

국립국어원 2019-01-61

발 간 등 록 번 호

11-1371028-000807-01

말뭉치 통합 검증

사업 책임자
최 기 선



제 출 문

국립국어원장 귀하

국립국어원과 체결한 용역 계약에 따라 ‘말뭉치 통합 검증’에 관한 용역 보고서를 작성하여 제출합니다.

■ 사업 기간: 2019년 6월 25일 ~ 2020년 2월 29일

2020년 2월 29일

사업 책임자: 최 기 선 (한국과학기술원)

사업 수행자 한국과학기술원, 나라아이넷(주),
(주)언어과학

사업 책임자 최기선

사업 참여자 강규영, 강소연, 강아름, 김건영, 김건태,
김동환, 김민지, 김선영, 김선주, 김세은,
김소희, 김아름, 김유겸, 김은경, 김주상,
김지성, 김진동, 김태우, 김태인, 김한결,
남상하, 노소은, 노영빈, 목정수, 박석원,
박승희, 박용배, 박지용, 박진호, 박형진,
박혜린, 박혜승, 백채원, 서반석, 손지은,
송상현, 송영숙, 신용남, 심지수, 안기경,
안상민, 연규동, 오규환, 오탈환, 유민애,
이경원, 이민호, 이상희, 이신복, 이의중,
이하은, 이호진, 장고은, 장원철, 장하연,
정용빈, 정유남, 정유성, 정혜린, 최원석,
최윤지, 최진, 최현수, 함영균, 허인영,
허철훈, 홍은영 (총 68명)

<사업 수행자>
 한국과학기술원 공동 수급체
 한국과학기술원, 나라아이넷(주), (주)언어과학

사업 책임자	최기선 (한국과학기술원 전산학부 석좌교수)
사업 참여자	강규영 (서울대학교 국어국문학과 박사수료)
	강소연 (고려대학교 언어학 석사)
	강아름 (고려대학교 언어정보연구소 연구교수)
	김건영 (고려대학교 국어국문학과 박사수료)
	김진태 (한국과학기술원 전산학부 석사과정)
	김동환 (한국과학기술원 정보전자연구소 연구원)
	김민지 (서울대학교 국어국문학과 석사과정)
	김선영 (서울대학교 국어학 박사)
	김선주 (미국 하와이주립대학교 언어학 석사)
	김세은 (한국과학기술원 시맨틱웹첨단연구센터 연구원)
	김소희 (고려대학교 언어학 박사)
	김아름 (서울대학교 국어국문학과 박사수료)
	김유겸 (서울대학교 국어국문학과 박사수료)
	김은경 (대전대학교 정경학부 빅데이터 사이언스 전공 교수)
	김주상 (서울대학교 국어학 박사)
	김지성 (한국과학기술원 전산학부 박사과정)
	김진동 (한국과학기술원 전산학부 겸임교수)
	김태우 (인하대학교 한국학연구소 연구교수)
	김태인 (서울대학교 국어학 박사)

김한결 (서울대학교 국어국문학과 박사과정)

남상하 (한국과학기술원 전산학부 박사수료)

노소은 (고려대학교 언어학 석사)

노영빈 (한국과학기술원 정보전자연구소 연구원)

목정수 (서울시립대학교 국어국문학과 교수)

박석원 (연세대학교 언어정보학 협동과정 석사과정)

박승희 (나라아이넷(주) 전무이사)

박용배 (서울시립대학교 국어학 박사)

박지용 (서울대학교 국어국문학과 박사수료)

박진호 (서울대학교 국어국문학과 교수)

박형진 (가천대학교 한국어문학과 교수)

박혜린 (고려대학교 언어학과 석사수료)

박혜승 (서울대학교 국어국문학과 박사수료)

백채원 (서울대학교 국어학 박사)

서반석 (서울대학교 국어학 박사)

손지은 (고려대학교 국어국문학과 박사수료)

송상헌 (고려대학교 언어학과 교수)

송영숙 (경희대학교 국어국문학과 박사수료)

신용남 (서울대학교 국어국문학과 박사수료)

심지수 (경희대학교 국어국문학과 박사수료)

안기경 (서울시립대학교 국어국문학과 박사수료)

안상민 (한국과학기술원 인공지능연구소 연구원)

연규동 (연세대학교 인문학연구원 교수)

오규환 (동덕여자대학교 국어국문학과 교수)

오태환 (연세대학교 국어국문학과 박사과정)

유민애 (서울대학교 국어교육연구소 객원연구원)

이경원 (고려대학교 언어학과 석사수료)

이민호 (한국과학기술원 전산학 석사)

이상희 (서울시립대학교 국어국문학과 박사과정)

이신복 (서울시립대학교 국어국문학과 박사수료)

이의종 (서울대학교 국어국문학과 박사수료)

이하은 (한국외국어대학교 정보통신공학과)

이호진 ((주)언어과학 기업부설 연구소 소장)

장고은 (서울대학교 국어국문학과 박사수료)

장원철 ((주)언어과학 상무이사)

장하연 (미국 남가주대학교 언어학과 박사수료)

정용빈 (한국과학기술원 전산학부 석사과정)

정유남 (고려대학교 강사)

정유성 (한국과학기술원 시맨틱웹첨단연구센터 연구원)

정혜린 (서울대학교 국어국문학과 박사수료)

최원석 (나라아이넷(주) 기업부설 연구소 부소장)

최윤지 (인하대학교 한국어문학과 조교수)

최진 (서울대학교 국어국문학과 박사수료)

최현수 (연세대학교 언어정보학 협동과정 석사과정)

함영균 (한국과학기술원 전산학부 박사과정)

허인영 (고려대학교 국어국문학과 박사수료)

허철훈 (한국과학기술원 전산학부 석사과정)

홍은영 (서울대학교 국어국문학과 박사수료)

말뭉치 통합 검증

국립국어원의 ‘4차 산업혁명 대비 국어 빅데이터(말뭉치) 구축’ 사업의 일환으로 진행된 본 사업의 목적은 체계적이고 일원화된 검증을 통하여 국가 주도 구축 대규모 언어 자원의 품질을 높이는 데에 있다.

본 사업에서는 7개 층위 분석 말뭉치 (형태 분석, 어휘의미 분석, 개체명 분석, 주격 무형대용어 복원, 상호참조 해결, 구문 분석, 의미역 분석), 원문 수집 자료 2종 (문어 디지털 자료, 신문 기사 디지털 자료), 원시 말뭉치 4종(구어, 메신저 대화, 일상대화, 웹 말뭉치)을 검증하였다.

이 중에서 가장 중요한 7개 층위 분석 말뭉치의 내용 검증으로, 각 분석 대상 말뭉치의 7% 분량을 검증용 말뭉치로 구축하고 이를 검증 대상 분석 말뭉치와 비교하는 방법으로 검증을 수행하였다.

분석 말뭉치 검증은 주석 형식 검증, 내용 검증, 일관성 검증, 다층위 활용을 위한 통합 검증의 네 단계 검증을 수행하였다.

문어 디지털 자료와 신문 디지털 자료 등 원문 수집 자료는 세 단계의 형식 검증(인코딩 검사, XML 형식 검사, 데이터 유효성 검사)을 수행하였다. 형식 검증을 모두 통과한 원문 수집 자료는 문어 말뭉치와 신문 말뭉치로 구축하였다.

마지막으로 원시 말뭉치(구어, 메신저 대화, 일상 대화, 웹 언어)에 대해서는 네 단계의 형식 검증(인코딩 검사, XML 형식 검사, 데이터 유효성 검사, 발화 요소 오류 탐지)을 수행하였다.

이러한 일련의 과정을 통하여 대규모로 구축된 다양한 국어 빅데이터(말뭉치)의 품질을 제고함으로써 언어 처리 연구 및 산업적 개발에 유용한 자료로 활용될 수 있을 것으로 기대된다.

주요어: 분석 말뭉치 검증, 검증용 분석 말뭉치 구축, 원문 자료 검증, 원시 말뭉치 검증

차 례

제 1 장 서론

1. 사업 목적	3
2. 사업 수행 범위	5
3. 사업 수행 절차	7
4. 사업 수행 일정	9

제 2 장 분석 말뭉치 검증

1. 검증 대상 및 절차	13
2. 검증용 말뭉치 구축	17
3. 형식 및 내용 검증	55
4. 일관성 검증	87
5. 통합 검증	91

제 3 장 원문 자료 검증 및 원시 말뭉치 검증

1. 원문 자료 검증	95
2. 원시 말뭉치 구축	99
3. 원시 말뭉치 검증	101

제 4 장 결론 및 제언

1. 결론	109
2. 제언	111

표 차례

<표 1> 사업의 주요 범위와 과업 내용	5
<표 2> 사업 수행 경과	9
<표 3> 검증 대상 분석 말뭉치의 종류와 수량	13
<표 4> 문어 분석 말뭉치 구축 대상 원시 말뭉치 기초 통계	13
<표 5> 구어 분석 말뭉치 구축 대상 원시 말뭉치 기초 통계	13
<표 6> 분석 말뭉치 검증 세부 사항	14
<표 7> 검증 단계별 입력 및 출력	16
<표 8> 형태 분석 주석 내용	21
<표 9> 어휘의미 분석 주석 내용	22
<표 10> 개체명 분석 주석 내용	23
<표 11> 주격 무형대용어 복원 주석 내용	23
<표 12> 상호참조 해결 복원 주석 내용	24
<표 13> 구문 분석 주석 내용	25
<표 14> 의미역 분석 주석 내용	25
<표 15> 층위별 주석 단위	32
<표 16> 문어 분석 말뭉치 구축 규모 대비 검증용 말뭉치 구축 비율	35
<표 17> 구어 분석 말뭉치 구축 규모 대비 검증용 말뭉치 구축 비율	36
<표 18> 의미역 분석 말뭉치에서의 집중 검토 대상 서술어 목록	53
<표 19> 의미역 분석 말뭉치에서의 주요 불일치 사례	54
<표 20> 층위별 분석 말뭉치 검증 현황	56
<표 21> 형식 검증 및 주석 내용 검증 대상 통계	58
<표 22> 1차 형식 검증 세부 내용	59
<표 23> 1차 형식 오류 모듈 자료 구조	60
<표 24> 1차 형식 오류 코드	61
<표 25> 2차 형식 오류 검증 모듈 자료 구조	62

