지 않지만, ‘대해’를 주어로 복원 대상 술어로 삼지 않는다. 따라서 ‘대해’의 주어를 따로 복원하지 않았다.

[실제 태깅 예시]

(3) 엘지그룹 고위임원은 5일 오전 세종시로 계열사 이전 검토설에 관한 언론보도에 대해 어이없다는 반응을 <고위임원이> 보였다.

‘고위임원’은 ‘대해’에 의존하므로, ‘보였다’의 주어로 ‘고위임원’을 복원하였다.

※ 구문 분석) 고위임원은 → NP_SBJ 대해
 대해 → VP 보였다.

(4) 서울시교육청은 김정자 교육위원의 현황 파악 요구에 따라 지난 11일부터 각급 사립학교 현장에 공문을 <서울시교육청이> 내려보내 ‘서울시교육청 소속 공무원의 사학기관 재취업 현황’을 <서울시교육청이> 조사하고 있다.

주어 ‘서울시교육청’은 ‘따라’에 의존하므로 ‘내려보내’와 ‘조사하고’의 주어로 ‘서울시교육청’을 복원하였다.

※ 구문 분석) 서울시교육청은 → NP_SBJ 따라
 따라 → VP 내려보내
 내려보내 → VP 조사하고

(5) 시위가 격화되자 경찰은 시위대에 최루탄과 고무총, 물대포를 이용해 무력 진압을 시도했고, 이 모습들이 사회관계망서비스(SNS)를 통해 <모습들이> 알려지면서 국제적인 관심을 끌었다.

‘모습들’은 ‘통해’에 의존한다. 따라서 ‘알려지면서’의 주어로 ‘모습들’을 복원하였다.

※ 구문 분석) 모습들이 → NP_SBJ 통해
 통해 → VP 알려지면서

(6) 성전환자 평등센터의 마라 케이슬링 이사는 성명을 통해 “내 친구 걸스팬이 백악관 직원으로 임명된 것은 그의 활동에 감동을 받아온 나를 비롯한 수많은 사람들을 고무시키고 있다”고 <이사가> 밝혔다.

모문의 주어 ‘이사’가 ‘통해’에 의존하는 것으로 보고, ‘밝혔다’의 주어로 ‘이사’를 복원하였다.

※ 구문 분석) 이사는 → NP_SBJ 통해
 통해 → VP 밝혔다.

(7) 일본은 1980년대 6년에 걸친 치열한 협상을 통해 미국으로부터 사용후 핵연료의 재처리를 포함한 핵 활동에 대한 ‘포괄적 (사전)동의’를 확보한 상태다.

위의 문장에서 ‘핵 활동에 대한’의 ‘대한’은 구문 분석상 지배하는 주어가 없다. 이 경우 주어를 복원하지 않기로 정하였으므로 어떠한 주어로도 복원하지 않았다.

(8) 한국형 전투기 개발 사업(KFX·보라매사업)과 관련해, 지난해 5월 주철기 청와대 외교안보수석이 관련 전문가들로부터 미국의 기술 이전 불허 가능성을 이미 보고받았다는 <한겨레>의 보도(10월 7일치 1면)에 대해 청와대가 해명에 나섰다.

‘관련해’는 주어가 생략되어 있지만, 복원 대상 술어에서 제외되어 주어를 복원하지 않았다.

(9) 하지만 기술 이전 불허를 미리 알았는지 여부와 당시 국방부 장관이던 김관진 국가안보실장의 기종 변경 과정 개입 등 핵심 쟁점에 대해서는 구체적 언급을 피했다.

‘대해서는’은 주어 복원 대상 술어가 아니므로 주어를 복원하지 않았다.

(10) 설사 정책의 변화가 있더라도 현 외교안보 라인을 통해 이뤄질 것이라는 이야기다.

‘통해’는 주어 복원 대상 술어가 아니므로 주어를 복원하지 않았다.

(11) 정부 당국자는 “정부 내에 최근 북한이 남쪽과의 비핵화 회담을 수용하는 등 남북대화 재개를 위해 ‘유화적인’ 자세를 보이는 것은 정부의 ‘원칙 있는’ 태도

때문이라는 인식이 많다”고 전했다.

‘위해’는 주어 복원 대상 술어가 아니므로 주어를 복원하지 않았다.

5) 관형절 내포문

한국정보통신기술협회(TTA) 구문 분석에서 모문과 관형절 내포문의 주어가 동일한 경우, 그 주어는 모문의 서술어에 의존한다. 따라서 내포문 서술어에 대한 주어를 복원해야 한다.

(1) 그는 마라톤 풀코스를 11번 <그가> 완주한 달리기 애호가이다.

(1)에서 공통 주어인 ‘그’는 모문의 서술어 ‘애호가이다’에 의존한다. 내포문의 서술어 ‘완주한’은 복원 대상 술어가 되므로, 주어로 ‘그’를 복원하였다.

또한 내포문이 두 개 이상일 경우에는, 인접성 우선에 따라 선행하는 두 절의 의존 관계가 가능한지 먼저 파악하고, 무형 대용어 복원 분석도 이에 따랐다. 이 두 원칙을 고려하면 결국 관형절의 위치에 따라 무형 대용어 복원 술어가 결정된다.

[실제 태깅 예시]

(2) 몬테네그로 출신의 이 **남성** 역시 크로아티아→슬로베니아→오스트리아 등 전형적인 ‘난민 루트’를 <남성이> 거친 것으로 확인됐다.

(2)의 문장은 모문과 관형절 내포문의 주어가 ‘남성’으로 동일하다. 따라서 ‘남성’은 모문의 서술어인 ‘확인됐다’에 의존하므로, 내포문 서술어 ‘거친’이 주어 복원 대상 술어이다.

한편, 내포문이 두 개 이상일 경우 즉, 세 개 이상의 절로 구성된 문장에서는 선행하는 두 절의 의존 관계가 가능한지 먼저 파악하고, 무형 대용어도 복원 분석도 이에 따랐다. 따라서 관형절의 위치에 따라 무형 대용어 복원 술어가 결정된다.

(3) **모스테파이**는 2010년 무렵 경범죄를 <모스테파이가> 저지른 후 급진적인 벨기

에 출신 이슬람 지도자를 만나 과격 이슬람주의에 <모스테파이가> 빠진 것으로 <모스테파이가> 알려졌다.

(3)의 문장을 분석할 때에는 선행하는 두 절의 의존 관계가 가능한지 먼저 파악해야 한다. 선행하는 두 절 즉, ‘모스테파이는 2010년 무렵 경범죄를 저지른 후 급진적인 벨기에 출신 이슬람 지도자를 만나’의 결합은 해석 및 구문 분석이 가능하므로 두 절의 의존 관계를 우선 분석한다. 이때 선행하는 절이 관형절이므로 ‘모스테파이’는 후행하는 부사절 술어인 ‘만나’에 의존한다. 이에 따라 ‘저지른’, ‘빠진’, ‘알려졌다’의 주어로 ‘모스테파이’를 각각 복원하였다.

(4) 67세 흑인 여성이 약 2m 아래 지하로 <여성이> 추락하는 사고를 당했다고 9일 (현지 시각) 보도했다.

위의 문장에서 주어 ‘여성’은 후행 인용절 서술어 ‘당했다고’에 의존한다. 따라서 선행 관형절 서술어 ‘추락하는’이 복원 대상 술어가 된다.

(5) 식료품 가격은 OECD 평균보다 3% 높으며 특히 고기 우유 치즈 달걀 과일 등이 비싼 것으로 <가격이> 나타났다.

반면, (5)와 같이 처음 출현하는 절이 관형절이 아니면 공통 주어는 선행하는 술어에 의존한다. (5)에서는 부사절이 먼저 출현하였으므로 선행절 간의 관계를 파악하여 ‘식료품 가격은 OECD 평균보다 3% 높으며 특히 고기 우유 치즈 달걀 과일 등이 비싸다’가 먼저 의존 관계를 맺고, 다시 이 전체가 ‘나타났다’와 의존 관계를 맺는 것으로 분석한다. 따라서 주어 ‘가격’은 내포문 서술어 ‘높으며’에 의존하고, 모문의 서술어 ‘나타났다’의 주어로 ‘가격’을 복원하였다.

6) 보조 용언 및 장형 사동

보조 용언 ‘놓다, 대다, 말다, 앓다, 있다, 주다, 하다’ 등은 주어 복원 대상 술어에서 제외하였으며, 보조 용언 목록은 『표준국어대사전』을 참조하였다.

(1) 덤으로 주는 티켓 값은 정부 예산으로 <정부가> 지원해 준다.

(1)에서 본용언인 ‘지원해’는 주어 복원 대상 술어이므로 주어인 ‘정부’를 복원하지만, 보조 용언 ‘준다’는 복원 대상 술어에서 제외한다.

(2) 리오넬 메시(바르셀로나)를 <누군가> 떠올리게 하는 폭발적인 드리블과 뛰어난 결정력으로 ‘지메시’라 불리는 지소연은 이번 월드컵에서 든든한 공격 파트너와 함께한다.

특히, 장형 사동 ‘-게 하다’의 경우, ‘본용언+보조 용언’을 하나의 VP로 보지 않고, ‘본용언’의 의미에 해당하는 주어를 복원하였으며, 후행 보조 용언의 주어는 복원하지 않았다. (2)에서 ‘떠올리게’의 주어는 일반적인 사람으로, 문서 내에 나타나지 않아 ‘누군가’로 복원하였다.

[실제 태깅 예시]

(3) 최근 세종문화회관 대극장에서 개막해 연일 매진행진을 <모차르트가> 이어가고 있는 ‘모차르트’는 대표적인 ‘비엔나(오스트리아) 뮤지컬’이다.

‘이어가고 있는’과 같이 ‘본용언+보조 용언’의 구성인 경우, 보조 용언은 무형 대용어 복원 분석 대상에서 제외한다. 따라서 ‘이어가고’의 주어는 복원했지만, ‘있는’의 주어는 복원하지 않았다.

(4) 우리는 각 나라의 시장에 대해 잘 <우리가> 알지 못합니다.

‘알지 못합니다’에서 ‘못하다’는 보조 용언이므로, ‘알지’의 주어만 복원하였다.

(5) 대학도 졸업하기 전에 수천만원 빚을 <여러분이> 지게 하는 대한민국을 갈아엎을 힘과 권리가 여러분에게 있다.

장형 사동 표현인 ‘-게 하다’에서도 본용언에 대해서만 주어를 복원하였다. 따라

서 ‘지게’의 주어로 ‘여러분’을 복원하고, ‘하는’의 주어는 복원하지 않았다.

(6) 그러나 한달여가 지난 뒤 피트니스클럽이 휴업을 하게 <무언가> 돼서 더 이상 이용할 수 없게 <무언가> 됐다.

‘-게 하다’와 달리 ‘-게 되다’의 ‘되다’는 『표준국어대사전』에 본용언으로 등재되어 있다. 따라서 이때 ‘되다’는 주어 복원 대상 술어이며, (6)의 ‘돼서’, ‘됐다’는 문서 내 적절한 선행어가 없어 <무언가>로 복원하였다.

7) 의사 보조 용언

한국정보통신기술협회(TTA) 지침에서 ‘-ㄴ 예정이다/전망이다/계획이다/처지이다’ 등 의사 보조 용언으로 처리한 목록은 보조 용언과 같이 주어 복원 대상 술어에서 제외한다.

(1) 세월호 선체 인양 작업이 마무리되면 세월호를 목포신항으로 <해수부가> 웁길 예정이다.

‘-ㄴ 예정이다’는 의사 보조 용언으로 분류되어 주어 복원 대상에서 제외된다. 따라서 ‘웁길’의 주어만 복원하였다.

[실제 태깅 예시]

(2) 당장 내년부터 2군 리그에 <엔씨소프트가> 참가할 계획이다.

‘-ㄴ 계획이다’는 의사 보조 용언이므로 주어 대상 술어에서 제외된다. 따라서 (2)에서는 ‘참가할’의 주어만 복원하였다.

(3) 지엠대우는 이번 새 모델을 역전의 발판으로 삼을 계획이다.

‘-ㄴ 계획이다’도 ‘-ㄴ 예정이다’와 마찬가지로 의사 보조 용언으로 간주한다.

‘지엠대우’는 ‘삼을’에 의존하며, ‘-ㄴ 계획이다’는 구문 분석상 주어가 없지만 무형 대응어를 복원하지 않는다.

(4) 이 소송은 서울시의 2차 조사를 이유로 일시 중단된 상태다.

‘-ㄴ 상태다’도 의사 보조 용언으로 보아 복원 대상이 아니다. 주어 ‘소송’은 ‘중단된’에 의존하는 것으로 분석한다.

(5) 올해부터 한국에서도 이런 징벌적 손해배상 제도가 도입될 전망이다.

‘-ㄴ 전망이다’는 추측을 나타내는 의사 보조 용언으로 보아 주어를 복원하지 않았다.