표 차례

<표 26> 2차 형식 오류 코드	64
<표 27> 층위별 분석 말뭉치 형식 검증 오류 유형 및 통계	67
<표 28> 형태 분석 내용 오류 검증 항목	69
<표 29> 형태 분석 내용 오류 검출 모듈 자료 구조	70
<표 30> 형태 분석 내용 검증 오류 코드	70
<표 31> 형태 분석 말뭉치 형식 및 주석 내용 검증 대상 통계	70
<표 32> 형태 분석 말뭉치 내용 검증 오류 유형 및 통계	71
<표 33> 형태 분석 말뭉치 주석 내용 검증 점수	71
<표 34> 어휘의미 분석 주석 내용 오류 검증 항목	72
<표 35> 어휘의미 분석 내용 오류 검출 모듈 자료 구조	72
<표 36> 층위별 주석 내용 검증 오류 코드	72
<표 37> 어휘의미 분석 말뭉치 형식 및 주석 내용 검증 대상 통계	73
<표 38> 어휘의미 분석 말뭉치 주석 내용 검증 결과	74
<표 39> 어휘의미 분석 말뭉치 주석 내용 검증 점수	74
<표 40> 개체명 분석 주석 내용 오류 검증 항목	74
<표 41> 주석 내용 오류 검출 모듈 자료 구조	77
<표 42> 층위별 주석 내용 검증 오류 코드	75
<표 43> 개체명 분석 말뭉치 형식 및 주석 내용 검증 대상 통계	75
<표 44> 개체명 분석 말뭉치 주석 내용 검증 오류 유형 및 통계	75
<표 45> 개체명 분석 말뭉치 주석 내용 검증 점수	76
<표 46> 주격 무형대용어 복원 주석 내용 오류 검증 항목	76
<표 47> 주석 내용 오류 검출 모듈 자료 구조	76
<표 48> 층위별 주석 내용 검증 오류 코드	77
<표 49> 주격 무형대용어 복원 말뭉치 형식 및 주석 내용 검증 대상 통계	77

표 차례

<표 50> 주격 무형대용어 복원 분석 말뭉치 주석 내용 검증 오류 유형 및 통계	78
<표 51> 주격 무형대용어 복원 말뭉치 주석 내용 검증 점수	78
<표 52> 상호참조 해결 주석 내용 오류 검증 항목	79
<표 53> 주석 내용 오류 검출 모듈 자료 구조	80
<표 54> 층위별 주석 내용 검증 오류 코드	80
<표 55> 상호참조 해결 말뭉치 형식 및 주석 내용 검증 대상 통계	81
<표 56> 상호참조 해결 분석 말뭉치 주석 내용 검증 오류 유형 및 통계	81
<표 57> 상호참조 해결 말뭉치 주석 내용 검증 점수	82
<표 58> 구문 분석 주석 내용 오류 검증 항목	82
<표 59> 주석 내용 오류 검출 모듈 자료 구조	82
<표 60> 층위별 주석 내용 검증 오류 코드	83
<표 61> 구문 분석 말뭉치 형식 및 주석 내용 검증 대상 통계	83
<표 62> 구문 분석 말뭉치 주석 내용 검증 오류 유형 및 통계	83
<표 63> 구문 분석 말뭉치 주석 내용 검증 점수	84
<표 64> 의미역 분석 주석 내용 오류 검증 항목	84
<표 65> 주석 내용 오류 검출 모듈 자료 구조	84
<표 66> 층위별 주석 내용 검증 오류 코드	85
<표 67> 의미역 분석 말뭉치 형식 및 주석 내용 검증 대상 통계	85
<표 68> 의미역 분석 말뭉치 주석 내용 검증 오류 유형 및 통계	86
<표 69> 층위별 주석 일관성 검증 모델	88
<표 70> 층위별 일관성 검증 결과	89
<표 71> 층위별 검증용 말뭉치 통합 검증 결과	91
<표 72> 층위별 검증 대상 분석 말뭉치 통합 검증 결과	92
<표 73> 원문 말뭉치의 스키마 유형	96
<표 74> 하위 검사 적합 판정에 따른 검증 결과 유형	97

표 차례

<표 75> 원문 말뭉치 유형별 검증 결과: 파일 수 (단위: 개)	98
<표 76> 원시 말뭉치 구축 결과: 파일 수	100
<표 77> 원시 말뭉치의 스키마 유형별	102
<표 78> process 시트	104
<표 79> parseLog 시트	104
<표 80> validLog 시트	104
<표 81> 원시 말뭉치 유형별 검증 결과: 파일 수	105

그림 차례

[그림 1] 사업 추진 배경 및 목표	4
[그림 2] 분석 말뭉치 검증 절차	8
[그림 3] 분석 말뭉치 검증 과정	16
[그림 4] 검증용 말뭉치 전체 구축 절차	18
[그림 5] 워크벤치 요구 사항 및 구현 사항	20
[그림 6] 형태 분석 주석 작업 화면	21
[그림 7] 어휘의미 분석 주석 작업 화면	22
[그림 8] 개체명 분석 주석 작업 화면	23
[그림 9] 주격 무형대용어 복원 주석 작업 화면	24
[그림 10] 상호참조 해결 주석 작업 화면	24
[그림 11] 구문 분석 주석 작업 화면	25
[그림 12] 의미역 분석 주석 작업 화면	26
[그림 13] 초별 주석 말뭉치 오류 예시: 선택한 어절과 다른 작업 대상	26
[그림 14] 워크벤치 오류 예시: 작업 창 누락 오류	27
[그림 15] 형태 분석 검수 작업 화면	28
[그림 16] 어휘의미 분석 검수 작업 화면	28
[그림 17] 개체명 분석 검수 작업 화면	29
[그림 18] 주격 무형대용어 복원 검수 작업 화면	29
[그림 19] 상호참조 해결 검수 작업 화면	30
[그림 20] 구문 분석 검수 작업 화면	30
[그림 21] 의미역 분석 검수 작업 화면	31
[그림 22] 워크벤치 관리자 페이지	31
[그림 23] 검증용 말뭉치 주석 과정	32
[그림 24] 1차 검수 대상	33
[그림 25] 2차 검수 대상	33

그림 차례

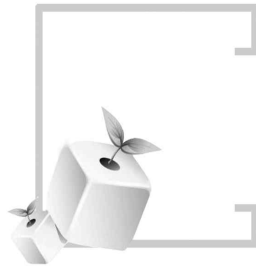
[그림 26] 1차 검수 예시	34
[그림 27] 2차 검수 예시	35
[그림 28] 형태 분석 지침 ‘3.가.1).라).(3)’	38
[그림 29] 형태 분석 지침 ‘3.가.2).다).(1)’	38
[그림 30] 형태 분석 검증용 말뭉치 수정 - ‘거’	38
[그림 31] 형태 분석 검증용 말뭉치 수정 - ‘이거, 그거, 저거’	39
[그림 32] 형태 분석 검증용 말뭉치 수정 - ‘NNπ’	39
[그림 33] 형태 분석 검증 말뭉치 수정 - ‘E _κ ’	40
[그림 34] 어휘의미 분석 오류 - 의미 번호 오분석	41
[그림 35] 어휘의미 분석 오류 - ‘개방’의 <우리말샘> 검색 결과	41
[그림 36] 어휘의미 검증 말뭉치 수정 - ‘개방’	42
[그림 37] 어휘의미 분석 지침 ‘3.라’	42
[그림 38] ‘헌법재판소’의 <우리말샘> 검색 결과	43
[그림 39] 어휘의미 검증 말뭉치 수정 - 구 등재어	43
[그림 40] 주격 무형대용어 복원 지침 ‘2.나.(4), (5)’	45
[그림 41] 주격 무형대용어 분석 말뭉치 수정 ‘누군가’	45
[그림 42] 주격 무형대용어 복원 지침 ‘1.나.(ㄷ)’	46
[그림 43] 주격 무형대용어 분석 말뭉치 수정	46
[그림 44] 주격 무형대용어 복원 지침 ‘1.가’	47
[그림 45] 상호참조 해결 지침 ‘2.1.1’	47
[그림 46] 상호참조 해결 말뭉치 수정 ‘제’	48
[그림 47] 상호참조 해결 지침 ‘2.1.5’	49
[그림 48] 상호참조 해결 말뭉치 수정	49
[그림 49] 주격 무형대용어 복원 지침 ‘2.1.16’	50
[그림 50] 상호참조 해결 말뭉치 수정	50

그림 차례

[그림 51] 구문 분석 말뭉치 수정 예시 - 명사구 열거 오분석	52
[그림 52] 형식 검증 및 주석 내용 검증 과정	55
[그림 53] 주석 일관성 검증 흐름도	87
[그림 54] 원문 말뭉치의 스키마 예시	97
[그림 55] 원시 말뭉치의 스키마 예시	103

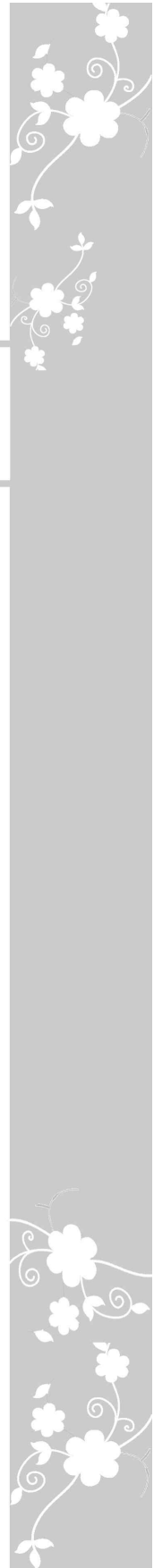
부록 차례

〈부록 1〉 층위별 초별 주식 말뭉치 JSON 구조	115
〈부록 2〉 주식 워크벤치 입력 및 출력 (JSON)	117
〈부록 3〉 층위별 분석 말뭉치 구축 지침 검토 의견	119



제 1 장

서 론



1. 사업 목적

본 사업의 목적은 국가 주도 ‘국어 빅데이터 구축 사업’의 하나로 체계적이고 일원화된 검증을 통한 국가 주도 구축 대규모 언어 자원의 품질을 높이는 데에 있다.

본 사업단에서는 다양한 국어 빅데이터 자료로서 층위별 7개 층위 분석 말뭉치 (형태 분석, 어휘의미 분석, 개체명 분석, 주격 무형대용어 복원, 상호참조 해결, 구문 분석, 의미역 분석), 원문 수집 자료 2종 (문어 디지털 자료, 신문 디지털 자료), 원시 말뭉치 6종 (구어, 메신저 대화, 일상대화, 웹 말뭉치 외 원문 수집 자료 2종으로 만든 문어 말뭉치, 신문 말뭉치 포함) 등을 검증한다. 일련의 검증 과정을 통하여 구축된 말뭉치가 전자 자료로서 적합한지, 분석 내용은 적합한 지 등을 검증하여 자료의 품질을 높인다.

분석 말뭉치 검증을 위하여 분석 말뭉치의 검증 기준이 되는 7개 층위 검증용 분석 말뭉치(이하 검증용 말뭉치)를 구축한다. 검증용 말뭉치를 검증 기준으로 삼아 검증용 말뭉치와 검증 대상 분석 말뭉치를 비교 검증한다.

두 말뭉치의 비교 검증 결과는 국립국어원의 검토를 거쳐 검증 대상 분석 말뭉치 구축 사업단에 전달된다. 검증 대상 분석 말뭉치 구축사업단은 두 말뭉치의 비교 검증 결과를 토대로 검증 대상 분석 말뭉치를 반복적으로 수정, 보완하여 검증 대상 분석 말뭉치의 품질을 높인다. 검증용 말뭉치는 대규모의 분석 말뭉치를 평가할 수 있는 기준이 될 뿐만 아니라, 고품질의 언어 주석 자료로 활용할 수 있다.

분석 말뭉치 검증을 위해 네 단계의 검증 체계를 확립하고, 이를 실제 분석 말뭉치 검증에 적용한다. 분석 말뭉치 검증은 먼저 주석 형식과 내용, 일관성을 층위별로 검증한 후 다층위 분석 말뭉치로의 활용을 위한 말뭉치 통합 검증으로 구성한다. 본 사업단에서 분석 말뭉치 검증 체계를 구축하고 이를 활용하여 검증을 수행함으로써 향후 다른 분석 말뭉치 주석 시 검증 절차로서 활용할 수 있다.

문어 디지털 자료와 신문 디지털 자료 등 원문 수집 자료는 세 단계의 형식 검증(인코딩 검사, XML 형식 검사, 데이터 유효성 검사)을 수행한다. 형식 검증을 모두 통과한 원문 수집 자료는 문어 말뭉치와 신문 자료 말뭉치로 구축된다.

본 사업단에서 구축한 문어 말뭉치와 신문 말뭉치와 원시 말뭉치 구축사업단에서 구축한 구어, 메신저 대화, 일상대화, 웹 원시 말뭉치는 네 단계의 형식 검증(인코딩 검사, XML 형식 검사, 데이터 유효성 검사, 발화 요소 오류 탐지)을 수행한다.

이러한 일련의 과정을 통하여 새로 구축되는 국어 빅데이터 자료의 품질을 높이고 지속적인 관리를 가능하게 하여 양질의 언어 자원 구축 환경을 제공한다.



[그림 1] 사업 추진 배경 및 목표

2. 사업 수행 범위

본 사업의 주요 과업 내용은 ‘4차 산업혁명 대비 국어 빅데이터(말뭉치) 구축’ 사업을 통해 구축된 대규모 말뭉치의 품질을 제고하기 위해 검증 방법론을 제시하고 이에 따른 검증을 수행하는 것이다.

본 사업의 수행 범위는 크게 세 가지로 제시되어 있다. 첫째는 2019년 말뭉치 구축 사업을 통해 구축된 층위별(형태, 어휘의미, 개체명, 상호참조, 주격 무형 대용어 복원, 구문, 의미역) 분석 말뭉치를 대상으로 주석 형식 및 내용, 일관성 검증 및 통합 검증을 수행하는 것이다. 둘째는 신문 기사 및 단행본 등의 문어 디지털 원문 수집 자료의 형식 오류를 검증한 후 원시 말뭉치로 구축하는 것이며, 마지막으로 방송, 일상 대화 등의 구어 원시 말뭉치, 메신저 대화 및 웹 원시 말뭉치를 대상으로 형식 오류를 검증하는 것이다.

제안요청서에 명시된 본 사업의 주요 범위와 자세한 과업 내용은 아래 <표 1>과 같다.

주요 사업 범위	과업 내용																
분석 말뭉치 검증	- 검증 대상 분석 말뭉치 규모																
	<table><tr><th>층위</th><th>수량</th></tr><tr><td>형태 분석</td><td>300만 어절 (문어 200만, 구어 100만)</td></tr><tr><td>어휘의미 분석</td><td>300만 어절 (문어 200만, 구어 100만)</td></tr><tr><td>개체명 분석</td><td>300만 어절 (문어 200만, 구어 100만)</td></tr><tr><td>주격 무형 대용어 복원</td><td>300만 어절 (문어 200만, 구어 100만)</td></tr><tr><td>상호참조 해결</td><td>300만 어절 (문어 200만, 구어 100만)</td></tr><tr><td>구문 분석</td><td>200만 어절 (문어 200만)</td></tr><tr><td>의미역 분석</td><td>200만 어절 (문어 200만)</td></tr></table>	층위	수량	형태 분석	300만 어절 (문어 200만, 구어 100만)	어휘의미 분석	300만 어절 (문어 200만, 구어 100만)	개체명 분석	300만 어절 (문어 200만, 구어 100만)	주격 무형 대용어 복원	300만 어절 (문어 200만, 구어 100만)	상호참조 해결	300만 어절 (문어 200만, 구어 100만)	구문 분석	200만 어절 (문어 200만)	의미역 분석	200만 어절 (문어 200만)
	층위	수량															
	형태 분석	300만 어절 (문어 200만, 구어 100만)															
	어휘의미 분석	300만 어절 (문어 200만, 구어 100만)															
	개체명 분석	300만 어절 (문어 200만, 구어 100만)															
	주격 무형 대용어 복원	300만 어절 (문어 200만, 구어 100만)															
	상호참조 해결	300만 어절 (문어 200만, 구어 100만)															
	구문 분석	200만 어절 (문어 200만)															
	의미역 분석	200만 어절 (문어 200만)															
	- 검증 방법론 수립 및 시행																
	<table><tr><th>검증 구분</th><th>실시 범위</th></tr><tr><td>형식 검증</td><td>분석 말뭉치 납품 분량 전수 검증</td></tr><tr><td>내용 검증</td><td>분석 말뭉치 납품 분량의 10% 표본 검증</td></tr><tr><td>일관성 검증</td><td>분석 말뭉치 납품 분량 전수 검증</td></tr><tr><td>통합 검증</td><td>분석 말뭉치 납품 분량 전수 검증</td></tr></table>	검증 구분	실시 범위	형식 검증	분석 말뭉치 납품 분량 전수 검증	내용 검증	분석 말뭉치 납품 분량의 10% 표본 검증	일관성 검증	분석 말뭉치 납품 분량 전수 검증	통합 검증	분석 말뭉치 납품 분량 전수 검증						
검증 구분	실시 범위																
형식 검증	분석 말뭉치 납품 분량 전수 검증																
내용 검증	분석 말뭉치 납품 분량의 10% 표본 검증																
일관성 검증	분석 말뭉치 납품 분량 전수 검증																
통합 검증	분석 말뭉치 납품 분량 전수 검증																
- 검증용 분석 말뭉치 구축: 층위별 검증 대상 말뭉치의 10% 이상																	
- 검증 단계별 일정 계획 수립 및 시행																	

	※ 분석 말뭉치 구축 사업의 구축 일정 및 분량에 맞추어 통합 검증 증을 포함하여 4단계로 나누어 검증	
원문 수집 자료 검증 및 원시 말뭉치 구축	- 검증 대상 원문 자료 규모	
	자료 구분	수량
	신문	12억여 어절
	문어	3.4억여 어절
	- 문어 디지털 원문 수집 자료의 형식 오류 검증 후 원시 말뭉치 구축	
원시 말뭉치 검증	- 검증 대상 원시 말뭉치 규모	
	원시 말뭉치 구분	수량
	일상대화	1,000시간 이상
	구어/준구어	1.54억여 어절
	메신저 대화	30만 발화 이상
	웹	누리소통망(SNS) 200만 발화, 블로그 1만 페이지, 게시판 1만 페이지, 리뷰 10만 리뷰 이상
	- 방송, 일상대화 등 구어 원시 말뭉치, 메신저 대화와 웹 말뭉치를 대상으로 형식 오류 검증	

〈표 1〉 사업의 주요 범위와 과업 내용

3. 사업 수행 절차

본 사업은 분석 말뭉치 검증, 원문 자료 검증 및 원시 말뭉치 구축, 원시 말뭉치 검증의 세 가지 과업으로 구성되어 있다.

첫 번째 과업인 분석 말뭉치 검증을 수행하기 위해서는 주석 내용 검증의 기준이 되는 검증용 분석 말뭉치 구축이 선행되어야 한다.

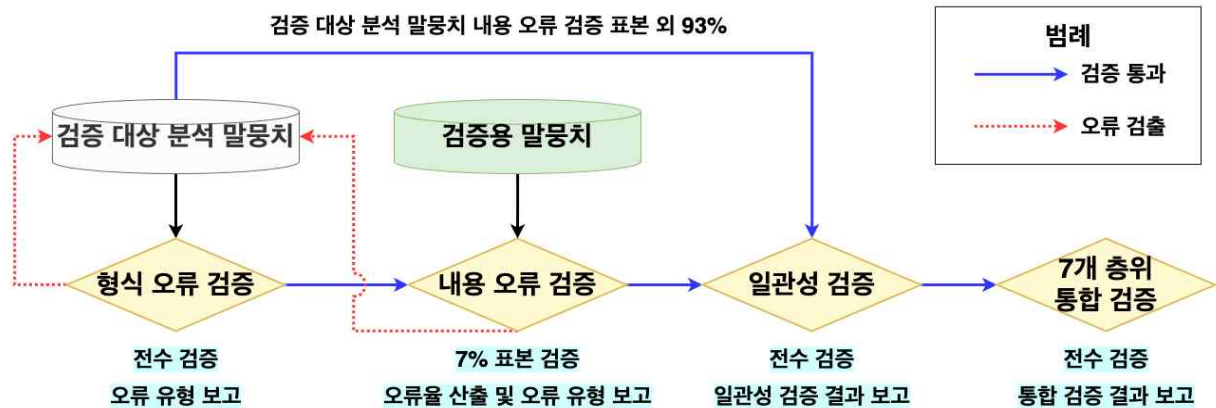
국립국어원에서 분석 대상 원시 말뭉치를 전달받아 내용 검증의 표본으로 협의한 검증 대상 7%(21만여 어절)를 검증용 말뭉치 구축 대상으로 확정하였다.

검증용 말뭉치 구축 대상의 원시 말뭉치는 주석자의 주석 전에 한국어 자동 분석 도구¹⁾를 사용하여 형태, 구문 분석 등을 정보를 부착하는 초별 분석을 수행하였다. 자동 분석 도구를 활용한 초별 분석 결과는 주석자가 층위별 분석 말뭉치 주석 시 참고 자료로 활용할 수 있도록 워크벤치를 통해 제공되었다.