그 외 ‘-ㄴ 이유다’, ‘-ㄴ 생각이다’ 등도 의사 보조 용언으로 간주했다.

의사 보조 용언 리스트는 한국정보통신기술협회(TTA)의 지침을 참조하고, 보조 용언과 유사한 기능을 하는 양상 의미를 가진 용언류 3개를 추가하였다. 위의 (2)~(5)의 예가 해당된다.

의사 보조 용언 리스트	
TTA 지침에 제시된 의사 보조 용언	-ㄴ 수/리(가) 있다/없다, -ㄴ/ㄹ {것 터 뿐 따름 모양 지경 참 중 노 릇 예정 길}이다, -ㄴ {만 법 듯}하다, -는 말이다, -ㄴ/ㄹ 듯(도) 하다, -ㄴ/ㄹ 것 같다, -ㄴ 것을(걸) 그랬다, -어서는 안 된다, - 고 해서, -든지 하다
본 사업에서 추가한 의사 보조 용언	-ㄴ 전망이다/계획이다/처지이다

2.3.2. 구어

주격 무형 대용어 복원 말뭉치는 문어 텍스트뿐 아니라 구어 텍스트 또한 구축 대상으로 한다. 구어 분석의 기본 방향은 문어와 동일하나, 주어가 대화문 내에 드러나지 않는 경우가 빈번하여 일반적인 사람이나 사물을 가리키는 ‘누군가’, ‘무언가’로 복원되는 빈도가 매우 높다. 또한 ‘걔(개는)’, ‘누굴(누구를)’, ‘학꾼(학교는)’ 등의 축약형의 등장도 빈번하였는데, 축약형의 경우에는 원래 형태를 주어로 복원하였다. 이외에도 담화 장치와 복원 대상 술어의 구분이 모호하다는 점 등 구어 말뭉치가 갖는 특수성에 근거하여 구어 세부 지침을 별도로 수립하였다.

1) 발화 단위

본 사업에서 구어 말뭉치 분석은 발화 단위를 기준으로 한다. 동일 화자가 발화한 하나의 문장처럼 생각되어도 원시 말뭉치에서 다른 발화 단위를 구성하면 별도의 분석 대상으로 간주한다.

(1)

1	이~ 하늘에 관문인 인천국제공항이
2	국민들의 자부심으로 <인천국제공항이> <u>떠오르고</u> 있습니다.

(1)의 용례는 본래 하나의 문장이지만, 발화 단위에 따라 1과 2로 구분되므로 각각의 발화 단위를 기준으로 무형 대용어를 복원한다. 따라서 ‘떠오르고’의 주어로 ‘인천국제공항’을 복원하였다.

[실제 태깅 예시]

(2)

1	그러니까 그런 진영에도 후보를 내겠다고 하면 진영들은
2	일월 삼십일일까지 한 사람이나 두 사람들 실무자를 <진영들이> <u>내 갖고</u>

(2)는 1-2번이 하나의 문장을 이루는 것으로 보인다. 만약 1-2번이 하나의 발화 단

위로 제시되었다면 ‘내갓고’는 ‘진영들’을 주어로 취하는 것으로 구문 분석했을 것이다. 그러나 현 말뭉치에서는 두 개의 발화 단위로 분리되어 있으므로 각각을 분석 대상으로 삼는다. ‘내갓고’에 의존하는 주어가 없으므로 선행 발화의 ‘진영들’을 주어로 복원하였다.

(3)

1	그까 그런 점에서 반기문 총장이 친동생과 조카 가
2	그것도 외국검찰에 <조카가> <u>기소됐다</u> 카는 거는

(3)도 동일 화자가 발화한 하나의 문장이지만, ‘기소됐다’의 주어인 ‘조카’가 선행 발화에 속하므로 ‘기소됐다’는 무형 대용어 복원 술어이다.

(4)

1	최순실 씨가
2	직접 <씨가> <u>지시한</u> 것 아니냐.

‘최순실 씨가’와 ‘직접 지시한 것 아니냐’의 발화 단위가 다르므로, ‘지시한’의 주어를 ‘최순실 씨’의 중심어인 ‘씨’로 복원하였다.

(5)

1	그래서 해경 은
2	개혁안을 <해경이> <u>준비하고</u> 있었는데 바로 그 다음 날

(5)에서도 마찬가지로 1과 2의 발화 단위가 다르므로, ‘준비하고’의 주어를 ‘해경’으로 복원하였다.

2) 선행어

가. 인칭 통일

구어 말뭉치의 특성상 화자나 청자가 선행어일 경우가 많으며, 이때 문어 분석과 마찬가지로 인칭에 유의한다. 1인칭에 해당하는 대명사나 1인칭을 지칭하는 보통 명사가 문서 내에 없는 경우에는 화자를 ‘누군가’로 복원한다. 마찬가지로 2인칭에 해당하는 대명사나 2인칭을 지칭하는 보통 명사가 문서 내에 없는 경우에도 청자를 ‘누군가’로 복원한다.

(1)

P1	<u>내가 어제 명동에 갔었어. 거기서 그 사람을 <u>봤지</u> 뭐야.</u> <내가>
P2	<u>누굴 <u>봤는데</u> ? 궁금해. 알려줘.</u> <누군가> (←P1을 가리키는 ‘너’가 문서 내 없음)
P1	<u>그 티비에 <u>나오는 사람</u> 김철수!</u> <사람이>
P2	<u>아 어땠어?</u> <김철수가>
P1	<u>티비랑 <u>똑같았어</u> . 넌 본 적 없지?</u> <김철수가>
P2	<u>응 <u>없어</u> .</u> <누군가> (←P2를 가리키는 ‘나’가 문서 내 없음)

(1)에서는 P1과 P2의 인칭에 맞도록 선행어를 복원하였으며, 1인칭 혹은 2인칭을 지칭하는 명사가 없는 경우에는 ‘누군가’로 복원하였다. ‘김철수’는 제삼자이므로 이름 그대로 복원하였는데, ‘티비에 나오는 사람 김철수’는 동격 명사구이므로 해당 발화에서는 ‘사람’을 복원하였다.

[실제 태깅 예시]

(2) 근데 니가 혹시 창피해서 못 <니가> 일어나면 내 남방을 <나는> 주겠다. 남방을 <니가> 덮고 <니가> 가라.

(2)는 1, 2인칭 대명사가 선행어로 분석된 사례이다. 문장 내에 ‘니’와 ‘나’가 선행어로 존재하므로 무형 대용어 복원 술어에 각각의 선행어를 복원하였다.

(3) 아까 영상에서도 나왔던 그 아드님에게 피임 교육을 초등학교 때부터 <누군가> 시키셨다면서요?

‘시키셨다면서요’의 생략된 주어는 청자로, 2인칭 대명사나 보통 명사로 주어를 복원해야 한다. 해당 문서에는 2인칭 명사구가 없어서 <누군가>로 복원하였다.

(4)

P1	여러분 교육과학기술부 이주호 장관을 박수로 <여러분이> 맞이하겠습니다.
	현장을 많이 <장관이> 다니신다고 제가 들었습니다.
	근황을 좀 잠깐 <장관이> 설명해주신다면.
P2	기업들 많이 <제가> 찾고 있고요.
P1	자 오늘 이제 본격적인 대답을 <우리가> 나누기 전에

P1 화자의 발화에서는 P2 화자를 일반 명사인 ‘장관’으로 복원하였다. 네 번째와 다섯 번째 발화의 서술어 ‘찾고’와 ‘나누기’는 주어 복원 대상 술어이므로 문서 내에서 선행어로 등장하는 대명사 ‘저’와 ‘우리’로 주어를 각각 복원하였다.

나. 축약형

문서에서 선행어와 조사 등이 결합하여 축약형으로 표기된 경우, 형태소 원형을 기준으로 주어를 복원한다.

(1) 그러고보면 충분히 어 <이야기가> 있을 법한 이야긴데.

(1)에서 ‘이야긴’은 ‘이야기’로 원형 복구하여 ‘있을’의 주어로 복원하였다.

[실제 태깅 예시]

(2)

P1	이 양팔 보시면 윤기가
	기름에다 볶았다면
	하얗게 <양파가> <u>일어났을텐데</u>

‘일어났을텐데’의 생략된 주어는 ‘양파’이다. 본 사업에서는 조사 결합 축약형은 형태소를 분리하여 원형을 복원하므로 ‘양팔’이 아닌 ‘양파’로 복원하였다.

(3)

P1	삼 분의 일은 출소잡 니다.
	남들보다 더 열심히 <출소자가> <u>하고요</u>

(3)에서 ‘하고요’는 주어 복원 대상 술어이므로, 선행어 ‘출소잡’을 원형 복구한 ‘출소자’로 주어를 복원하였다.

(4)

P1	네 봄의 전령사를 <요리가> <u>대표하는</u>
	마지막 요린 데요

‘요린’을 ‘요리’로 원형 복구하여 ‘대표하는’의 주어로 복원하였다.

(5)

P1	이게 정치적 오염으로부터
	<이거는> <u>자유롭지</u> 않다.

(5)에서는 ‘이게’를 ‘이거’로 원형 복구하여 주어를 복원하였다.

(6) 아님 **자길** <자기가> 알릴 수 있다든지

(6)에서는 ‘자길’을 ‘자기’로 원형 복구하여 ‘알릴’의 주어로 복원하였다.

다. 후행하는 주어의 복원

주어가 서술어에 후행하는 경우는 첫째, 주어가 한 문장 내에서 도치된 경우 둘째, 발화 단위가 분리되어 후행 발화 단위에 주어가 등장하는 경우이다.

먼저, 주어가 문장 내에서 도치되어 서술어에 의존하는 주어가 없을 경우, 해당 서술어는 주어 복원 대상이 된다. 이때 문장 내의 도치된 주어를 선행어로 삼는다.

(1) 커트라인에 <점수대가> 몰려 있어요 점수대가

※ 구문 분석) 커트라인에	→ NP_AJT	몰려
몰려	→ VP	있어요
있어요	→ VP	점수대가
점수대가	→ NP_SBJ	ROOT

‘몰려 있어요’의 주어는 ‘점수대’이지만, 주어가 도치되어 있어 구문 분석상 ‘몰려’에 의존하는 주어는 없다. 따라서 ‘몰려’는 주어 복원 대상 술어이며, 같은 문장 내의 ‘점수대’를 선행어로 삼아 주어를 복원하였다.

다음으로, 하나의 발화 단위 내에 주어가 없을 경우에는 연속된 발화 단위가 하나의 문장처럼 보이더라도 별도의 분석 대상으로 간주한다.

(2)

P1	제대로 <사람이> <u>합시다</u> 소개도 안 <사람이> <u>해줘</u> 심지어
	아니 뭐 <u>사람이</u>

(2)는 두 개의 발화 단위로 이루어져 있다. 발화 흐름상 하나의 문장으로 발화하였으나 휴지 구간 등으로 인해 별도의 발화 단위로 분리되었을 것이다. 구어에서는 발화 단위를 분석 기준으로 삼으므로 ‘합시다’와 ‘해줘’에 의존하는 주어가 없고, 이에 대한 무형 대응어를 복원해야 한다. 기본 분석 원칙에 따라 선행 발화에 선행어가 있다면 그것으로 복원하고, 선행 발화에 선행어가 없다면 후행 발화에서 선행어를 탐색한다. (2)에서는 후행하는 선행어인 ‘사람’으로 주어를 복원하였다.

[실제 태깅 예시]

(3) 편히 <코리아이버슨이> 잠드세요 코리아이버슨

‘잠드세요’의 주어인 ‘코리아이버슨’이 동일 문장 내에서 서술어에 후행하므로 ‘잠드세요’의 주어를 ‘코리아이버슨’으로 복원하였다.

(4) 저도 군청에 있다 왔는데 <거리가> 가까운 줄 알았어요 거리가

구문 분석상 ‘가까운’에 의존하는 주어가 없으므로 ‘거리’를 선행어로 복원하였다.

(5) 뭐 <오빠가> 먹고 싶은 거 없어 오빠는?

‘먹고’의 주어인 ‘오빠’가 서술어 뒤로 도치되어 있으므로 ‘먹고’는 주어 복원 대상 술어가 된다. 따라서 ‘먹고’의 주어를 ‘오빠’로 복원하였다.

(6)

P2	인제 거기를 이제 이~ 봉사 활동을 <대통령이> <u>하는데</u>
	<u>대통령</u> 께서

(6)은 두 개의 발화 단위로 이루어져 있으며, 선행 발화의 ‘하는데’에 의존하는 주어가 존재하지 않는다. 따라서 후행 발화의 ‘대통령’으로 주어를 복원하였다.

3) ‘그렇다’ 부류

‘그렇죠’를 복원 대상 술어로 볼 것인지에 관한 기준은 명확하지 않다. 본 사업에서는 선행 문맥 전체를 지시하는 의미이면 담화 장치로 보아 분석 대상에서 배제한다. 반면에 특정 명사를 지시하는 의미이면 문서 내 선행어를 주어로 복원한다.