이와 동시에 주석자들이 작업할 작업 환경인 워크벤치를 구축하기 위한 설계 작업과 개발 작업을 진행하였다. 주석자들은 개발된 워크벤치를 활용하여 층위별 초별 주석 말뭉치의 주석 내용을 수정, 보완하는 방식으로 주석 작업을 수행하였다. 주석 작업은 동일한 주석 단위에 대해 두 명의 주석자가 작업한 후 두 명의 주석 결과 일치 여부에 따른 검수 절차를 거쳐 검증용 말뭉치를 구축하였다.

분석 말뭉치 검증은 [그림 2]와 같이 전체적으로 형식 검증, 검증용 말뭉치와의 비교를 통한 내용 검증, 일관성 검증, 통합 검증으로 이루어진다. 검증 대상 분석 말뭉치가 들어 오면 먼저 형식 검증을 실시하여 말뭉치의 주석 형식 및 태그, 원문 일치 여부 등을 전수 검사한다. 형식 검증을 통과한 말뭉치는 검증용 말뭉치와의 비교를 통한 내용 검증을 수행한 후, 검증용 말뭉치와 일치하지 않는 주석 내용을 오류로 판단하여 층위별 검증 지표에 따라 오류 보고서를 작성한다. 그러나 검증용 말뭉치에 오류가 포함될 가능성이 있으므로 오류 보고서를 검토하여 검증용 말뭉치 주석을 수정하는 단계를 거친 뒤 내용 재검증을 실시하고 이 결과를 토대로 재작성한 오류 보고서를 국립국어원으로 발송한다. 국립국어원의 오류 보고서 검토, 승인을 받은 후 형식 및 내용 검증 결과 보고서를 작성하여 국립국어원에 보고한다.

1) 한국전자통신연구원(ETRI) 언어지능그룹의 지원으로 한국어 분석 도구인 ‘엑소브레인 한국어 분석 툴킷 v3.0’을 사용하였다.



[그림 2] 분석 말뭉치 검증 절차

일관성 검증은 내용 검증에서의 표본 검증의 한계를 보완하기 위한 것으로, 검증 대상 분석 말뭉치가 전체적으로 고르게 주석되었는지를 확인한다. 먼저 층위별 검증용 분석 말뭉치를 학습 데이터로 입력하여 층위별 자동 분석기를 학습한다. 이 분석기를 활용하여 검증 대상 원시 말뭉치를 자동으로 분석한 후 이 자동 분석된 말뭉치를 기준으로 검증 대상 말뭉치와 비교하여 검증 대상 말뭉치의 오류율이 전체적으로 고르게 분포하는지를 확인하는 방식으로 진행하였다.

7개 층위의 분석 말뭉치는 모두 동일한 원시 말뭉치를 대상으로 주석하였다. 구축 주체가 다른 7개 층위 분석 말뭉치를 다층위 말뭉치로 활용하기 위해서는 기본적으로 문장 수, 어절 수 등의 기초 통계량이 동일해야 한다. 통합 검증에서는 이를 확인하는 절차를 수행하였다.

두 번째 과업인 원문 자료 검증 및 원시 말뭉치 구축에서는 수집된 문어 및 신문 자료를 대상으로 자료 검증과 원시 말뭉치 구축을 수행하였다. 원문 자료 검증은 인코딩 검사 및 XML 형식 검사, 데이터 유효성 검사를 포함한다. 원시 말뭉치 구축은 검증 단계에서 검출된 오류를 수정한 원문 자료에 문단 정보를 추가하며, 구축 지침에 명시된 형식 정의를 준수하여 원시 말뭉치를 생성하는 작업이다. 원문 자료의 검증 및 원시 말뭉치 구축 작업을 위해 Node.js로 개발한 CLI 도구를 사용하였다.

세 번째 과업인 원시 말뭉치 검증은 일상 대화, 구어/준구어, 웹 언어, 메신저 대화 등 총 4종의 원시 말뭉치를 대상으로 자료의 무결성을 검증하는 작업이다. 주요 검사 항목은 원문 자료 검증과 마찬가지로 인코딩 검사, XML 형식 검사, 데이터 유효성 검사이며 말뭉치 특성에 따라 추가적인 오류 탐지를 수행하였다. 다종의 원시 말뭉치를 일괄 검증하기 위해 Node.js로 개발한 CLI 도구를 사용하며 각 말뭉치의 구성과 형식적 특성에 맞춰 검증을 실시하였다.

위의 절차와 일련의 과정들은 국립국어원과의 유기적인 협조를 통해 수행하였다.

4. 사업 수행 일정

본 사업은 당초 2019년 6월 25일부터 12월 31일까지 6.2개월의 기간 동안 수행하는 것으로 계약되었다. 그러나 본 사업의 검증 대상이 되는 분야별, 층위별 말뭉치 구축 사업의 계약 및 지침 확정, 구축 일정의 지연 등으로 2020년 2월에 구축이 완료되는 분석 말뭉치의 검증까지 완료하기 위해 2020년 2월 29일까지로 사업 기간을 2개월 연장하여 수행하였다.

구분	예정	수행 일정	비고
사업 기간	2019. 6. 25. ~ 2019. 12. 31. (6.2 개월)	2019. 6. 25. ~ 2020. 2. 29. (8.2개월)	사업 기간 연장
주석 체계 확립	2019. 7월 초 ~ 8월 말	2019. 7월 초 ~ 10월 중순	데이터 입수 및 주석 체계 확립 등
검증용 말뭉치 구축	2019. 7월 초 ~ 11월 말	2019. 9월 중순 ~ 2020. 2월 중순	전 층위 기준
분석 말뭉치 검증	2019. 8월 초 ~ 12월 중순	2019. 10월 초 ~ 2020. 2월 말	전 층위 기준
원문 수집 자료 검증 및 원시 말뭉치 구축	2019. 10월 초 ~ 12월 중순	2019. 10월 초 ~ 2020. 1월 중순	전 원문 수집 자료 기준
원시 말뭉치 검증	2019. 8월 초 ~ 12월 중순	2019. 9월 초 ~ 2020. 2월 중순	전 원시 말뭉치 기준

〈표 2〉 사업 수행 경과

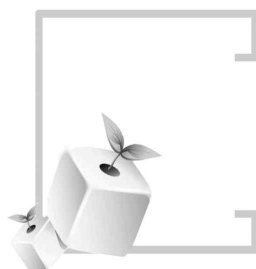
2019년 7월 11일 착수 보고를 기점으로 사업이 시작되었으며, 종료 보고는 2020년 2월 25일에 진행하였다. 정상적인 사업의 수행을 위하여, 매월 월간 보고서를 제출하였으며, 특히 사항에 대한 별도의 수시 보고를 수행하였다.

분석 말뭉치 검증은 층위별 일정에 따라 문, 구어 최소 4회(시범 검증 포함)에서 최대 6회까지 이루어졌다. 분석 말뭉치 검증은 2019년 9월 30일 형태 분석 시범 검증을 시작으로 2020년 2월 27일까지 층위별 분석 말뭉치 검증을 수행하였다.

분석 말뭉치 검증에 사용한 검증용 말뭉치는 2019년 8월 22일 형태 분석 검증용 말뭉치 구축을 시작으로 하여, 2020년 2월 15일에 구축을 완료하였다. 검증용 말뭉치는 일차적으로 구축 완료한 이후에 분석 말뭉치 검증 결과를 기반으로 하여 수정, 보완을 거쳐 2020

년 2월 27일에 국립국어원에 납품하였다.

원문 자료 검증 및 원시 말뭉치 구축, 원시 말뭉치 검증은 사업 개시로부터 12월까지
국립국어원의 원시 말뭉치 구축 지침을 분석하고 원문 자료 수집 사업자와 원시 말뭉치
구축 사업자의 표본 자료를 검토하였다. 원문 자료를 원시 말뭉치로 변환하는 작업은 1
월에 진행되었다. 구축된 말뭉치에 대한 검증 및 구축 작업은 1월부터 2월까지 진행되었
으며, 구축 사업자의 말뭉치 납품 일정에 따라 차례로 진행되었다. 원문 자료 및 원시 말
뭉치를 검증하여 보고한 오류를 구축 사업자가 수정하면 재검증을 진행하였다.



제 2 장

분석 말뭉치 검증



1. 검증 대상 및 절차

분석 말뭉치 검증은 ‘4차 산업혁명 대비 국어 빅데이터(말뭉치) 구축’ 사업을 통해 구축된 분석 말뭉치가 정해진 말뭉치 주석 체계에 따라 정해진 형식과 지침에 따라 주석되었는지 등을 확인하여 분석 말뭉치 품질을 검증하는 과정이다.

검증 대상이 되는 분석 말뭉치는 동일한 원시 말뭉치에 7종의 언어 정보가 별개로 부착된 말뭉치로 종류와 분량은 <표 3>과 같다.

분석 종류	수량
형태 분석	300만 어절 (문어 200만, 구어 100만)
어휘의미 분석	300만 어절 (문어 200만, 구어 100만)
개체명 분석	300만 어절 (문어 200만, 구어 100만)
주격 무형 대용어 복원	300만 어절 (문어 200만, 구어 100만)
상호참조 해결	300만 어절 (문어 200만, 구어 100만)
구문 분석	200만 어절 (문어 200만)
의미역 분석	200만 어절 (문어 200만)

<표 3> 검증 대상 분석 말뭉치의 종류와 수량

말뭉치의 장르는 문어의 경우 신문 기사이며, 구어의 경우는 공적 독백, 공적 대화, 일상 대화로 구성되어 있다. 문어와 구어의 기초 통계는 <표 4>, <표 5>와 같다.

	문서	문단	문장	어절
문어	7,265	62,440	150,085	2,000,215

<표 4> 문어 분석 말뭉치 구축 대상 원시 말뭉치 기초 통계 (단위: 개)

	장르	파일	발화	어절
구어	공적 독백	106	20,202	135,578
	공적 대화	292	143,799	632,112
	사적 대화	23	58,780	233,540
합계		421	222,781	1,001,230

<표 5> 구어 분석 말뭉치 구축 대상 원시 말뭉치 기초 통계 (단위: 개)

분석 말뭉치 검증은 네 단계로 구성되어 있으며, 단계별로 아래의 내용에 초점을 맞추어 검증을 진행하였다.