‘그렇다’ 부류의 주어 복원은 기본적으로 문어 분석과 동일하다. 우선 선행어 명사를 찾고, 문서 내 선행어 명사를 찾지 못할 때에 한하여 ‘무언가’로 태깅한다.

또한 동일 발화 내, 선행 발화, 후행 발화 순으로 선행어를 탐지한다. ‘그렇다’ 부류에는 ‘그렇죠, 그렇군요, 그렇습니다, 그래요, 그랬더니, 그쵸’ 등이 해당된다.

(1)

P1	멸치 국물은 뭐 그~ 기본적으로 사실은 낼 줄 아셔야 되는
P2	네. <국물이> <u>그렇죠</u> .

위 문장에서 ‘그렇죠’의 주어는 선행 발화의 주어인 ‘멸치 국물’과 동일하므로 생략된 주어를 ‘국물’로 복원하였다.

[실제 태깅 예시]

(2) 어~ <대통령이> 그렇기 때문에

‘그렇다’ 부류의 주어 복원 시, 문서 내 구체적 대상이 있는 경우 그것을 선행어로 삼는다. (2)에서는 ‘대통령’이라는 구체적 명사구를 주어로 복원하였다.

(3)

P1	좀 익혀주시면은 국물 맛 이 시원하거든요.
P2	<맛이> <u>그렇군요</u> .

(3)에서는 ‘그렇군요’의 주어를 선행어 ‘맛’으로 복원하였다.

(4)

P1	늘푸른한국당 이 일단 원외정당이지않아요?
P2	<늘푸른한국당이> <u>그렇습니다</u> .

(4)에서는 ‘그렇습니다’의 주어를 선행 발화의 주어인 ‘늘푸른한국당’으로 복원하였다.

(5)

P1	분권형 개헌 을 하려면 결국에는 다른 당하고 같이 하셔야 하는 거 아닌가요?
P2	<개헌이> <u>그렇습니다</u> .

(5)에서는 ‘그렇습니다’의 주어를 선행 발화의 ‘개헌’으로 복원하였다.

(6)

P1	그게 일반적인 관계에서는 사실 상대방 저렇게 왜 행동하는지
	별로 고민 안 <누군가> 하잖아요.
P2	<누군가> <u>그렇죠</u> .

맞장구의 의미를 지니는 ‘그렇다’의 경우에도 특정 명사를 지시하는 것으로 해석되면 문맥에 따라 <누군가> 혹은 <무언가>로 복원한다. (6)의 ‘그렇죠’는 앞선 발화 내용인 ‘(사람들이) 고민을 안 한다’에 맞장구치는 것으로, ‘사람들이 그렇다’의 의미를 나타낸다. 하지만 문서 내에 일반적인 사람을 지칭하는 선행어가 없으므로 <누군가>로 주어를 복원하였다.

(7)

P1	법관의 자의적이거나 주관적인 그런 판단이 개입될 여지가 있다.
P2	음. 법관의 재량에 따라서.
P3	<무언가> <u>그렇죠</u> .

(7)의 ‘그렇죠’는 ‘재판이 그렇죠’ 혹은 ‘사람이 그렇죠’ 등으로 해석된다. 여기에서는 생략된 선행어를 ‘재판’ 등으로 판단하여 ‘그렇죠’의 주어를 ‘무언가’로 복원하였다.

4) 띄어쓰기 오류

본 사업의 구어 말뭉치는 문어 말뭉치에 비해 띄어쓰기 오류가 빈번하였는데, 발화자나 전사자에 따라 휴지 구간이나 띄어쓰기에 차이가 발생하는 것으로 보인다. 이에 띄어쓰기 오류(마크업 기호 문제 포함)에 대해서는 바른 띄어쓰기를 기준으로 선행어와 술어의 범위를 지정하였다. 단, 잘못 띄어 쓴 경우는 복수의 어절에 대해 선행어와 술어 범위를 지정하였으나, 잘못 붙여 쓴 경우는 어절의 일부만 범위를 지정하지는 않았다.

- (1) 김철 수는 밥 먹고, 커피 <김철 수가> 마셨어.
- (2) 난 밥 먹고, 커피 <나는> 마 셴어.
- (3) 난 밥 먹고, 커피 <나는> 마시고있어.

(1)의 주어 ‘김철 수’는 본래 ‘김철수’로 붙여 써야 한다. 따라서 바른 띄어쓰기를 기준으로 ‘김철 수’ 전체를 선행어로 지정하여 ‘마셨어’의 주어로 복원한다. (2)의 서술어 ‘마 셴어’ 역시 ‘마셨어’로 붙여 써야 하는데 띄어쓰기 오류가 발생한 경우로, 바른 띄어쓰기를 기준으로 ‘마 셴어’ 전체를 서술어로 지정한다. 반면, (3)의 ‘마시고있어’는 본용언+보조 용언 구성으로, 본래 ‘마시고 있어’로 띄어 써야 하지만 잘못 붙여 쓴 경우이다. 해당 경우에는 어절의 일부인 ‘마시고’만 지정하지 않고, ‘마시고있어’ 어절 전체를 서술어로 지정한다.

[실제 태깅 예시]

- (4) 초반에 싸이렌이 울렸을 때 그걸 엄격하게 <누군가> 들여다 보는

서술어 ‘들여다 보는’의 올바른 띄어쓰기는 ‘들여다보는’이다. 이에 두 어절로 되어 있는 ‘들여다 보는’ 전체를 하나의 술어로 범위를 지정하였다.

- (5) 왜 자기 모습을 뿔뿔히 <흑자들이> 드러 내지 못하고

서술어 ‘드러 내지’의 올바른 띄어쓰기는 ‘드러내지’이다. 따라서 하나의 술어로 범위를 지정하였다.

(6) 사실 외국 <누군가> 나가보면 전혀 안 <누군가> 쳐다 보거든요?

서술어 ‘쳐다 보거든요’의 올바른 띄어쓰기는 ‘쳐다보거든요’이다. 붙여 써야 하므로 하나의 술어로 범위를 지정하였다. 반면, ‘나가보면’은 ‘나가 보면’으로 띄어 써야 하지만 어절 전체를 복원 대상 술어로 지정하였다.

(7) 각각의 집에서 <습관을> 해오던 습관을 버리고

서술어 ‘해오던’은 ‘해 오던’으로 띄어 써야 하지만 복원 대상 술어로 본용언인 ‘해’만 분리하여 지정하지 않는다. 어절 단위를 기준으로 ‘해오던’ 전체를 서술어로 지정하였다.

주격 무형 대용어 복원 구축 지침은 과제 수행 기간 동안 지속적으로 수정, 보완되었다. 지침 구축 시에는 문어와 구어의 매체 특성에 관한 연구가 선행되어야 하며 각 사용역의 특성을 반영해야 한다. 또한 국립국어원의 구문 분석 말뭉치 구축, 의미역 말뭉치 구축, 말뭉치 검증 사업 등 유관 사업과의 통합을 고려하였다. 전산언어학적 활용뿐만 아니라 언어학적 연구 자료로서 가치 있는 성과물을 창출하기 위해 본 분석 지침의 세부 사항은 꾸준히 보완 및 정제되어야 한다.

3. 데이터 구축 수행 도구 활용

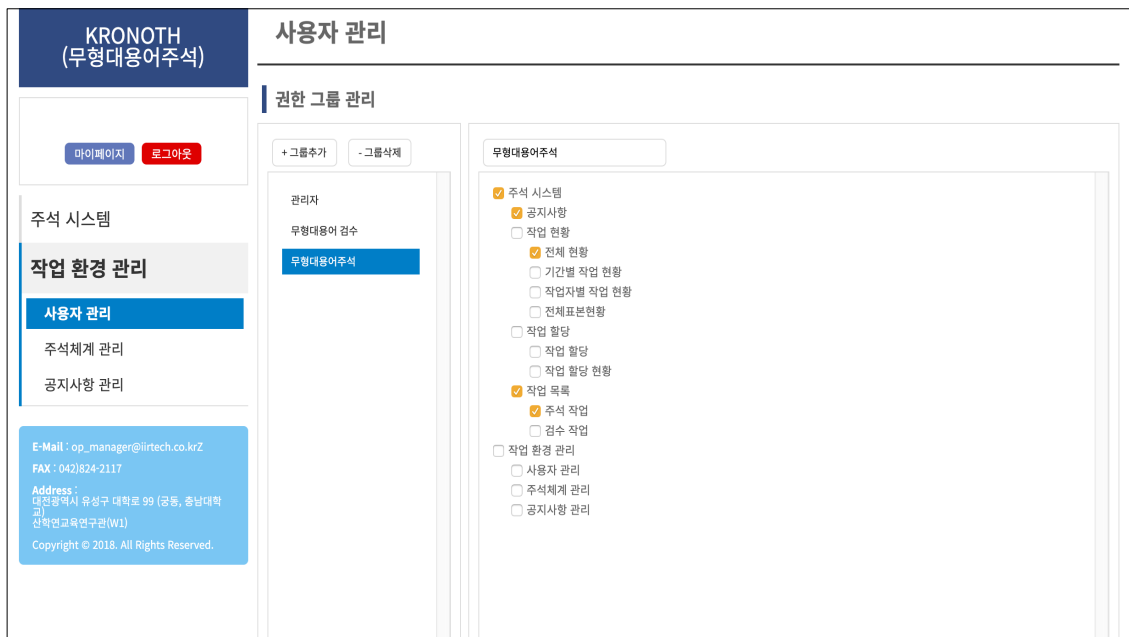
3.1. 시스템 설치 및 구성

웹 시스템인 아마존 웹 서비스(Amazon Web Service, AWS) 클라우드 서비스에 시스템을 설치하여 안정적인 클라우드 환경을 확보하였으며, 데이터 보관의 시간적, 공간적 제약 없는 작업 환경을 구성하였다.

3.2. 자료 보안 및 외부 인력 접근 제어

회원 가입 형태가 아닌 컨소시엄 내 수행 인력에 한해서만 접근 사용자 권한을 발급했다. 미승인(외부 인력) 사용자의 접근은 원천적으로 불가능하다. 외부 해킹 등에 따른 원시 혹은 가공 데이터 접근에 대한 클라우드 서비스 차원의 침입 방지 보안 서비스가 기본적으로 제공되어 자료 보안에 도움이 된다.

작업 권한 부여 시 승인된 도구 사용 권한은 작업된 데이터에 대한 외부 접근을 차단하므로, 납품 시 최고 관리자(사업 책임자)외에는 데이터 반출이 불가능하다.



<작업 권한자 접근 화면/기능 - 크로노스(KRONOTH) 주석 시스템>

3.3. 구축 도구 활용

3.3.1. 원시 데이터 검사

원시 데이터 수령 이후, 구축 도구를 활용하여 파일 열기와 파일 용량 확인을 실시하여 수령 파일 자체의 무결성을 확인하고, 단위 파일당 시스템 허용 용량을 검사하였다. 또한 주식 부착 여부 확인으로 원시 데이터의 메타 정보 등 작업을 위한 최소한의 데이터 부착 상태를 점검하였다.

분석 표지 무결성의 확인을 통하여, 부착된 분석 표지가 정상적으로 부착된 상태인지 주식 형식을 검사하고, 분석 표지 기준 확인으로 부착된 분석 표지가 사업 검증, 납품 기준에 맞게 부착되었는지 여부를 알아보는 주식 검사를 실시하였다.

예외 문자, 문자 코드를 확인하여, 파일 내 포함된 문서가 정상적인지를 확인하는 내용 형식 검사를 실시하였으며, 추출, 변환, 적재(Extract, transform, load, ETL)를 실행하여, 원시 데이터 파일을 확인했다. 정상적인 파일로 확인되면 파일 내용의 내보내기 및 시스템 변환, 탑재 작업을 수행하였다.

3.3.2. 승인된 사용자 시스템 사용 등록

시스템 사용 시에는 발주 기관의 보안 요구 사항을 준수하였다. 이르테크 공동수급체에 소속된 구축 작업자에게 비밀 유지 서약서, 보안 각서 등의 인적 보안서를 받았다. 또한 작업자의 아이디와 비밀번호를 수령 및 등록하여 작업 권한을 부여하였다.

등록된 작업자에게만 시스템 접속 URL을 제공하며 등록이 승인된 작업자만 진입이 허용된다. 시스템 최초 로그인 시 작업자는 개인별 비밀번호를 반드시 변경해야 한다. 이때 최고 관리자는 작업자별 작업 권한에 따라 시스템 접근 제어 실행 권한을 다시 부여하며, 작업자에게는 승인된 작업 화면만 활성화되고, 접근도 가능하다.