- 형식 검증: 정해진 주식 형식에 맞게 말뭉치가 구축되었는가?
- 주식 내용 검증: 정해진 주식 지침에 따라 말뭉치가 분석되었는가?
- 주식 일관성 검증: 분석 말뭉치 전체에 걸쳐서 일관된 주식 양상을 보이는가?
- 통합 검증: 분석 말뭉치가 다층위 말뭉치로써 사용될 수 있는가?

검증 단계	검증 개요		검증 내용 및 방법
1차 형식 검증	대상	분석 말뭉치 전량	- 주식 표준 형식 준수 검사(JSON 유효성 검사)
	목적	주식 표준 형식 준수 여부 확인	
	기준	국립국어원 분석 말뭉치 JSON 형식 v1.9	
2차 형식 검증	대상	분석 말뭉치 전량	- 주식 표준 형식 key별 value의 자료형 및 내용 유효성 검사
	목적	주식 표준 형식 기준 value 유효성 검사	
	기준	주식 표준 형식 key별 value 기준	
주식 내용 검증	대상	분석 말뭉치 표본 (검증 대상의 7%)	- 검증 대상 말뭉치 주식 내용과 검증용 말뭉치 주식 내용 비교 - 불일치 주식에 대한 정오 판별
	목적	주식 지침 대비 주식 내용 검사	
	기준	7개 층위별 주식 지침	
주식 일관성 검증	대상	분석 말뭉치 전량	- 검증용 말뭉치로 자동 분석기 학습 - 자동 분석기로 표본 이외 분량 주식 - 자동 주식 검증용 말뭉치와 검증 대상 말뭉치 주식 내용 비교 - 구간별 주식 일치율 경향성 평가
	목적	분석 말뭉치 내 주식 일관성 검사	
	기준	자동 주식 검증용 말뭉치 대비 검증 대상 말뭉치의 구간별 주식 일치율과 주식 일치율 평균값 대비 신뢰 구간 (99% confidence interval, confidence alpha=0.01) 포함 여부 검사	
통합 검증	대상	분석 말뭉치 전량	- 원시 말뭉치 대비 문서, 문단, 문장, 어절 정보 비교
	목적	다층위 말뭉치 활용성 검사	
	기준	분석 대상 원시 말뭉치 (문어, 구어)	

<표 6> 분석 말뭉치 검증 세부 사항

검증의 첫 단계로 검증용 말뭉치와 검증 대상 분석 말뭉치를 비교하여 형식 검증한다. 형식 검증은 <표 6>과 같이 두 단계로 나뉘어진다. 1차 형식 검증은 검증 대상 분석 말뭉치의 전량을 대상으로 하여 표준 구축 형식을 준수하였는지 여부를 검사한다. 본 단계의 검증 기준은 분석 말뭉치 표준 구축 형식인 ‘국립국어원 말뭉치 형식 v1.9’을 따른다. 본 단계에서는 표준 입력 형식 준수 여부를 검사하기 위해 표준 구축 형식에 정의된 JSON 구조 및 자료형의 유효성을 검사한다.

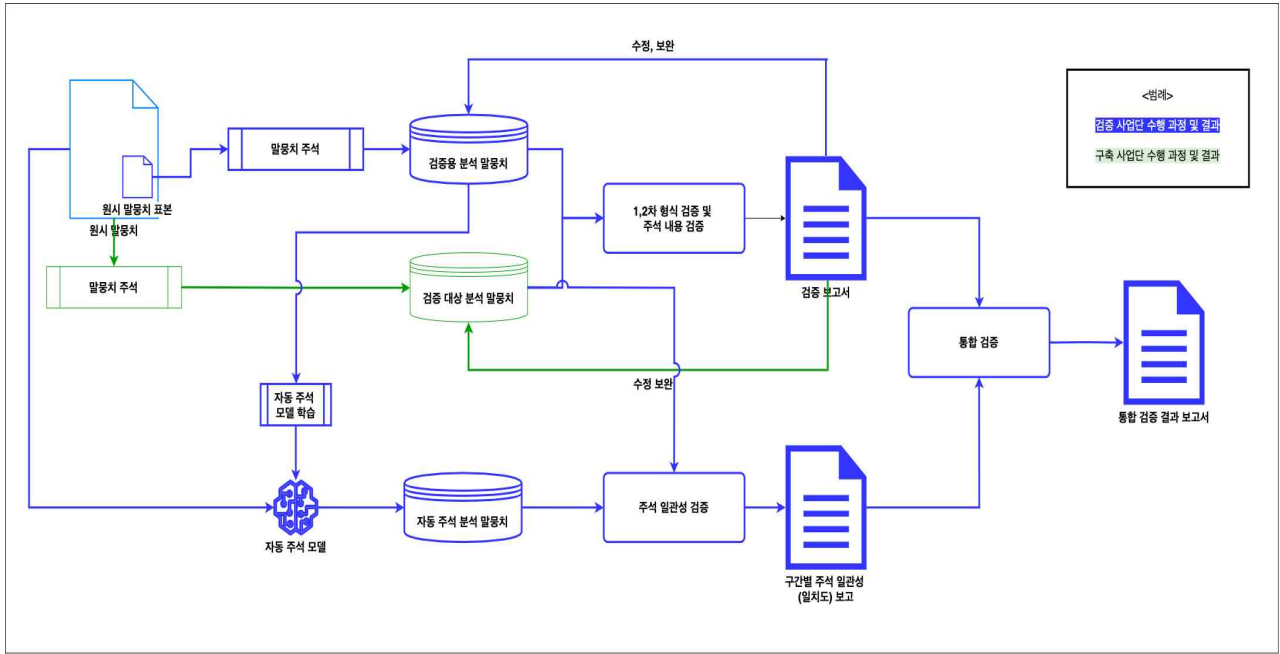
2차 형식 검증 역시 검증 대상 분석 말뭉치 전량에 대하여 표준 구축 형식 내 value 유효성 검사를 한다. 검증 기준은 1차 형식 검증과 마찬가지로 표준 구축 형식을 따르며, 표준 구축 형식에 정의된 key 별 value가 정의된 자료형과 주석 내용을 준수하였는지 검사한다. 본 단계에서 검사하는 value의 주석 내용은 층위별로 정해진 주석 표지(태 그 세트 등) 이외의 것들을 사용하였거나, 잘못된 어절 번호 사용 등 해당 층위에서 주석될 수 없는 주석 내용이 있는지 확인하는 작업이다.

주석 내용 검증은 형식 검증을 통과한 검증 대상 분석 말뭉치 중 검증용 말뭉치와 일치한 주석 단위 표본(7% 범위)에 대하여 두 말뭉치를 비교하여 불일치 주석 단위를 검출하는 과정이다. 불일치 내용의 정오 판별의 기준은 각 층위의 분석 지침을 준용한다.

주석 일관성 검증은 검증 대상 분석 말뭉치 전량에 대하여 말뭉치의 주석 일관성을 검사하는 과정이다. 주석 일관성 검증은 검증용 말뭉치를 학습하여 만든 자동 주석 모델을 사용하여 원시 말뭉치를 자동 주석한 주석 일관성 검증용 분석 말뭉치(이하 자동 주석 말뭉치)와 검증 대상 분석 말뭉치를 비교하여 주석 일치도의 경향성을 분석한다. 자동 주석 말뭉치와 검증 대상 분석 말뭉치를 10개 구간으로 나누어 각 구간의 주석 일치도를 구하고, 10개 구간의 주석 일치도의 평균값의 99% 신뢰 구간(confidence alpha = 0.01) 내에 각 구간의 주석 일치도가 포함되는지 확인한다. 99% 신뢰 구간을 벗어나는 구간은 다른 구간에 비해서 주석 일관성이 낮은 구간으로 평가하였다.

통합 검증은 검증 대상 분석 말뭉치 전량에 대하여 구축된 분석 말뭉치가 다층위 분석 말뭉치로 활용되는 데 문제가 없는지를 검사한다. 검증 기준은 분석 대상 문, 구어 원시 말뭉치이며, 검증 기준 대비 문서, 문단, 문장, 어절 정보 등을 비교하여 일치하는지 검사한다.

분석 말뭉치의 전체 검증 과정은 [그림 3]과 같다.



[그림 3] 분석 말뭉치 검증 과정

형식과 주석 내용 검증 단계를 포함한 검증 전 단계의 입력과 출력은 <표 7>과 같다.

검증 단계	입력	출력	비고
1차 형식 검증	- 검증 대상 분석 말뭉치	- 1차 형식 검증 결과	- 1차 형식 검증, 2차 형식 검증, 주석 내용 검증 결과는 한 로그 파일(txt)로 출력
2차 형식 검증	- 검증 대상 분석 말뭉치	- 2차 형식 검증 결과	
주석 내용 검증	- 입력1: 검증용 분석 말뭉치 - 입력2: 검증 대상 분석 말뭉치	- 주석 내용 불일치 결과	
주석 일관성 검증	- 입력1: 자동 주석 말뭉치 - 입력2: 검증 대상 분석 말뭉치	- 말뭉치 구간별 주석 일치도	- 10구간 분할
통합 검증	- 입력1: 원시 말뭉치 - 입력2: 검증 대상 분석 말뭉치	- 문서, 문장, 어절 통계 - 문서, 문장, 어절 불일치 결과	- 로그 파일(txt)로 출력

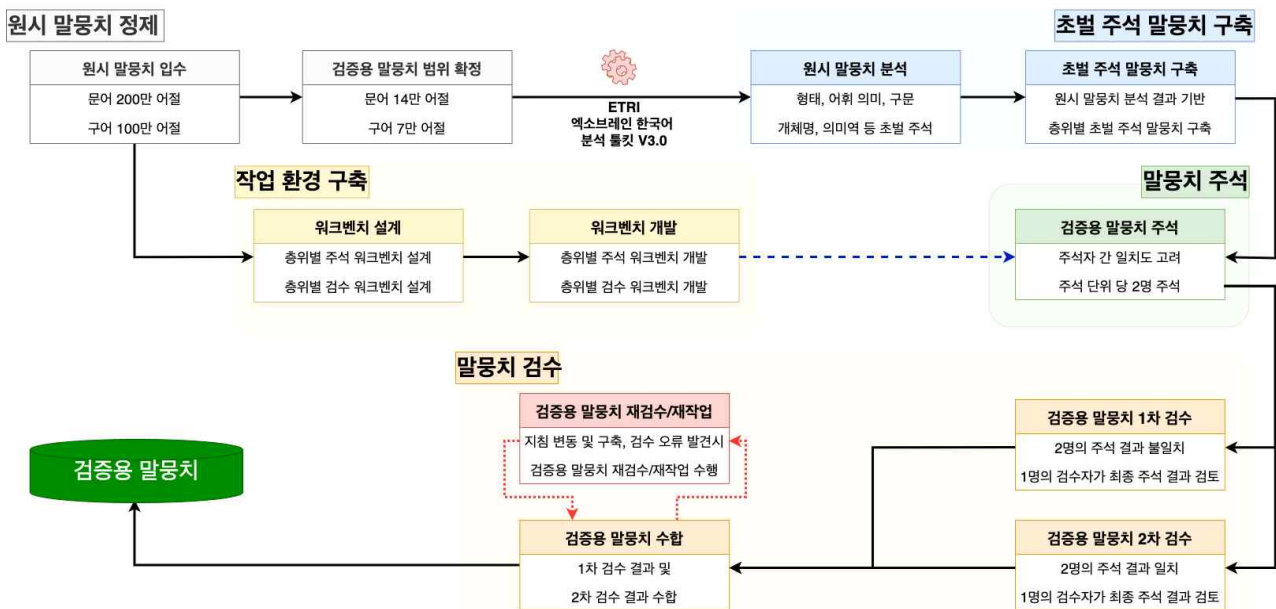
<표 7> 검증 단계별 입력 및 출력

2. 검증용 말뭉치 구축

검증용 말뭉치는 검증 대상 말뭉치와의 비교 검토를 통해 주석 내용을 검증하기 위해 검증 대상 말뭉치 중 표본으로 추출한 일부를 분석하여 정답 세트로 구축하는 것이다. 검증의 신뢰도를 담보하기 위해서는 검증용 말뭉치의 품질을 확보하는 것이 무엇보다 중요하다.

2.1. 구축 절차

검증용 말뭉치 전체 구축 절차는 [그림 4]와 같다.



[그림 4] 검증용 말뭉치 전체 구축 절차

우선, 구축 대상 원시 말뭉치(문어 200만 어절, 구어 100만 어절)를 입수한 후 검증용 말뭉치 구축 범위(문어 14만 어절, 구어 7만 어절)를 확정하였다.

다음으로 검증용 말뭉치 구축 범위의 원시 말뭉치를 한국전자정보통신연구원(ETRI)의 엑소브레인 한국어 분석 툴킷 v3.0을 활용하여, 분석 말뭉치 구축을 위한 참고 자료로 활용하기 위한 초별 주석 말뭉치를 구축하였다. 초별 주석 말뭉치의 주석 결과를 활용하여 총위별로 주석 작업에 필요한 정보를 주석자에게 제공하였다.

7개 총위의 의존성과 다총위 말뭉치 구축을 고려하여 한국전자통신연구원 언어지능그룹에서 본 사업을 위해 제공한 한국어 분석 도구인 ‘엑소브레인 한국어 분석 툴킷 v3.0’을 사용하여 원시 말뭉치를 분석하였다. 이를 활용하여 형태소, 동음이의어, 개체명, 구문, 개체명, 의미역 분석 결과 등을 분석한 원시 말뭉치 분석 결과를 얻었다.

이 분석 결과를 기반으로 총위별 JSON 형식의 초별 주석 말뭉치를 제작하였다. 초별 주석 말뭉치는 주석자들이 작업할 대상의 말뭉치로 원시 말뭉치 분석 결과에서 총위별로 필요한 정보만을 추출하여 만든 말뭉치이다. 총위 별로 구축한 초별 주석 말뭉치에 포함

된 세부 내용은 <부록 1>에 수록되어 있다. 사업 수행 계획 당시에는 7개 층위의 의존성에 기반을 두어 타 층위의 주석 결과를 활용하고자 <부록 1>과 같이 층위별 초별 주석 말뭉치를 구성하였으나 사업 수행 과정에서 층위별 주석 일정이 조정되면서 일부 초별 주석 말뭉치 정보는 엑소브레인 한국어 분석 툴킷 v3.0의 결과를 그대로 초별 주석 말뭉치에 사용하게 되었다.

초별 주석 말뭉치 주석 정보를 활용하여 층위별 검증용 말뭉치 구축 주석자들은 새로운 정보를 주석하거나 이미 부착된 주석 정보를 수정하는 방식으로 검증용 말뭉치를 구축하였다. 검증용 말뭉치 주석과 검수가 끝난 주석 단위를 하나로 모아 검증용 말뭉치를 완성한다.

2.2. 구축 환경

검증용 말뚝치를 구축하기 위해서 작업용 워크벤치를 온라인 환경²⁾에서 구축하였다. 워크벤치는 [그림 5]와 같이 층위별로 주식용 워크벤치와 검수용 워크벤치를 구축하였으며, 주식 작업의 진행 상황을 살펴볼 수 있는 관리자 페이지로 구성하였다. 보안을 유지하기 위하여 개인 계정을 생성하여 들어가기로 하면 개인에게 할당된 층위의 작업만 할 수 있게 작업공간을 구축하였다.



[그림 5] 워크벤치 요구 사항 및 구현 사항

2.2.1. 주식 워크벤치

검증용 말뚝치 주식 워크벤치는 층위별 초별 주식 말뚝치를 기반으로 하여 주식자들이 층위별 검증용 말뚝치를 구축하였다. 주식 워크벤치의 입력과 출력은 <부록 2>에 수록되어 있다.

형태 분석 주식 워크벤치 내에서 층위별로 주식한 내용은 아래 <표 8>과 같으며, 작업 화면은 [그림 6]과 같다.

2) www.crowdworks.kr

층위	주석 내용	주석 방법
형태 분석	형태소 분절 수정	- 초별 주석 말뭉치 내 morps 내용 수정 - 형태소 분절 결과를 병합하거나 나눔
	형태소 태그 부여	- 수정한 분절 결과에 대한 형태소 태그 부여 - 형태소 태그 47개 ³⁾

〈표 8〉 형태 분석 주석 내용

작업하는데 어려운 점이 있으신가요?

⑩작업 가이드

다음 문장을 읽고 어절 내의 형태소 분절이 잘못된 부분을 찾아 고쳐주세요.

“**휠체어**를 탄 장애인 이 공중에서 단식농성까지 하는 건 제가 처음일걸요.”

선택 어절	형태소 태그	보류/삭제	행 추가
형태소 분절 결과	형태소 태그	보류/삭제	행 추가
“	SS	작업보류	<div>+</div> <div>전체 결과보기</div>
휠체어	NNG	작업보류 삭제	
를	JKO	작업보류 삭제	

우리말샘

[그림 6] 형태 분석 주석 작업 화면

어휘의미 분석 주석 워크벤치 내에서 층위별로 주석한 내용은 아래 〈표 9〉와 같으며, 작업 화면은 [그림 7]과 같다. 어휘의미 주석 작업을 위해서 워크벤치 내에 우리말샘 API를 연동하여 주석자가 작업 시에 우리말샘을 쉽게 검색할 수 있도록 하였다.

3) 본 사업에서 사용한 형태 분석 말뭉치 구축 지침 사용한 형태소 태그 중 관형사(MM)류는 엑소브레인 한국어 분석 툴킷 V3.0에 사용된 형태소 태그와 달리 성상 관형사(MMA), 지시 관형사(MMD), 수 관형사(MMN)로 세분화하였다.

층위	주석 내용	주석 방법
어휘의미 분석	어휘의미 분석 단위 수정	<ul style="list-style-type: none"> - 어휘의미 분석 대상은 체언으로(NNG, NNB, NNP, NP, NR, XR)로 한정 - 형태소 분석 결과를 토대로 어휘의미 번호 부여 단위 수정 - 형태소 분석 결과를 어휘의미 번호 부여 단위를 기준으로 병합하거나 나눔
	어휘의미 번호 부여	<ul style="list-style-type: none"> - 우리말샘 기준 어휘의미 번호 부여 - 미등재어(777), 어의 번호 없음(888) 등 예외 처리 번호 부여

<표 9> 어휘의미 분석 주석 내용

[그림 7] 어휘의미 분석 주석 작업 화면

개체명 분석 주석 워크벤치 내에서 층위별로 주석한 내용은 아래 <표 10>과 같으며, 작업 화면은 [그림 8]과 같다.

층위	주석 내용	주석 방법
개체명 분석	개체 부여 대상 선정	- 어절 단위 개체 범위 선택 - 형태소 주석 결과를 활용하여 개체가 아닌 어절 내 조사 등을 삭제
	개체명 태그 부여	- 개체 부여 대상에 개체명 태그 부여 - 개체명 태그 15개

<표 10> 개체명 분석 주석 내용

다음 문장을 읽고 개체를 찾아 재정의 해주세요.

스페인 상원, 자치권 박탈 확정... 美·EU "스페인 정부 노력 지지"

선택 어절	스페인	
형태소 분석 결과	형태소 태그	선택여부
스페인	NNP	<input checked="" type="checkbox"/> 변경불가

작업 데이터			
스페인	LC	보류하기	삭제
상원	CV	보류하기	삭제
美	LC	보류하기	삭제
EU	OG	보류하기	삭제
스페인	LC	보류하기	삭제
정부	OG	보류하기	삭제

우리말샘 개체 추가하기

[그림 8] 개체명 분석 주석 작업 화면

주격 무형대용어 복원 주석 워크벤치 내에서 층위별로 주석한 내용은 아래 <표 11>과 같으며, 작업 화면은 [그림 9]와 같다.

층위	주석 내용	주석 방법
주격 무형대용어 복원	문장 내 주어 복원 대상 서술어 확인	- 구문 분석 결과 구문 분석 태그가 VP, VP_MOD인 어절이 주어에 해당하는 어절이 문장 내에 존재하는지 확인
	주어 복원	- 복원 대상 서술어의 주어를 문서 내에서 선택함 - 문서 내에서 주어가 드러나지 않으면 영주어를 주석함

<표 11> 주격 무형대용어 복원 주석 내용

문장
1. [이집트 민주화 시위 8일째...학울락 놀란 지구촌] 中... 이집트 관련 인터넷 검색 차단
2. ?살입-빈부격차 달은 끝...텐안면사태 악용 떠올려
3. 중국이 이집트에서 벌어지고 있는 대규모 시위사태의 불통이 중국으로 뻗지 않을까 바짝 긴장하고 있다.
4. 10년 만에 루이비통 백백가랑로 대형 쇼핑몰 조방제는 사치품이 아니라서 중국, 이집트가 너무 달아간 데모이라고 CNN도 아시요 보도했다

현재 문장에서 선택된 서술어의 주어가 모두 문장 내에 있습니까?
☐ or ☒

자세히 보기

어절	태그
?살입-빈부격차	NP_OBJ
달은	VP_MOD
끝...	NP
텐안면사태	NP_OBJ
악용	NP_OBJ
떠올려	VP

[그림 9] 주격 무형대용어 복원 주석 작업 화면

상호참조 해결 주석 워크벤치 내에서 층위별로 주석한 내용은 아래 <표 12>와 같으며, 작업 화면은 [그림 10]과 같다.