3.3.3. 등록된 원시 말뭉치 데이터의 분배

등록된 원시 말뭉치는 최고 관리자 권한으로 시스템 접근 시에만 작업 배분이 가능하며, 일별 수행 가능한 할당량과 주간 및 월간 공정량을 확인하여 최고 책임자가 작업을 할당할 수 있다. 최고 관리자가 할당한 작업은 분석 표지를 부착하는 작업과 검수하는 작업으로 구분되며, 관리자는 할당된 작업들의 실시간 상태 확인이 가능하다.

3.3.4. 할당된 작업 수행

작업자가 시스템에 접근할 때, 첫 번째 보이는 페이지는 공지 사항 페이지로, 진입 시 변경된 지침 등의 공지 사항을 확인하고 작업에 임하게 된다.

작업자는 현재 할당된 작업 및 할당된 작업 중 미수행, 미완료(임시 저장 상태)된 작업 현황을 확인하여, 선택한 작업의 작업 창으로 이동하고 작업을 수행한다. 작업 중 임시 저장을 할 수 있으며, 작업 완료 시 작업 내용을 저장할 수 있다. 작업을 완료하여 저장한 문서는 다시 작업할 수 없다.

작업 창 진입 시 원문 데이터 및 기본적인 형태소 분석 정보 등을 제공하여 작업의 편의성을 높였으며, 보류 기능과 검토 요청 기능을 보완하여, 모호한 기준의 작업일 경우 지침을 재확인하거나 상위 검수자들에게 검토를 요청하도록 하여 작업의 정확도를 높였다. 작업자가 작업 완료 후 저장한 결과 데이터는 격리되어 보존된다.

3.3.5. 수행 현황 확인과 관리

수행 현황을 확인하고 관리하는 것은 최고 관리자 권한으로 시스템에 접근할 때만 가능하다. 최고 관리자는 현재 수행 완료된 작업의 현황을 확인하고, 공정률을 확인할 수 있으며, 현재 수행 중인 작업과 남은 작업 현황을 확인할 수 있다. 기간별, 수행 작업자별 수행 현황도 확인이 가능하다. 이런 과정을 거쳐 완료된 작업 대상의 데이터를 내보내고, 납품(1차, 2차, 최종)한다.

3.3.6. 작업 결과 데이터 관리

작업 결과 데이터는 최고 관리자 권한 외에는 접근(열람, 조회, 추출 등)이 불가능한 보안 권한 체계로 구성되어 있다. 최고 관리자는 사업 책임자 부재 시 작업 배정 등을 위해 사업 책임자 외 1명으로 최소로 구성했다. 이로써 데이터의 유출 등 인적 유출 위험을 최소화했다.

데이터 납품을 위한 구축 결과 데이터 추출은 시스템 로그를 통해 이력 관리되며, 이에 대한 모든 접근자는 추적이 가능하다.

3.4. 구축 절차

지침 요소의 명확성 및 작업자 판단의 일관성을 제공하기 위하여 선행어 탐색 범위, 선행어 적용 단위, 대용어 복원 대상, 대용어 복원 위치를 기본으로 하며, 작업자는 다음과 같은 단계를 거쳐 주격 무형 대용어를 복원했다.



<주격 무형 대용어 복원 말뭉치 구축 도구 사용 단계>

3.4.1. 전처리 단계

국립국어원에서 제공하는 구축 대상 원시 말뭉치에 문단과 문장 정보를 재구성하여 주격 무형 대용어 복원에 적합한 정보가 출력되도록 전처리 작업을 수행하였다. 이 전처리 단계는 문장, 단락, 형태소 분석, 개체명 정보 생성을 하는 단계이며, 필요 시 보완이나 수정도 가능하다. 이 단계를 수행함으로써 작업자가 주격 무형 대용어 복원 작업에만 집중할 수 있도록 하였다.

1) 서술어 자동 탐지

형태 분석 기반으로 서술어를 탐지하여, 대상 서술어가 될 수 있는 서술어 후보들을 시각화하여 제공하는 단계이다.

형태분석 및 구문-의미역 기반 서술어 탐지			
작업 도구 사용 절차 위치			
1	1	SSO NNG XSN JKG NNG SYNG NNG NNG JC NNG SE NNP NNG XSV EC MAG NNG SSC "과학자 ¹ 를 ² 위 ³ 귀국 - 귀국 기관 과 ⁴ 교류 - 노벨상 ⁵ 배출 하려면 반드시 필요"	무형대응어 복원
2	1	SSO NNP XSN NNG JKO NNG XSV EC NNG JKO VV EC NNG JC NNG JKO NNG XSV ETM NNG JKS NNG XSV EF SF "노벨 상 수상자 ¹ 를 배출 하려면 국경을 뛰어넘어 ² 인지와 연구비를 확보 하려는 노력 이 필요 하 니다." "노벨상 수상자를 배출하려면 국경을 뛰어넘어 인지와 연구비를 확보하려는 노력이 필요합니다."	배출하려면 x 확보하려는 x 필요하는 x 필요합니다.
	2	MAJ NNG JKB NNP JKO VV ETM NNG JKO VV ETM NNB JX NNG VCP EF SF SSC "다만 과학자 ¹ 에게 노벨상 을 받으라는 압력 을 넣 는 것 은 글을 이 니다 . "	받는 x
	1	NNG NNP NNG NNG JKO VV ETM NNP NNP NNG JKG NNP NNP NNG SSO SN SSC JX SN NNB C NNP NNP NNP 지난해 노벨 화학 상 을 받은 이스라엘 바이오파인 연구소 의 아다 요나트 박사 (71) 는 17 일 서울 강남구 삼성동	받은 x 해아 x
	2	NNG JKB VV ETM SY SYNG NNG JKB VV EC SY NNG JKB NNP JKB NNP NNG JKS VV EC MAG VV EC VV EC 코엑스에서 열린 ' 2023 세계 석학 에게 묻는다 ' 간담회에서 한국에서 노벨상 수상자 가 나오려면 어떻게 하 아야 하 는지	받은 x 열린 x 찾았다.
3	1	VV ETM NNG JKB MM MAG NNG XSV EF EF SF 문 는 질문 에 이 같이 답 하 였다 . 문는 질문에 이같이 답했다.	자문 탐지된 서술어 후보 시각화 자문 탐지된 서술어 후보 목록화 임정남 x

2) 서술어 추가

시각화된 서술어 목록을 통해 작업자는 누락된 서술어를 확인하여 추가할 수 있다.

직관적인 추가 방법을 사용하므로 작업자가 선택한 서술어는 바로 도구에 반영되어 기존 선택된 서술어와 함께 시각화된다.

직관적 화면 구성을 통한 서술어 추가

작업 도구 사용 절차 위치

1	1	"과학자들의 귀국·외국기관과 교류... 노벨상 배출하려면 반드시 필요"
2	1	"노벨상 수상자를 배출하려면 국경을 뛰어넘어 인재와 연구비를 확보하려는 노력이 필요합니다."
	2	다만 과학자에게 노벨상을 받으려는 압력을 낮추는 것은 금물입니다."
3	1	지난해 노벨 화학상을 받은 이스라엘 바이츠만연구소의 아다 요나트 박사(71)는 17일 서울 강남구 삼성동 코엑스에서 열린 '세계 석학에게 묻는다' 간담회에서 한국에서 노벨상 수상자가 나오려면 어떻게 해야 하는지 묻는 질문에 이같이 답했다.
	2	요나트 박사는 생화학분자생물학회 주최로 열린 연례 국제학술대회에 참석하기 위해 한국을 찾았다 .
4	1	요나트 박사는 "이스라엘은 외국에서 성공적으로 연구하고 있는 과학자를 돌아오게 하려고 엄청난 노력을 투자 하고 있다"고 말했다. "인재가 귀국하면 다른 나라의 연구 동향을 파악하는 것은 물론 문제를 해결하는 방식도 습득할 수 있게 된다"고 말했다.
5	1	해외 인재의 귀국은 국제적인 연구 교류를 위해 선(善)순환 구조를 만드는 데도 도움이 된다.
	2	외국 석학들과 함께 연구한 과학자가 고국에 돌아와 후학을 양성하면 자연스레 해외와 교류할 기회가 주어진다.

서술어 탐색 무한대용어 복원

문단	문장	서술어	삭제
2	2	낮추는	x
3	1	받은	x
3	1	열린	x
3	1	해야	x
3	1	묻는	x
3	2	열린	x
3	2	찾았다.	x
4	1	연구하고	x
4	1	엄청난	x
4	1	한다	x
4	1	파악하는	x
4	1	해결하는	x
4	1	습득할	x
5	2	연구한	x

누락된 서술어 추가 (Drag)

3) 서술어 삭제

탐지된 서술어 중 선행어가 이미 있거나 작업 대상이 아닌 서술어 후보는 마우스를 한 번 눌러서 해제할 수 있다. 해제된 서술어는 추가할 때와 마찬가지로 구축 도구에 바로 시각화되어 반영된다.

클릭 한 번으로 목록 삭제

작업 도구 사용 절차 위치

1	SSO NNG XSN JKG NNG SY NNG NNG JC NNG SE NNP NNG XSV EC MAG NNG SSC "과학자들의 귀국·외국기관과 교류... 노벨상 배출하려면 반드시 필요 "
1	SSO NNP XSN NNG JKO NNG XSV EC NNG JKO VV EC NNG JC NNG XSV ETM NNG JKS NNG XSV EF SF "노벨상 수상자를 배출하려면 국경을 뛰어넘어 인재와 연구비를 확보하려는 노력이 필요합니다."
2	MAJ NNG JKB NNP JKO VV ETM NNG JKO VV ETM NNB JX NNG VCP EF SF SSC 다만 과학자에게 노벨상을 받으려는 압력을 낮추는 것은 금물입니다."
1	NNG NNP NNG NNG JKO VV ETM NNP NNP NNG JKG NNP NNP NNG SSO SN SSC JX SN NNBC NNP NNP NNP 지난해 노벨 화학상을 받은 이스라엘 바이츠만연구소의 아다 요나트 박사(71)는 17일 서울 강남구 삼성동 코엑스에서 열린 '세계 석학에게 묻는다' 간담회에서 한국에서 노벨상 수상자가 나오려면 어떻게 해야 하는지 VV ETM NNG JKB MM MAG NNG XSV EF EF SF 묻는 질문에 이같이 답했다. 묻는 질문에 이같이 답했다.

서술어 후보 목록 제거 (Click)

서술어 탐색 무한대용어 복원

문단	문장	서술어	삭제
1	1	배출하려면	x
2	1	배출하려면	x
2	1	필요	x
2	2	확보하려는	x
2	2	받으려는	x
2	2	낮추는	x
3	1	받은	x
3	1	열린	x
3	1	해야	x
3	1	묻는	x
3	2	열린	x
3	2	찾았다.	x
4	1	연구하고	x
4	1	엄청난	x
4	1	한다	x
4	1	파악하는	x
4	1	해결하는	x

3.4.3. 복원 작업 단계

1) 무형 대용어 복원 작업 전환

서술어를 탐색하는 단계가 끝나면 작업자는 본격적으로 무형 대용어를 복원하는 작업으로 들어간다. 선행어를 선택하고 복원할 수 있는 ‘무형 대용어 복원’ 작업창으로 전환한다.

선행어 선택, 복원을 위한 작업창으로 전환

작업 도구 사용 절차 위치

번호	선행어	내용
1	1	"과학자들의 귀국-외국기관과 교류... 노벨상 <u>배출하려면</u> 반드시 필요"
2	1	"노벨상 수상자를 <u>배출하려면</u> 국경을 <u>뛰어넘어</u> 인재와 연구비를 <u>확보하려는</u> 노력이 필요합니다.
	2	다만 과학자에게 노벨상을 <u>받으려는</u> 압력을 <u>넣는</u> 것은 금물입니다."
3	1	지난해 노벨 화학상을 <u>받은</u> 이스라엘 바이츠만연구소의 아다 요나트 박사(71)는 17일 서울 강남구 삼성동 코엑스에서 <u>열린</u> '세계 석학에게 묻는다' 간담회에서 한국에서 노벨상 수상자가 나오려면 어떻게 <u>해야</u> 하는지 묻는 질문에 이같이 답했다.
	2	요나트 박사는 생화학분자생물학회 주최로 <u>열린</u> 연례 국제학술회에 참석하기 위해 한국을 <u>찾았다</u> .
4	1	요나트 박사는 "이스라엘은 외국에서 성공적으로 <u>연구하고</u> 있는 과학자를 돌아오게 하려고 <u>엄청난</u> 노력을 <u>한다</u> "며 "해외 인재가 귀국하면 다른 나라의 연구 동향을 <u>파악하는</u> 것은 물론 문제를 <u>해결하는</u> 방식도 <u>습득할</u> 수 있게 된다"고 말했다.
1	1	해외 인재의 귀국은 국제적인 연구 교류를 위해 선(善)순환 구조를 만드는 데도 도움이 된다.