층위	주석 내용	주석 방법
상호참조 해결	개체 선택	- 어절 단위의 개체 범위 선택 - 형태소 주석 결과를 활용하여 개체가 아닌 어절 내 조사 등 삭제
	군집 생성	- 선택된 개체와 동일한 공지시 관계에 있는 명사구의 군집 생성

<표 12> 상호참조 해결 복원 주석 내용

IMF, 올 한국성장률 전망 0.25%p 낮추듯
[한겨레] "금융시장 취약...가계부채 문제"
국제통화기금(IMF) 협의단이 올해 우리나라 경제성장률 전망치를 종전 3.5%에서 0.25%포인트 사실상 하향 조정했다.

협의단은 지난달 말부터 2주동안 한국 경제 상황을 직접 살펴왔다. 세계경제 여건이 예상보다 나빠지는 상황에서, 한국 금융시장에 취약성이 존재하고 높은 수준의 가계부채가 문제가 될 수 있다는 게 근거다.

국제통화기금 협의단은 12일 오후 정부 과전청사에서 '2012년 아이엠에프-한국 연례협의' 결과를 발표하면서, 올해 한국 경제성장률이 지난 4월 발표 때보다 0.25%포인트 낮아질 것이라 예상했다.

국제통화기금은 애초 성장률 전망치로 3.5%를 제시한 바 있다.

국제통화기금은 "유럽 위기의 심화가 예상된다"며 "(한국의) 유럽에 대한 직접적인 위험 노출은 크지 않지만 위기 여파가 미국과 중국으로 전이될 경우 큰 영향을 받을 수 있다"고 밝혔다.

호이코르 평가단장은 "애초 올 하반기에 한국 경기가 회복될 것으로 전망했다"며 "그러나 세계 경기둔화 가능성이 커지면서 한국의 경기 회복도 애초 예상보다 한두 분기쯤 늦은 2013년 초에나 가능할 것"이라고 전망했다.

국내 금융시스템의 취약성에 대한 경고도 나왔다.

국제통화기금은 "한국이 아시아에서 가장 개방된 경제 중 하나이지만, 자본유출입 변동성 및 외화조달 리스크에 노출돼 있다"며 "한국 정부가 '태일 리스크'(예상치 못한 위험)에 맞서 비상 대책을 강구해야 한다"고 조언했다.

국제통화기금은 이어 "최근 가계에 대한 비은행 금융기관 대출이 급증하고 있고, 이 상황에 대한 긴밀한 모니터링과 시정조치가 필요하다"며 높은 가계부채에 대한 감시 강화도 요구했다.

상호참조해결집합 추가

IMF

그룹수정

-

+

국제통화기금(IMF) 협의단

그룹수정

-

+

올 한국성장률 전망

그룹수정

-

+

올해 우리나라 경제성장률 전망

메모 | 수정 | 삭제

지

0.25%p

그룹수정

-

+

0.25%포인트

메모 | 수정 | 삭제

0.25%포인트

메모 | 수정 | 삭제

우리나라

그룹수정

-

+

한국 금융시장에 취약성이 존재하...

그룹수정

-

+

근거

메모 | 수정 | 삭제

선택 어절

국제통화기금(IMF) 협의단이

선택 취소

형태소 분절

형태소 분절 결과	데이터 편집 포함 여부
국제	<input checked="" type="checkbox"/> 포함
통화	<input checked="" type="checkbox"/> 포함
기금	<input checked="" type="checkbox"/> 포함
(<input checked="" type="checkbox"/> 포함
IMF	<input checked="" type="checkbox"/> 포함
)	<input checked="" type="checkbox"/> 포함
협의	<input checked="" type="checkbox"/> 포함
단	<input checked="" type="checkbox"/> 포함
이	<input checked="" type="checkbox"/> 포함

데이터 편집

[그림 10] 상호참조 해결 주석 작업 화면

구문 분석 주석 워크벤치 내에서 층위별로 주석한 내용은 아래 <표 13>과 같으며, 작업 화면은 [그림 11]과 같다.

층위	주석 내용	주석 방법
구문 분석	지배소 주석	- 어절 간 지배소 주석
	구문 분석 태그 주석	- 구문 태그 및 기능 태그 주석 - 구문 태그 9개, 기능 태그 6개

<표 13> 구문 분석 주석 내용

문장

알뜰폰-지역 케이블 방송 '경쟁 제한성 높다' 의견 담긴 것으로 알려져

	구문태그	기능태그	지배소	전체보기
알뜰폰-지역	NP	▼	케이블	
케이블	NP	▼	방송	
방송	NP	▼	높다'	
'경쟁	NP	▼	제한성	
제한성	NP	▼	SBJ	
높다'	VP	▼	의견	
의견	NP	▼	SBJ	
담긴	VP	▼	MOD	
것으로	NP	▼	AJT	
알려져	VP	▼		

[그림 11] 구문 분석 주석 작업 화면

의미역 분석 주석 워크벤치 내에서 층위별로 주석한 내용은 아래 <표 14>와 같으며, 작업 화면은 <그림 12>와 같다.

층위	주석 내용	주석 방법
의미역 분석	의미역 부여 대상 서술어 선택	- 서술어 수정, 추가, 삭제를 통해서 의미역 부여 대상 서술어 선택
	서술어 번호 선택	- 격틀 정보를 활용하여 서술어 번호 선택
	서술어의 논항 주석	- 필수역과 부가역의 최상위 지배소에 논항 주석

<표 14> 의미역 분석 주석 내용

문장
위르달의 복지정책은 '빈곤이 빈곤을 낳는다'는 그의 학문적인 문제의식과 맞닿아 있다.

자세히보기

위르달의	복지정책은	'빈곤이	빈곤을	낳는다'는	그의	학문적인	문제의식과	맞닿아	있다.
NP_MOD	NP_SBJ	NP_SBJ	NP_OBJ	VP_SBJ	NP_MOD	VNP_MOD	NP_AJT	VP	VP
<input type="text"/>	<input type="text"/>	ARG0	ARG1	날.01	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input type="text"/>	ARG1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	ARG2	맞대다 000102	ARGM-AUX

[그림 12] 의미역 분석 주석 작업 화면

검증용 말뭉치 주석 작업과정에서 워크벤치 상에서 주석자로부터 보고된 오류를 두 가지 유형으로 분류할 수 있고 해당 유형에 따라 수정, 보완 조치하였다.

○ 유형 1: 초별 주석 말뭉치의 오류

초별 주석 말뭉치 오류의 경우, 일괄적으로 원시 말뭉치를 분석하여 초별 주석 말뭉치를 구축하는 과정에서 발생한 오류이다. 대표적인 예시로 [그림 13]과 같이 엑소브레인 한국어 분석 툴킷 V3.0을 사용하여 문장 분할이 잘못되어 문장 및 어절 ID가 반복되는 경우가 발생하였다. 이러한 오류의 경우, 초별 주석 말뭉치를 직접 수정하여 오류를 바로잡았다.

다음 문장을 읽고 어절 내의 형태소 분절이 잘못된 부분을 찾아 고쳐주세요.

형태소분절 & 태그

2008년, 세계적인 투자은행 리먼브러더스가 파산한 것을 계기로 전 세계에 금융 위기가 몰아닥쳤다. 미국과 유럽 등 각국 정부는 심각한 불황을 타개하려고 엄청난 빚을 내서 경기 부양에 나선다. 최근 벌어지고 있는 국가 신용 위기는 어마어마한 규모의 재정 적자가 부메랑이 되어 돌아온 것이다.

선택 어절	파산한
형태소 분절 결과	형태소 태그
것	NNB
을	JKO

행 추가

+

전체 결과보기

선택한 어절과 다른 결과가 나옴

우리말샘

이전 어절

다음 어절

저장하고 다음으로

작업불가

[그림 13] 초별 주석 말뭉치 오류 예시: 선택한 어절과 다른 작업 대상 (형태 분석 검수 예시)

○ 유형 2: 워크벤치의 오류

[그림 14]와 같이 워크벤치에서 오류가 발생할 경우, 워크벤치 페이지 기능 수정 및 보완을 먼저 수행하여 작업자의 불편을 최소화하고자 노력하였다.

[그림 14] 워크벤치 오류 예시: 작업 창 누락 오류 (어휘의미 분석 예시)

2.2.2. 검수 워크벤치

검증용 말뭉치 검수 워크벤치는 주석 워크벤치의 기능에 두 명의 주석자의 결과를 보여주어 주석 결과를 수정 및 보완할 수 있는 기능을 추가하였다.

형태 분석 검수 워크벤치의 경우 [그림 15]와 같이 대상 문단에서 한 어절을 선택할 경우 두 주석자의 형태소 분석 결과, 형태소 태그가 입력창 아래에 나타난다. 이를 참고하여 검수자는 해당 문단의 최종 형태소 분석 결과를 입력한다.

다음 문장을 읽고 여결 내의 형태소 분석이 잘못된 부분을 찾아 고쳐주세요.

당첨금에 대한 과세 역시 강화된다. 경마·소싸움 등에서 받은 배당금이 200만 원을 초과하면 세금을 내도록 했다. **그동안은** 배팅액(10만원 이하)의 100배를 초과할 때만 세금을 매겨왔다.

선택 어절: 그동안은

형태소 분석 결과	형태소 태그	보류/삭제	행 추가
그	NNG	작업보류	+ 전체 결과보기
동안	NNG	작업보류	
은	JX	삭제	

우려할 점

[그림 15] 형태 분석 검수 작업 화면

어휘의미 검수 워크벤치의 경우 [그림 16]과 같이 대상 문단에서 두 주석자가 주석한 결과가 화면 우측에 나타난다. 이를 참고하여 검수자는 해당 문단의 최종 어휘의미 분석 결과를 입력한다.

통신은 또 김 위원장이 이날 평양 백화원 국빈관에서 왕 부장을 면담하는 자리에서 북한은 **한반도** 비핵화를 위해 노력하고 있다며 **“한반도** 경제의 긴장상태를 원치 않는다”고 말했다고 전했다. 그는 또 **“중국과 함께 협조와** 조화를 이뤄 6자회담을 부단히 진전시켜 나가야 한다”고 말했다는 것.