복원 작업 화면 전환

무형대용어 복원

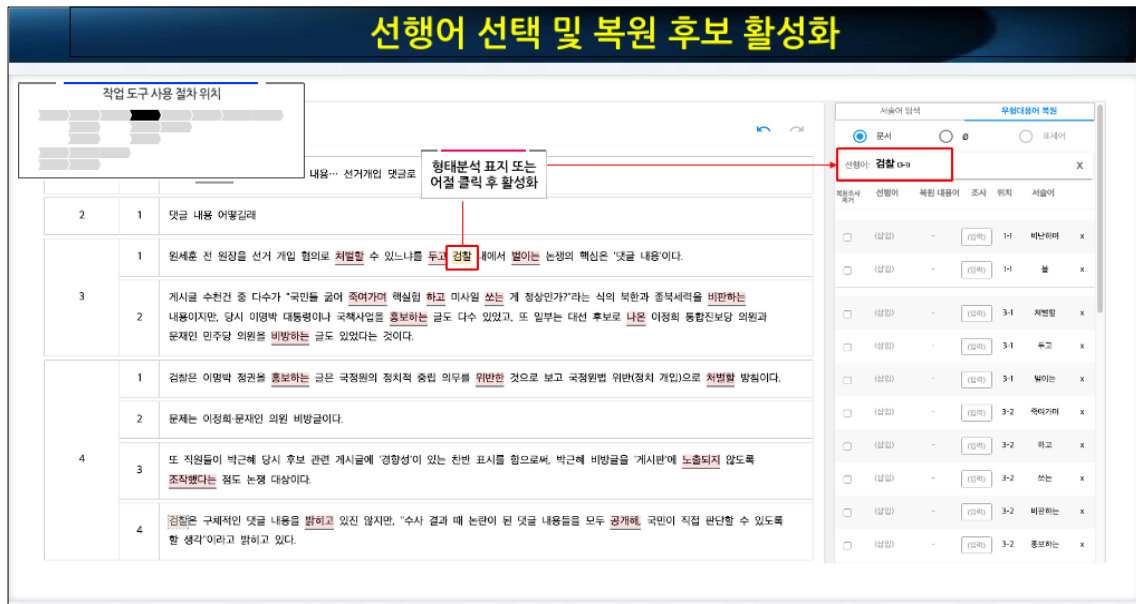
선행어: X

복원 대용어

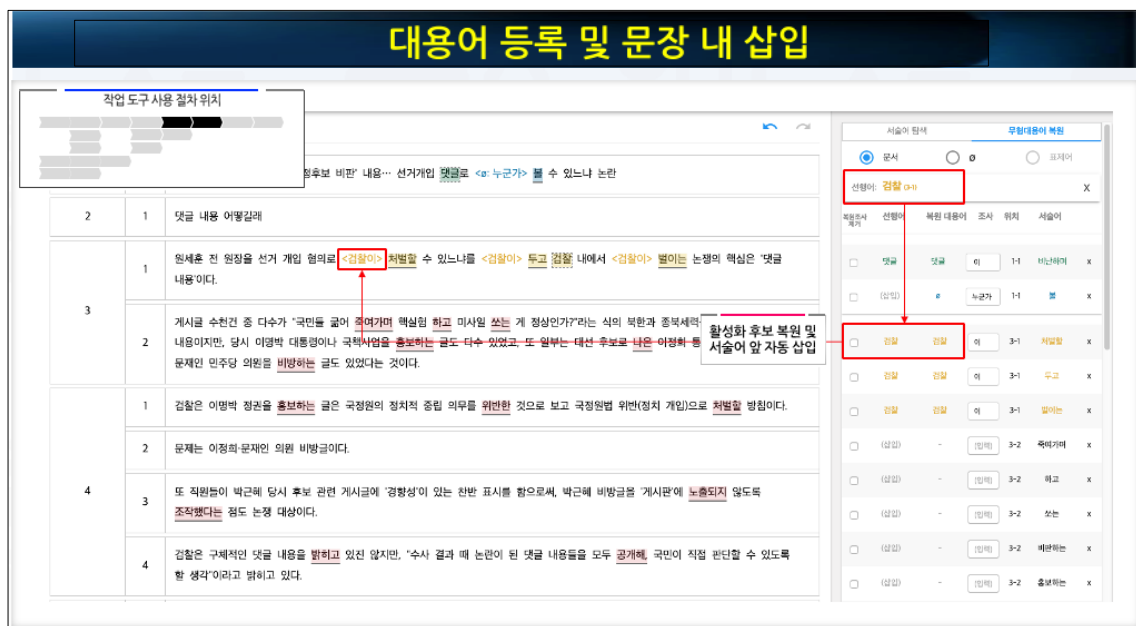
복원 대용어	복원 대용어	복원 대용어	복원 대용어	복원 대용어	복원 대용어	복원 대용어	복원 대용어	복원 대용어	복원 대용어
<input type="checkbox"/> (선행어)	-	(입력)	1-1	배출하려면	x				
<input type="checkbox"/> (선행어)	-	(입력)	2-1	배출하려면	x				
<input type="checkbox"/> (선행어)	-	(입력)	2-1	뛰어넘어	x				
<input type="checkbox"/> (선행어)	-	(입력)	2-1	확보하려는	x				
<input type="checkbox"/> (선행어)	-	(입력)	2-2	받으려는	x				
<input type="checkbox"/> (선행어)	-	(입력)	2-2	넣는	x				
<input type="checkbox"/> (선행어)	-	(입력)	3-1	받은	x				
<input type="checkbox"/> (선행어)	-	(입력)	3-1	열린	x				

2) 복원 대용어 등록 및 문장 삽입

복원할 대용어를 본문 창에서 선택하면 오른쪽 상단 선행어 입력 창에 선행어 후보가 활성화된다.



계속해서 왼쪽 작업창 목록 중 무형 대용어를 삽입할 위치를 선택하면, 클릭한 위치에 활성화된 선행어 후보가 자동 삽입되어 대상 서술어의 선행어가 된다.



3) 복원 대용어의 조사 편집

삽입된 문장의 문법에 맞게 복원 대용어의 조사를 편집할 수 있다.

삽입된 문장의 문법에 맞게 복원 대용어 조사 편집

작업 도구 사용 절차 위치

1	1	스트로스칸 "정복력"이었지만 "도덕적 잘못"
2	1	사건 이후 첫 인터뷰
3	1	경선출마 "노 코멘트"
4	1	도미니크 스트로스칸 국제통화기금(IMF) 전 총재가 호텔 여종업원과의 성관계 혐의와 관련해 "공격이나 강제는 없었다"면서도 "도덕적 잘못"이라고 <총재가> 인정했다.
	2	그는 "대통령 선거 출마를 <그가> 원했지만, 현재의 사회당 후보 경선에 대해서는 <그가> 언급하지 않겠다"고 밝혔다.
5	1	스트로스칸은 이번 사건 뒤 처음으로 18일 오후 8시(현지 시간), 프랑스 사영 방송인 TF1 텔레비전과 인터뷰했다.
	2	이 자리에서 그는 "호텔 여종업원이 모든 것에 대해 거짓말을 했다"고 말했다.
	3	그는 자신을 <검사> 기소했다가 기소를 취소한 사이러스 밴스 <검사의> 보고서를 인용하며, 왜 자신에 대한 기소가 취소됐는지를 설명했다.

서울어 탐색 무협대용어 복원

2 문서 0 표제어

선행어: 그 14-2

복원 대용어 조사 위치

총재 (입력) 4-1 인정했다. x

그 2-2 원했지만. x

그 2-2 언급하지 x

의미에 맞게 조사 편집

5-2 말했다. x

5-3 기소했다가 x

5-4 은 x

5-4 강요한 x

5-4 주장했다. x

4) 부정 주어 및 문서 외 선행어 복원 지원

일반적인 주어는 흔히 생략되는 한국어의 특성을 고려해 부정 주어 및 문서 외 선행어로 '누구나, 무언가'를 선택할 수 있다.

부정주어, 문서 외 선행어 복원 지원

작업 도구 사용 절차 위치

2	1	"청사안 관저 <국가정보원장이> 두고 노출 위험 곳 <국가정보원장이> 거주" <누군가> 지적
3	1	원세훈 국가정보원장이 서울 내국동 국정원 청사 내 관저 대신에 도곡동 티워밸리스 옆 18층짜리 오피스의 일부를 개조해 관저로 <국가정보원장이> 사용하고 있는 것으로 <누군가> 드러났다.
4	1	18일 여권 관계자와 국정원의 말을 <누군가> 종합하면 원 <원장>은 지난해 7월께 이 빌딩의 한 층(248평)을 개조해 가족이 함께 <원장이> 살고 있다.
	2	이 빌딩은 국정원 산하 국가안보전략연구소 소유로 연구소가 12~18층을 쓰고 있으며, 티워밸리스와 대림이크로타운 등이 들어 있는 호화 단지 안에 자리잡고 있다.
	3	정치권에선 원 원장이 외부 접근이 쉽지 않은 내국동 관저보다 스포츠센터와 쇼핑시설 등 각종 근린시설이 많은 도곡동에 살고 싶어했던 것으로 보고 있다.
	4	여권 관계자는 "동선이 공개되지 않아야 할 국정원장이 이렇게 외부에 노출되기 <장소가> 위험 장소로 거처를 <국정원장이> 옮겼다는 건 <문제가> 큰 문제"라고 <관계자가> 말했다.
5	1	이 빌딩은 업무시설 및 근린생활시설로 등록돼 있어 주거용 시설을 <누군가> 짓기 위해선 용도 변경을 <누군가> 해야

서울어 탐색 무협대용어 복원

2 문서 0 표제어

선행어: 누구 14-2

복원 대용어 조사 위치

국가정보원장 2-1 두고 x

국가정보원장 2-1 거주 x

국가정보원장 2-1 지적 x

국가정보원장 3-1 사용하고 x

국가정보원장 3-1 드러났다. x

누군가 4-1 종합하면. x

원장 4-1 살고 x

장소 4-4 위험 x

국정원장 4-4 옮겼다는 x

문제 4-4 큰 x

5) 복수의 복원 대용어 적용

선행어 후보를 한번 등록하면 복원 시 여러 번 적용할 수 있게 해 최소의 작업으로 최대의 효율을 도모하였다.

1회 선행어 등록으로 복수의 복원 대용어 적용 가능(작업 최소화)

작업 도구 사용 절차 위치

선행어 1회 등록으로 여러번 복원 적용

선행어 등록: 그 (4-2)

3 그 그

5 그 그

7 그 그

8 그

그는 대통령 선거 출마를 <그가> 원했지만 현재의 사회당 후보 경선에 대해서는 <그가> 언급하지 않겠다고 <그가> 밝혔다.

그는 자신을 기소했다가 기소를 취소한 사이러스 벤스 법사의 보고서를 인용하며, 왜 자신에 대한 기소가 취소됐는지를 설명했다.

스트로스칸은 이번 사건 뒤 처음으로 18일 오후 8시(현지 시각), 프랑스 사영 방송인 TF1 텔레비전과 인터뷰했다.

이 자리에서 그는 "호텔 여중업원이 오든 것에 대해 거짓말을 했다"고 말했다.

그는 자신을 기소했다가 기소를 취소한 사이러스 벤스 법사의 보고서를 인용하며, 왜 자신에 대한 기소가 취소됐는지를 설명했다.

서울어 탐색: 무형대용어 복원

선행어: 그 (4-2)

복원 대용어: 조사 위치 서울어

총재 총재 가 4-1 인정했다. x

3 그 그 4-2 원했지만. x

5 그 그 4-2 언급하지 x

7 그 그 4-2 밝혔다. x

(상립) - (법학) 5-2 말했다. x

(상립) - (법학) 5-3 기소했다. x

(상립) - (법학) 5-3 취소한 x

(상립) - (법학) 5-3 설명했다. x

(상립) - (법학) 5-4 온 x

3.4.4. 자가 진단 단계

1) 보류

작업자가 구축 작업을 진행할 때 작업 내용에 대한 의문이 생기거나 지침을 확인하고 싶을 때 작업자는 해당 작업 문장의 어절에 대해 보류를 설정할 수 있다. 작업자는 지침 확인 후 보류 설정된 문장에 대해 작업을 재개할 수 있다.

작업 모호시 지침 확인 후 재작업 가능

작업 도구 사용 절차 위치

보류 (1)

2 "노벨상 수상자를 <국가가> 배출하려면 국경을 <국가가> 뛰어넘어 인제와 연계를 <국가가> 확보하려면 노력이 필요합니다.

다만 과학자에게 노벨상을 <과학자가> 받으라는 압력을 <국가가> 넣는 것은 균열입니다."

3 지난해 노벨 화학상을 <박사(71)> 받은 이소리엘 바이츠만연구소의 이다 요나트 박사(71)는 17일 서울 강남구 삼성동 코엑스에서 <간담회> 열린 '세계 석학에게 묻는다' 컨퍼런스에서 한국에서 노벨상 수상자가 나오려면 어떻게 <한국> 해야 하는지 <한국> 묻는 질문에 이렇게 답했다.

4 요나트 박사는 생화학분자생물학회 주최로 <국제학술대회> 열린 연례 국제학술대회에 참석하기 위해 한국을 <박사가> 찾았다.

5 요나트 박사는 "이스라엘은 외국에서 성공적으로 <과학자> 연구하고 있는 과학자를 돌아오게 하려고 <노력> 엄청난 노력을 <이스라엘> 한다"며 "해외 인재가 귀국하면 다른 나라의 연구 동향을 <인재> 파악하는 것은 물론 문제를 <인재> 해결하는 방식도 <인재> 습득할 수 있게 된다"고 말했다.

6 해외 인재의 귀국은 국제적인 연구 교류를 위해 선(先)순환 구조를 만드는 데도 도움이 된다.

7 외국 석학들과 함께 <과학자> 연구한 과학자가 고국에 돌아와 후학을 <과학자> 양성하면 자연스레 해외와 <과학자> 교류할 기회가 주어진다.

서울어 탐색: 무형대용어 복원

선행어: 그 (4-2)

복원 대용어: 조사 위치 서울어

총재 총재 가 4-1 인정했다. x

3 그 그 4-2 원했지만. x

5 그 그 4-2 언급하지 x

7 그 그 4-2 밝혔다. x

(상립) - (법학) 5-2 말했다. x

(상립) - (법학) 5-3 기소했다. x

(상립) - (법학) 5-3 취소한 x

(상립) - (법학) 5-3 설명했다. x

(상립) - (법학) 5-4 온 x

2) 검토 요청

작업자가 구축 작업 진행 시 판단 불가한 예외 상황이 발생할 경우 작업자는 상위
검수자에게 해당 문장에 대해 검토를 요청할 수 있다.

3) 자가 점검

작업자가 작업을 완료할 때, 작업 시 보류한 문장과 검토 요청해 놓은 문장에 대해 최종 자가 점검을 할 수 있도록 하여 작업의 정확도를 높였다.

작업 도구 사용 절차 위치

1. 작업 도구 사용 절차 위치

작업 보류, 검토 내역 별도 자가 점검

교류... 노벨상 <국가>가 **배출하려면** 반드시 필요"

배출하려면 국경을 <국가>가 **뛰어넘어** 인재와 연구비를 <국가>가 **확보하려면** 노력이 필요합니다.

No	구분	내용
2		다만 과학자 에게 노벨상을 <과학자>가 받으려 는 압력을 <국가>가 붙 는 것은 금물입니다."
3	1	지난해 노벨 화학상을 <박사(71)>가 받은 이스라엘 바이츠만연구소의 아다 요나트 박사(71) 는 17일 서울 강남구 삼성동 코엑스에서 <간담회>가 열린 '세계 석학에 묻는다' 간담회 에서 한국 에서 노벨상 수상자가 나오려면 어떻게 <한국>가 해야 하는지 <한국>가 물 는 질문에 이같이 답했다.
	2	요나트 박사 는 생화학박사생물학회 주최로 <국제학술대회>가 열린 연례 국제학술대회 에 참석하기 위해 한국을 <박사>가 찾았다 .
4	1	요나트 박사 는 '이스라엘은 외국에서 성공적으로 <과학자> 연구하고 있는 과학자 를 돌아오게 하려고 <노력>이 엄청난 노력'을 <이스라엘>이 한다 며 '해와 인제가 귀국하면 다른 나라의 연구 동향을 <인제>가 파악 하는 것은 물론 문제를 <인제>가 해결 하는 방식도 <인제>가 습득 할 수 있게 된다'고 말했다.
5	1	해와 인제의 귀국은 국제적인 연구 교류를 위해 선(善)순환 구조를 만드는 데도 도움이 된다.
	2	외국 석학들과 함께 <과학자>가 연구한 과학자 가 고국에 돌아와 후학을 <과학자>가 양성 하면 자연스레 해외와 <과학자>가 교류 할 기회가 주어진다.
	3	이런 교류를 통해 <과학도>들이 젊은 과학도 들이 해외로 진출해 <한국>가 좋은 연구 환경 에서 <과학도>들이 공부 할 수 있는 가능성도 높아진다는 것이다.

분류		검토 요청 (2)	
서울이 발의		무함마드가 확인	
백종우 씨가	선행이	복합 대응이	조사 위치 서울이
<input type="checkbox"/>	국가	국가	가 1-1 배출하려면 x
<input type="checkbox"/>	국가	국가	가 2-1 배출하려면 x

4) 작업 종료 재확인

작업 종료 단계 시 작업 완료 메시지와 함께, 선행어가 복원되지 않은 서술어 및 미작업 내역을 확인하게 함으로써 작업 오류를 최소화했다.

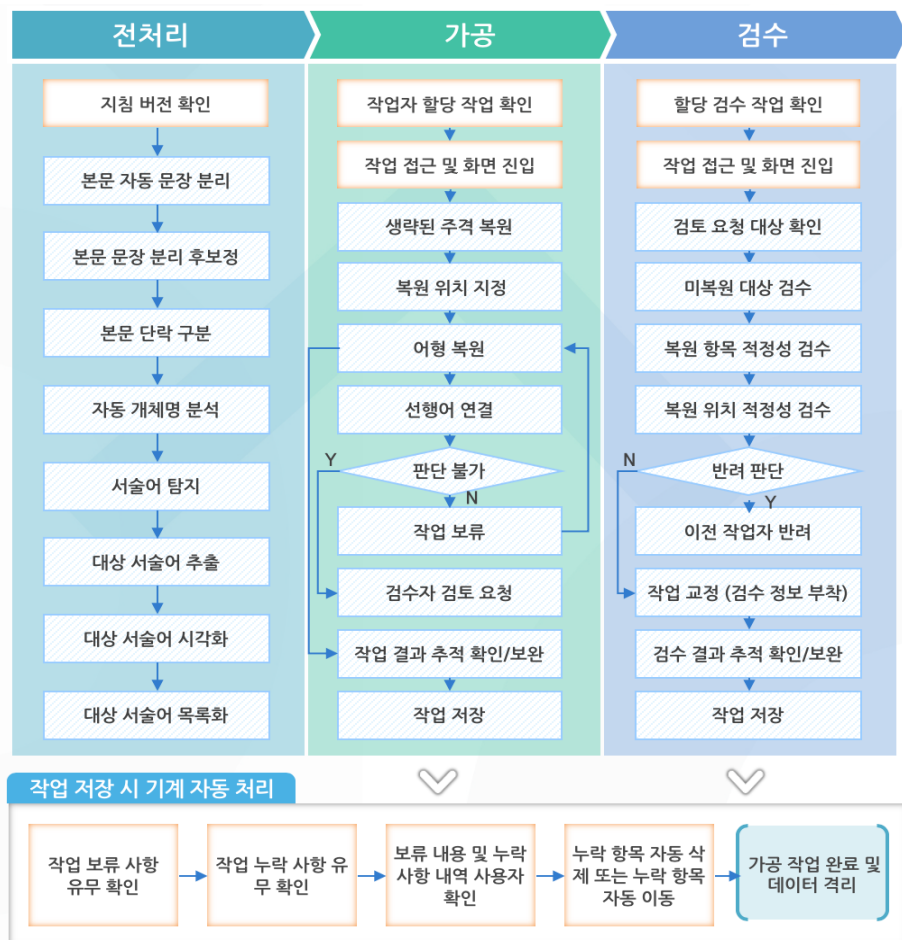


4. 말뭉치 구축 및 납품

4.1. 말뭉치 구축

4.1.1. 말뭉치 구축 절차

말뭉치의 구축을 위해 원본 데이터 형태 분석, 개체명 분석 및 문장, 문단 등 본문 주석을 자동으로 처리하고, 주격 무형 대용어 복원을 위한 서술어를 추출한다. 이와 같은 전처리 작업을 통해 보다 정확하고, 효율적인 가공 작업을 지원하고, 작업 중 작업 오류 및 검토 요청을 통해 부정확한 가공을 사전 배제하며, 작업 완료 시 자동 검사를 통해 미복원 선행어나 누락된 서술어가 없는지 작업 내용을 재확인 후 작업을 완료하는 절차로 진행한다.



〈주격 무형 대용어 복원 말뭉치 가공 절차도〉

4.1.2. 말뭉치 특성에 따른 구축

한국어는 필수격의 생략이 빈번하고, 어순이 비교적 자유롭다. 주격 무형 대용어 복원에서는 이런 한국어의 특성과 어감을 최대한 살리기 위해 전산 언어적 활용과 문법적 적합성을 모두 고려했다.

초기의 구축 도구 작업 시 한국어의 특성에 맞는 말뭉치 활용도를 높이기 위해, 무형 대용어(4가지 형태) 복원 시스템을 보완하였으며, 구어 말뭉치의 발화자 정보를 반영하였다.

The screenshot displays the KRONOTH annotation tool interface. The main window shows a text corpus with morphological annotations. The side panel, titled '무형대용어 복원' (Untyped Pronoun Restoration), lists various pronouns and their corresponding untyped pronouns for restoration. The interface includes a search bar, a list of pronouns, and a table for selecting the appropriate untyped pronoun for each instance.

복원조 사제거	선행어	복원 대용어	조사 위치	서술어
<input type="checkbox"/> 씨	씨	가	11-3	나온다
<input type="checkbox"/> 능	능	이	11-4	못한
<input type="checkbox"/> 제	제	(입력) 누군가 아무나 아무것	11-4	없는
<input type="checkbox"/> (삽입)	0	무언가	11-4	아닙니까
<input type="checkbox"/> 부부	부부	가	12-1	있고
<input type="checkbox"/> 부부	부부	가	12-1	있어

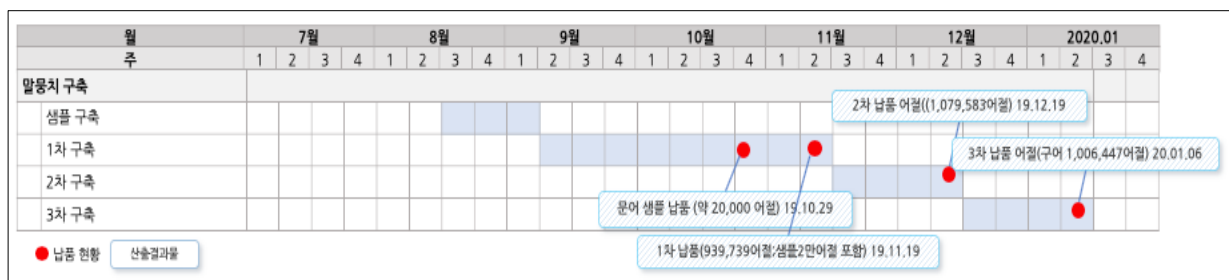
〈무형 대용어의 복원 - 크로노스(KRONOTH) 주석 시스템〉

4.1.3. 말뭉치 구축 기간

월	7월				8월				9월				10월				11월				12월				2020.01			
주	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
말뭉치 구축																												
샘플 구축																												
1차 구축																												
2차 구축																												
3차 구축																												

- 샘플 구축 기간은 내부 크로노스(KRONOTH) 주식 시스템의 구성과 설계를 위한 기간으로, 2019년 8월 19일부터 2019년 8월 27일까지 샘플 문서로 말뭉치를 구축하였다.
- 1차 구축 기간은 2019년 9월 9일부터 2019년 11월 15일까지이며, 시범 검증 문서를 포함하여 문어 939,739어절(3,406개 문서)을 구축하였다.
- 2차 구축 기간은 2019년 11월 16일부터 2019년 12월 18일까지이며, 문어 1,079,583어절(3,859개 문서)을 구축하였다.
- 3차 구축 기간은 2019년 12월 19일부터 2020년 1월 3일까지이며, 구어 시범 검증 20,342어절(55개 문서)과 구어 1,006,447어절(423개 문서)을 구축하였다.

4.2. 말뭉치 납품



4.2.1. 납품 어절 및 문서 수

- 주격 무형 대용어 복원 말뭉치 결과물로 납품된 총 어절은 3,025,769개(7,688개 문서)이며, 이 가운데 문어는 2,019,322어절(7,265개 문서), 구어는 1,006,447어절(423개 문서)이다.
- 납품은 총 3차로 이루어졌으며, 1차 납품 2019년 11월 19일, 2차 납품 2019년 12월 19일, 3차 납품 2020년 1월 6일이다.
- 1차 납품 시 문어 시범 검증 문서를 포함한 문어 939,739어절(3,406개 문서), 2차 납품 시 문어 1,079,583어절(3,859개 문서), 3차 납품 시 구어 시범 검증 문서를 포함한 전체 1,006,447어절(423개 문서)을 결과물로 납품하였다.

자료 유형	차수	어절	문서 수(개)	구축기간	납품일
문어	1차 (시범 포함)	939,739	3,406	19.9.9~19.11.15	19.11.19
	2차	1,079,583	3,859	19.11.16~19.12.18	19.12.19
	문어 총계	2,019,322	7,265		
구어	3차	20,342	55	19.12.19~20.1.3	20.1.6
		986,105	368		
	구어 총계	1,006,447	423		
총계		3,025,769	7,688		

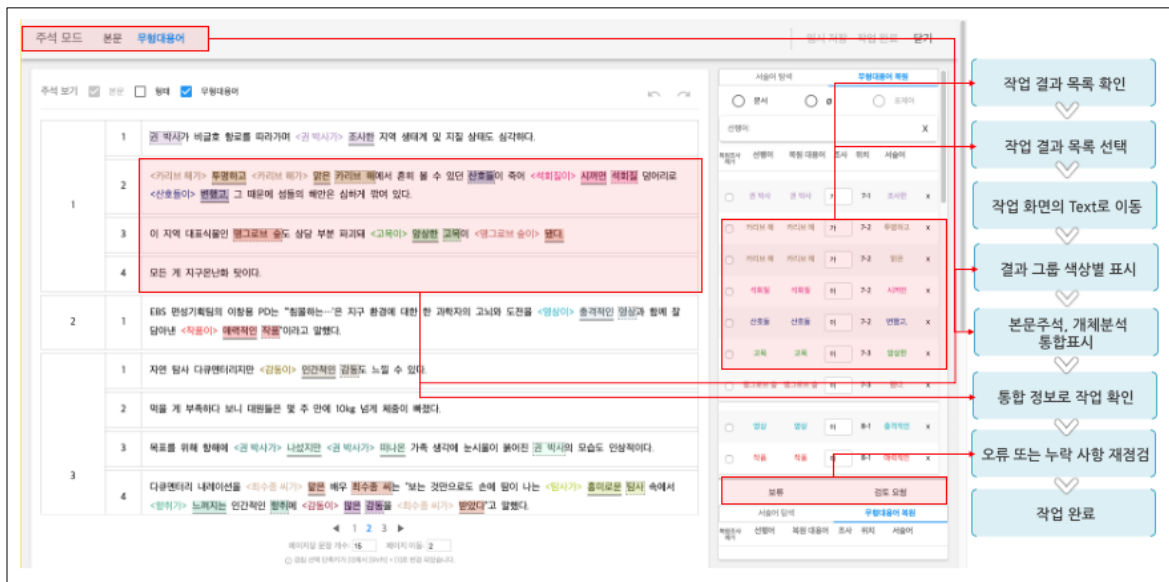
<주격 무형 대응어 자료 유형별 납품 현황>

5. 검증 및 산출물 보고

5.1. 내부 검증

5.1.1. 작업자 검증

작업자가 작업을 완료하기 전까지는 자신의 작업을 수정할 수 있다. 작업 창에서 오류 부분을 마우스로 누르면 편집 목록으로 이동된다. 작업 목록에서 항목을 선택하면 해당 문장의 위치로 자동 이동하여 작업자 스스로 직관적인 검증을 하도록 하였다.



〈작업자 작업 및 검증 화면 - 크로노스(KRONOTH) 주석 시스템〉

5.1.2. 기계적 검증

분석 말뭉치 구축을 위해 부착하는 주석 정보는 문서에 따라 주석의 양이 과해지면 작업 내용에 대한 추적이 어려워진다. 따라서 효율적인 구축 작업을 위해서 작업한 대상의 위치를 쉽게 추적 가능하도록 하는 작업과 동시에, 작업 완료 시 문서 내 미작업 대상에 대한 자동 검수를 통해 분석 대상 누락을 빠르게 점검하여 구축 작업의 효율성을 높이는 동시에 작업 품질을 높이는 역할을 하였다.



〈미작업 내용 및 서술어 누락에 대한 검증 - 크로노스(KRONOTH) 주석 시스템〉

5.1.3. 절차적 검증

검수 단계의 검증은 작업자 결과물에 대해 지침 준수 여부와 작업 결과물 사이의 불일치를 검출하는 단계이다. 단순 오류 검출뿐만 아니라 다수의 작업자 간 지침 해석이 다른 불일치 대상에 대한 처리 방침을 최종 결정하여 교정 작업이 이루어진다. 이 과정에서 데이터 변경, 작업 이력에 대한 통계 관리로 지침과 말뭉치의 품질을 개선하였다.

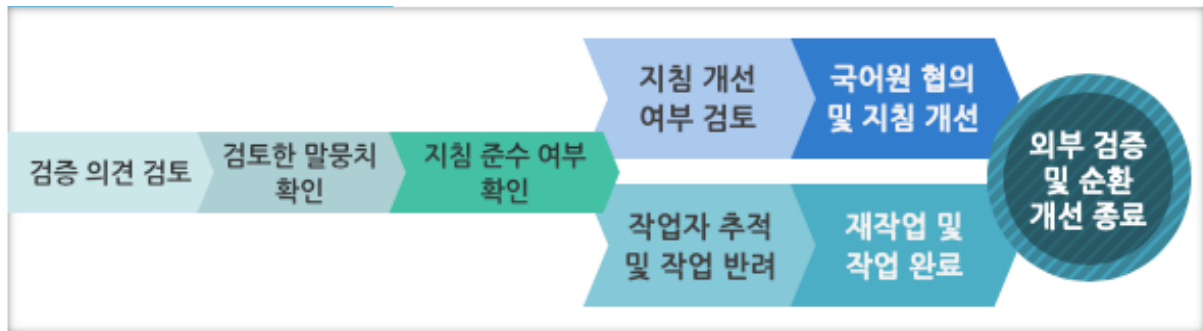
작업자는 작업 과정 중 지침이 모호하거나 적용하기 어려운 작업을 검수자에게 검토를 요청하고, 검수자는 작업자의 검토 요청 데이터에 주석을 부착하여 저장한다. 이 단계에서 완료된 데이터를 격리 저장하고 단계별로 역할을 부여하여 철저히 검증했다.

5.1.4. 관리적 검증

이 단계에서는 관리자가 작업의 전반적인 현황을 관리했다. 작업자의 작업 내용과 현황을 확인하거나 동일 시간 내 작업자별 수행량, 작업자별 검토 요청, 검수자별 수정 현황, 작업자와 검수자 작업 보류 이력 사항 등을 확인할 수 있다. 이런 과정을 통해 작업의 품질과 작업자의 수행 성과를 관리하고, 지침 보완 사항을 파악해 작업의 반력 및 지침 개선에 적용했다.

5.1.5. 활용성 검증

구축 도구 시스템에 저장된 데이터에 대하여 발주 기관과 협의하고, 데이터의 활용 및 평가 검증을 통해 객관적인 활용 의견을 수합하고, 해당 의견에 대한 지침의 검토와 구축된 말뭉치에 대한 검토를 통해 지침 개선이나 말뭉치 재작업 등의 순환 개선을 실시하였다.



<의견 검토 및 순환 개선 절차도>

5.2. 외부 검증

주격 무형 대용어 복원 문서의 품질을 검증하기 위해 형식 및 내용 오류를 분석하고, 이를 점수로 산출하였다.

5.2.1. 형식 오류 검증

1) 형식 오류 검증 내용

- 말뭉치 형식(json) 검증
- 제약 조건 검사

2) 형식 오류 검증 방법

구축 말뭉치 납품 분량 전수에 대한 형식 오류 검증 오류 유형

- FORMAT_ERROR_ZEROANAPHORA_NULL_ANTECEDENT : 태깅된 서술어의 선행어가 없는 경우
- FORMAT_ERROR_ZEROANAPHORA_WRONG_BEGIN : begin 값이 - 1보다 작은 경우
- FORMAT_ERROR_ZEROANAPHORA_WRONG_ORDER : begin 값이 end 값보다 큰 경우
- 정답 세트 범위 내: 내용 검증 실시
- 정답 세트 범위 외: 일관성 검증 실시(2차 검증부터)

5.2.2. 내용 오류 검증

1) 내용 오류 검증 내용 및 방법

내용 오류 검증에서는 형식 오류 검증을 통과한 구문 분석 말뭉치에 대해 발주 기관의 말뭉치 통합 검증 사업팀이 구축한 정답 세트와의 일치도를 검사하였다. 검사 항목은 복원 대상 서술어 검증과 선행어 검증으로 구분된다. 복원 대상 서술어 검증에서는 복원해야 할 서술어를 누락하지는 않았는지, 복원 대상이 아닌 서술어를 복원하지는 않았는지 여부를 확인하였다. 선행어 일치 검증에서는 복원된 주어의 어형과 문서 내 위치가 일치하는지 확인하였다.

검사 항목	오류 유형
복원 대상 서술어 일치 여부	<ul style="list-style-type: none"> • ZEROANAPHORA_PREDICATE_MISSED (복원 대상인 서술어를 누락) • ZEROANAPHORA_PREDICATE_OVER (복원 대상이 아닌 서술어를 복원)
선행어 일치 여부	<ul style="list-style-type: none"> • ZEROANAPHORA_ANTECEDENT_DIFFERENT (선행어 불일치)

<내용 오류 일치도 검사 항목>

검증 지표는 서술어 일치도와 선행어 일치도를 두루 살펴 산출하였으며, 검증 수식은 아래와 같다.

○ 검증 지표 : 서술어 일치도(F1) * 선행어 일치도(Accuracy)

2) 내용 오류 검증 결과

내용 오류 검증 결과, 문어와 구어의 검증 총점은 각각 0.7999, 0.6811로 나타났다.

	선행어 일치도	서술어 일치도	총점
문어	0.8719	0.9174	0.7999
구어	0.7320	0.9305	0.6811

문어 2,000,213어절(7,265개 문서)을 대상으로 내용을 검증한 결과, 선행어 일치도는 0.8719, 서술어 일치도는 0.9174로 총점은 0.7999였다. 오류 유형의 세부 사항은 아래 표와 같다.

오류 유형	개수	비율
ZEROANAPHORA_ANTECEDENT_DIFFERENT (선행어 불일치)	1999	0.416025
ZEROANAPHORA_PREDICATE_MISSED (복원 대상인 서술어를 누락)	1243	0.258689
ZEROANAPHORA_PREDICATE_OVER (복원 대상이 아닌 서술어를 복원)	1563	0.325286

<문어 문서의 오류 개수 및 비율>

구어 1,011,778어절(423개 문서)을 대상으로 내용을 검증한 결과, 선행어 일치도는 0.7320, 서술어 일치도는 0.9305로 총점은 0.6811이었다. 오류 유형의 세부 사항은 아래 표와 같다.

오류 유형	개수	비율
ZEROANAPHORA_ANTECEDENT_DIFFERENT (선행어 불일치)	3137	0.641907
ZEROANAPHORA_PREDICATE_MISSED (복원 대상인 서술어를 누락)	1140	0.233272
ZEROANAPHORA_PREDICATE_OVER (복원 대상이 아닌 서술어를 복원)	610	0.124821

<구어 문서의 오류 개수 및 비율>

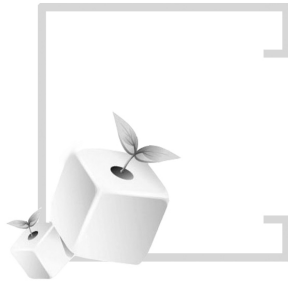
5.3. 산출물

5.3.1. 산출물 납품 형태

주격 무형 대용어 복원 말뭉치(300만 어절), 중간 산출물 등을 저장 매체(휴대용 저장 매체 3개)에 담아 제출하며, 사업 수행 과정 및 결과 요약 보고서 형태의 최종 보고서(인쇄본 30부)를 납품 완료 시(2020년 1월 15일)에 제출한다.

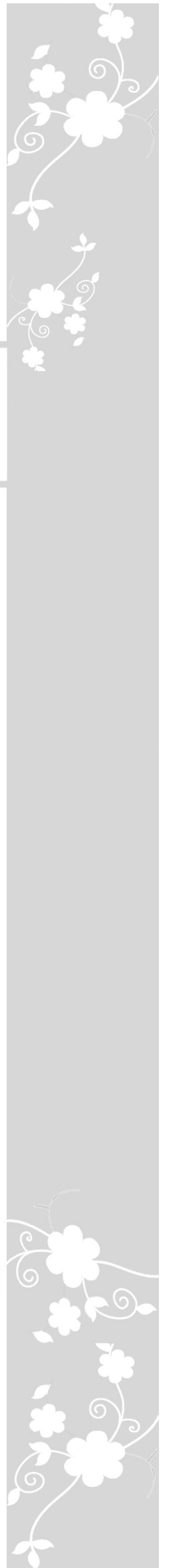
5.4. 사업 보고

- 착수 보고 2019년 8월 30일
- 중간 보고 2019년 12월 3일
- 최종 보고 2020년 1월 14일



제 3 장

향후 계획



1. 개선 방향

본 사업에서는 주격 무형 대용어 복원 분석 말뭉치를 구축하고, 산업계와 언어학계를 아우르며 실현 가능한 지침을 수립하고자 하였다. 그러나 기존의 구축 사례가 거의 없고 구축된 말뭉치에 대한 국어학적인 검토나 연구가 없어 복원 방법과 복원 대상, 범위 등을 정하기가 쉽지 않았다. 또한 본 사업에서는 주격 무형 대용어의 복원만을 고려하였고, 복원에 필요한 구문 분석 정보를 작업자가 직접 판단했기에 구문 분석 결과 정보에 따라 지침과의 불일치로 보이는 사례가 발생하기도 하였다.

그럼에도 불구하고 본 사업의 주격 무형 대용어 복원 말뭉치는 현재까지 구축된 가장 대량의 무형 대용어 복원 말뭉치라는 데에 의의가 있다. 다만, 목적어 등 타 문장 성분에 대해서는 분석하지 않았고, 구문 분석과 형태 분석 정보가 연계되지 않는 등 제한적인 정보가 제공된다는 한계가 있다. 또한 구어 자료는 문장이 아닌 발화 단위를 기본으로 삼았고, 메타 데이터인 실제 발화자가 아닌 선행어를 대상으로 복원하여 구어의 특성을 충분히 반영하지 못한 부분이 있다.

이러한 한계는 이후의 연속적인 사업 진행을 통해 발전될 것으로 기대한다. 또한 본 사업 결과물은 완전한 전자화의 토대를 갖췄으므로 개별 연구자의 연구와 그 성과 공개를 통해서도 보완될 수 있다.

1) 다른 문장 성분 복원으로의 확대

본 사업은 무형 대용어 말뭉치로써는 대규모인 문어 200만 어절, 구어 100만 어절을 구축하기 위해 주어 복원에만 집중했다. 우리말은 주어 외 문장 성분도 맥락에 따라 생략할 수 있으므로 다른 문장 성분으로도 분석이 확대되어야 한다.

이와 함께 주석 체계에 대한 논의를 재고할 필요가 있다. 본 사업보다 먼저 진행된 한국 전자통신연구원(ETRI)에서는 주격, 목적격, 부가격에 대한 복원 말뭉치를 진행하며 아래와 같은 표지 체계를 제시한 바 있다. 본 사업의 대규모 주격 무형 대용어 복원 말뭉치의 복원 양상을 분석하여 표지 체계를 상세화하고 그 적용 범위와 단계 등을 다각적으로 검토해야 한다.

	기호	설명
1	s	주격 생략
2	o	목적격 생략
3	a	부가격 생략
4	m	관형형용언의 주격, 목적격 생략 - ms, mo로 표기
5	b	선행어가 뒤에 존재함

2) 타 층위 정보와의 연계

무형 대용어 복원은 궁극적으로 구문 분석, 개체명, 어휘 의미 분석 등 다른 분석 층위와 연계해야 한다. 의미, 통사, 화용 정보가 연결되는 접점으로 기능하기 위해 지침과 자료 구조를 수정해야 할 것이다.

3) 구어 자료 처리의 개선

본 사업은 대량의 구어 말뭉치에 대해 주격 무형 대용어 복원 말뭉치를 구축했다. 구어 자료에 대한 분석과 활용은 최근에 확대되고 있다. 구어 자료는 문어체와 구어체라는 자료 자체의 차이뿐만 아니라, 메타데이터와 본문 주석(문장, 단락 등의 단위적 정보)과 전사 주석(발화 현상에 대한 주석) 등 자료의 표현과 의미가 문어 자료와 전혀 다르다. 따라서 향후의 개선에서는 문어 데이터와 구어 데이터 처리에 있어 기본적인 일관성은 유지하되 구어 자료의 특성을 고려한 처리 방안 또한 고려되어야 한다.

특히 무형 대용어 복원의 경우, 구어에서 주어가 대화문 내에 드러나지 않는 경우가 빈번하여 발화자를 ‘누군가’로 복원하는 경우가 많았으므로, 보다 정확한 복원을 위해 발화자 층위의 정보 제시 방안이 논의되어야 할 것이다.

4) 복원 정보화의 상세화

본 사업에서 무형 대용어 복원의 구축 결과는 다음과 같이 주격이 생략된 서술어와 그것의 선행어 정보로 구성되어 있다.

```

“predicate” : {
    “sentence_id” : “SRAK00138.2”,
    “form” : “말게“,
    “begin” : 4,
    “end” : 6,
    “word_ids” : [ 2 ] },
“antecedent” : {
    “sentence_id” : “SRAK00138.2”,
    “form” : “이종재“,
    “type” : “subject“,
    “begin” : 9,
    “end” : 12,
    “word_ids” : [ 4 ] }

```

여기에 앞서 언급한 목적격 등 다른 문장 성분의 복원 정보가 포함되는 것은 물론, 복원되는 요소의 삽입 위치, 다양한 선행어의 허용 여부(문서 내·외부 혹은 선·후행), 부정칭의 상세 분석 등 그동안 언어학적 논의에서 다루었던 사항들도 포함될 수 있다. 이러한 정보를 꾸준히 체계화하면 무형 대용어 복원 말뭉치의 활용성도 높아질 것이다. 이 중 일부는 이미 언어학 연구에서 논의되기도 했으나 말뭉치 구축에는 아직 반영되지 않았다. 이번 4차 산업혁명 대비 국어 빅데이터 구축 사업으로 전산 처리 분야의 최신 기술과 경험, 구축 지원 도구 등이 확산되어 향후 논의가 보완되고 확대될 것으로 기대한다.

2. 기대 효과

주격 무형 대용어 복원 말뭉치 구축을 위해, 기본적인 범위, 단위, 대상, 방법을 명확히 제시하고, 작업자가 일관되게 판단할 수 있도록 올바른 사례, 여러 예외 상황에 대한 검토와 실험을 진행하여 지침에 반영하였다. 이러한 지침에 따라 일관되게 작업할 수 있는 도구를 활용하여 주격 무형 대용어 복원 말뭉치를 구축함으로써, 4차 산업 및 언어적 연구에서 즉각적인 전산 처리가 가능한 말뭉치 구축에 이바지하고자 하였다.

무형 대용어 복원 말뭉치 연구는 2005년 이후 일본 이이다류(飯田龍), 이누이켄타로(乾健太郎)의 연구, 중국 양샤오펑(楊曉峰)의 연구를 비롯하여 미국 및 체코의 대용어 복원 연구(Issues in Anaphora Resolution(미국)/Anaphora Resolution(체코)) 등 그 양이 증가하고 있는 추세이다. 그러나 한국어는 격의 생략과 비실현이 빈번해 생략의 범위와 양상이 다양한 언어이므로 다른 언어의 무형 대용어 말뭉치를 그대로 가져와 언어 처리

에 활용하기에는 한계가 있었다. 그러나 본 사업에서 우리말로 된 대량의 무형 대용어 복원 분석 말뭉치를 구축함으로써 국내 대용어 복원 연구의 발판이 마련되었다. 우리말의 대용어 사용 양상이 잘 반영된 언어 자료를 활용하여 관련 연구가 활발히 이루어지기를 기대한다.

또한 본 사업은 구문 분석이나 의미역 등과 같은 구조, 의미 분석 기술의 개발과도 밀접한 관련이 있어 활용성이 높을 것으로 기대한다. 무형 대용어 사용으로 인한 해석의 모호성을 해소하여 구문 분석이나 의미역 분석 결과의 개선에도 기여할 수 있을 것이다.

참고문헌

- 김광희(2011), 대용표현, 『국어학』 60, 국어학회.
- 김영태 외(2018), 신경망모델을 이용한 무형대용어 해결 기법, 『한국정보과학회 학술발표 논문집』 12, 한국정보과학회.
- 류지희 외(2017), 한국어 생략어 복원 가이드라인, 『한글 및 한국어 정보처리 학술대회 논문집』 29, 한국어정보학회.
- 박진호(2007), 유형론적 관점에서 본 한국어 대명사 체계의 특징, 『국어학』 50, 국어학회.
- 임수종 외(2005), 백과사전 질의응답을 위한 생략된 표제어 복원에 대한 연구, 『한국정보과학회 학술발표 논문집』 32-2, 한국정보과학회.
- 황민국 외(2014), 무형대용어 해결 기술을 이용한 백과사전 표제어 복원, 『한글 및 한국어 정보처리 학술대회 논문집』 26, 한국어정보학회.
- 황민국 외(2015), Structural SVM을 이용한 백과사전 문서 내 생략 문장성분 복원, 『지능정보연구』 21-2, 한국지능정보시스템학회.
- Ge, N., Hale, J. and Charniak, E. 1998. A statistical approach to anaphora resolution. *Proceedings of the 6th Workshop on Very Large Corpora*.
- Iida, R., Inui, K., Takamura, H. and Matsumoto, Y. 2003. Incorporating contextual cues in trainable models for coreference resolution. *Proceedings of the 10th EACL Workshop on the Computational Treatment of Anaphora*.
- Iida, R., Inui, K. and Matsumoto, Y. 2005. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processsing(TALIP)*, 4-4.
- Iida, R., Inui, K. and Matsumoto, Y. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing(TALIP)*, 6-4.
- Iida, R. and Poesio, M. 2011. A cross-lingual ILP solution to zero anaphora resolution. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.
- Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M. and Jurafsky, D. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39-4.
- McCarthy, J. F. and Lehnert, W. G. 1995. Using decision trees for coreference resolution. *Proceedings of the 14th International Joint Conference on Artificial Intelligence(IJCAI)*.

- Okumura, M. and Tamura, K. 1996. Zero pronoun resolution in Japanese discourse based on centering theory. *Proceedings of the 16th International Conference on Computational Linguistics(COLING)*.
- Park, C., Choi, K-H., Lee, C. and Lim, S. 2016. Korean coreference resolution with guided mention pair model using the deep learning. *ETRI Journal*, 38-6.
- Yang, X., Zhou, G., Su, J. and Tan, C. L. 2003. Coreference resolution using competition learning approach. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics(ACL)*.
- Yang, X., Su, J. and Tan, C. L. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics(COLING-ACL)*.
- Yang, X., Su, J. and Tan, C. L. 2008. A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34-3.

<Abstract>

Constructing a Subject Zero Anaphora Resolution Corpus

This project aims to construct a zero anaphora resolution corpus in Korean. Zero anaphora resolution makes it possible to clearly identify the entity information in the document. Therefore, zero anaphora resolution clarifies the meaning of the document and improves the consistency of information. The results of resolution can be usefully utilized in fields such as information retrieval and extraction, question and answer, and machine translation. This project focused on constructing a corpus and establishing guidelines for the subject zero anaphora resolution. The major tasks and goals of the project as follows:

To establish guidelines for a subject zero anaphora resolution corpus

: While focusing on maintaining consistency and enhancing efficiency for Natural Language Processing, guidelines have been established so as not to deviate from general linguistics. Spoken language has different properties from written language. A subject, for instance, is unrealized frequently and, even though overt, it tends to be abbreviated. Therefore, an elaborated instructions for spoken language should be described. Besides, various exceptions were reviewed and tested continuously for written and spoken language both, and they were reflected in the entire guidelines for zero anaphora resolution.

To construct a subject zero anaphora resolution corpus

: Based on the guidelines, a subject zero anaphora resolution corpus with a

scale of three million words was constructed: two million words of written language and one million of 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. The subject zero anaphora resolution corpus was constructed through a process of internal and external verification. For the internal verification, a system allowed annotators to correct their own errors and a mechanical system automatically inspected unworked objects. For external verification, corpus form verification and content error verification were conducted. In addition, the annotators's results were compared with an answer set provided by National Institute of Korean Language. Through this process, the quality of the corpus was measured.

The corpus constructed in this project is the largest one for a zero anaphora resolution corpus in Korean, and is expected to contribute to the Fourth Industry Revolution and language research.

Keywords: zero anaphora, zero anaphora resolution, subject zero anaphora resolution, corpus

Project Director: Kwak Yongjin(IIR TECH)

사업 책임자	곽용진((주)이르테크 대표이사)
사업 참여자	이숙의(충남대학교 인문과학연구소 전임연구원) 김진수(충남대학교 국어국문학과 교수) 정해영((주)이르테크 선임연구원) 김정인(충남대학교 국어국문학과 박사 과정) 이정은(충남대학교 국어국문학과 박사 과정) 장지현(충남대학교 국어국문학과 박사 과정) 정민경(충남대학교 국어국문학과 석사 과정) 임보람(충남대학교 국어국문학과 석사 과정) 외 18명
담당 연구원	이승재(국립국어원 언어정보과장) 서셋별(국립국어원 언어정보과 학예연구사)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2020년 1월 14일

발행일: 2020년 1월 14일

인 쇄: 세종기획

※ 이 책은 국립국어원의 용역비로 수행한 ‘주격 무형 대용어 복원 말뭉치 구축’ 사업의 결과물을 발간한 것입니다.