선택 어절: 통신은

형태소 분석 결과	형태소 태그	어휘의미 번호	예외 상황	병합 여부		
통신	NNG	어휘의미 번호를 기입하여	어휘미동재여	형태미동재여	작업불가	<input type="checkbox"/> 병합여부
은	JX		작업불가			<input type="checkbox"/> 병합여부

어절단위 코멘트

우려할 점

분절	태그	어휘번호	예외 상황	분절	태그	어휘번호	예외 상황
통신	NNG	003		통신	NNG	001	
은	JX			은	JX		
또	MAG			또	MAG		
김	NNP	010		김	NNP	010	
위원장	NNG	001		위원장	NNG	001	
이	JKS			이	JKS		

닫기

[그림 16] 어휘의미 분석 검수 작업 화면

개체명 검수 워크벤치의 경우 [그림 17]과 같이 대상 문단에서 주석자4)가 주석한 결과가 화면 우측에 나타난다. 이를 참고하여 검수자는 해당 문단의 최종 개체명 분석 결과를 입력한다.

4) 개체명 분석 검증용 말뭉치는 한 주석 단위를 한 명의 주석자가 작업하였다. 따라서 검수는 한 주석자의 결과를 검수자가 확인하는 방식으로 이루어졌다. 이에 대해서는 1.4절에서 상술한다.

다음 문장을 읽고 개체를 찾아 재정의 해주세요.

형태소분절 & 태그

러시아의 남하를 구실로 한반도 침략을 정당화했던 전범자들의 후예 일본 주류 보수우익들이 소련이 사라진 지금 다시 중국을 주적으로 삼정하고 있다. 미국은 그런 일본과 함께 한국을 끌어들이 한-미-일 삼각(군사)동맹을 만들고 있다. 한반도가 또 서들의 주전장이 왜간다.

선택 어절			정당화했던			작업 데이터		
형태소 분절 결과	형태소 태그	선택 여부	러시아	LC	보류하기	삭제		
정당	XR	<input checked="" type="checkbox"/> 변경	전범자	CV	보류하기	삭제		
화	XSN	<input checked="" type="checkbox"/> 변경	후예	CV	보류하기	삭제		
하	XSV	<input checked="" type="checkbox"/> 변경	일본	LC	보류하기	삭제		
있	EP	<input checked="" type="checkbox"/> 변경	소련	OG	보류하기	삭제		
던	ETM	<input checked="" type="checkbox"/> 변경	중국	OG	보류하기	삭제		
			미국	OG	보류하기	삭제		
			일본	OG	보류하기	삭제		
			한국	OG	보류하기	삭제		
			한	OG	보류하기	삭제		
			미	OG	보류하기	삭제		
			일	OG	보류하기	삭제		
			한반도	LC	보류하기	삭제		
			한반도 침략	EV	보류하기	삭제		

우리말 선택 개체 추가하기

[그림 17] 개체명 분석 검수 작업 화면

주격 무형대용어 복원 검수 워크bench의 경우 [그림 18]과 같이 대상 문서에서 선택한 문장에서 두 주석자가 해당 문장을 주어 복원 대상으로 주석하였는지를 확인한다. 이를 참고하여 검수자는 해당 문단의 최종 주격 무형대용어 복원 주석 결과를 입력한다.

문장

1. * 모든 친구가 '선생님' / 가르침을 주고 받다
 2. * "오스트레일리아는 국토의 3분의 2가 건조기후예요.
 3. * 강수량이 500mm 이하죠.
 4. * 11월에는 오대기후이데 사환이 살기에 오대기후가 후계이데 건조기후가 후계이데"

현재 문장에서 선택된 서술어의 주어가 모두 문장 내에 있습니까? ☐ 예 ☒ X [참고데이터A: X](#) [참고데이터B: X](#)

자세히보기	
어절	태그
모든	DP
친구가	NP_SBJ
'선생님'	NP
/가르침을	NP_OBJ
주고	VP
받다	VP

[그림 18] 주격 무형대용어 복원 검수 작업 화면

상호참조 해결 검수 워크bench의 경우 [그림 19]와 같이 대상 문서에서 두 주석자가 주석한 공지시 관계 집합이 화면 좌측에 나타난다. 이를 참고하여 검수자는 해당 문서의 최종 상호참조 해결 주석 결과를 입력한다.

작업 1

무슬림들

이슬람 극단주의 무장단체의 공격을 받은

이슬람 극단주의 무장단체

케나

60명

북동부 국경 도시 만데라

버스에 타고 있던 무슬림인 압디 모하무드

트릭

트릭 운전사

중격을 가해 버스를 멈춰세운 무장괴한들

소말리아

케나

북동부

'알라 외에 다른 신은 없으며 무함마드는 알라'

알라 로바 만데라 주지사

기독교 등 비무슬림 승객들

작업 2

소말리아 무장단체 '알샤바'

무슬림 승객들

버스에 타고 있던 무슬림이 노 압디 모하무드

트릭 운전사

알라

28명이 숨지 노 이사진

일카에다 연계 조직이 노 알샤바

케나의 북동부

기독교 등 비무슬림 승객들

케나

이슬람 극단주의 무장단체의 공격을 받은

승객들

기독교인만 죽이려 하자 무슬림들 "우리도 죽여라"

이슬람 극단주의 무장단체의 공격을 받은 케나의 버스에서 무슬림 승객들이 기독교 등 비무슬림 승객들을 보호해, 자칫 대규모 살인으로 이어졌을 수 있는 참극을 막았다. 트위터(www.twitter.com, www.twitterkr.com)는 자기 명의의 게시물에 쓸 영문 주소와 이메일 등을 입력하는 것이 전부다.

NHN의 '미투데이'(www.me2day.net)나 다음의 '요즘'(yozm.daum.net), SK커뮤니케이션즈의 '커넥팅'(connect.nate.com) 등 국내 포털업체의 SNS는 기존 아이디를 활용할 수 있어 가입이 매우 쉽다.

소말리아 무장단체 '알샤바'

기자는 일단 트위터에 가입하면서 추천받은 트위터 리스트에서 뉴욕타임스 북리뷰와 영국 프리미어리그 프로축구팀 아스날의 트위터들(팔로잉)하기로 했다.

케나서 60명 태운 버스 공격

게시 가능한 글의 분량이 140자로 제한되기 때문에 기업이나 단체의 공식 트위터에서 제공하는 정보는 대부분 공식 홈페이지로 연결되는 인터넷 링크라는 사실은 아쉽다. 무슬림 승객들 종교 넘어 지향

무슬림 승객들 종교 넘어 지향

포털사이트에서 '김연아 트위터' '오바마 트위터' 같은 키워드로 검색을 해 이들의 트위터 주소를 얻는다.

누군가의 트위터를 일단 팔로잉하기 시작하면 그를 읽는 사람과 그가 읽는 사람을 따라가며 내 트위터 인맥도 급속히 확장된다.

21일 아침 케나 수도 나이로비에서 60여명을 태우고 북동부 국경 도시 만데라로 향하던 버스를 소말리아 무장단체 알샤바 대원들이 공격했다.

만데라까지 150km 남짓 남은 엘와크 인근이었다.

<포터> 통신 주 외신들의 보도를 종합하면, 중격을 가해 버스를 멈춰세운 무장괴한들은 승객들에게 하차하라고 명령하며 무슬림과 기독교인으로 갈라서라고 요구했다.

버스에 타고 있던 무슬림인 압디 모하무드 압디는 10여명의 알샤바 대원들이 무슬림 승객들에게 기독교인들과 떨어지라고 요구했으나, 승객들이 거부했다고 당시 상황을 설명했다.

그는 "무장단체 대원들이 총을 쏘겠다고 위협했지만 우리는 계속 거부했고 우리 형제 자매들을 보호했다"고 말했다.

그는 비무슬림들이 쉽게 식별될 수 없도록 버스에서 무슬림들이 일부에게 무슬림식 복장을

상호참조해결집합 추가

그룹수정

트릭 운전사

그룹수정

소말리아

그룹수정

북동부

그룹수정

'알라 외에 다른 신은 없으며 무함마드는 알라'

그룹수정

알라 로바 만데라 주지사

그룹수정

기독교인

그룹수정

승객들

그룹수정

28명이 숨지 노 이사진

그룹수정

[그림 19] 상호참조 해결 검수 작업 화면

구문 분석 검수 워크bench의 경우 [그림 20]과 같이 대상 문서에서 두 주석자가 주석한 구문, 기능 태그를 보여준다. 또한, 두 주석자의 지배소 주석 내용 각각 다른 색상의 화살표로 보여준다. 이를 참고하여 검수자는 해당 문서의 최종 구문 분석 주석 결과를 입력한다.

문장

이에 대해 한국방송 이사회 이길영 이사장은 <한겨레>와의 통화에서 "아직까지 정해진 건 없다"고 말했다. 이사회는 오는 18일 임시이사회에서 후임 사장 선임을 위한 절차 등을 논의할 예정이다.

	구문태그	기능태그	지배소	전체보기
이에	NP	이동노 NP	대해	<div>작업1</div> <div>작업2</div> <div>내 작업</div>
대해	VP	이동노 VP	말했다.	
한국방송	NP	이동노 NP	이사회	
이사회	NP	이동노 NP	이사장은	
이길영	NP	이동노 NP	이사장은	
이사장은	NP	SBJ	말했다.	
<한겨레>와의	NP	MOD	통화에서	
통화에서	NP	AJT	없다"고	
"아직까지	AP		없다"고	
정해진	VP	MOD	건	
건	NP	SBJ	없다"고	
없다"고	VP	CMP	말했다.	
말했다.	VP			

[그림 20] 구문 분석 검수 작업 화면

의미역 분석 검수 워크bench의 경우 [그림 21]과 같이 대상 문서에서 두 주석자가 주석한 의미역 주석 결과를 보여준다. 이를 참고하여 검수자는 해당 문서의 최종 의미역 분석 주석 결과를 입력한다.

0건
④ 작업가이드
⑤ 문의하기

문장
15일 문화재단과 인문학협동조합 공동주최로 열린 '신경숙 표절 사태와 한국문학의 미래' 토론회에서였다.

자세히보기
참고데이터 보기

15일	문화재단과	인문학협동조합	공동	주최로	열린	'신경숙	표절	사태와	한국문학의	미래	토론회에서였다.
NP_AJT	NP_CNJ	NP	NP	NP_AJT	VP_MOD	NP	NP	NP_CNJ	NP_MOD	NP	VNP
ARGM-TMP	ARGO	ARGO	ARGM-MNR	일.01							ARG1

작업자 정보
operator@1911.kr
작업자: @hanmail.net
작업환경: pc
일주일: null

복작업불가사용

복작업관리사용

참고데이터

작업자1

15일	문화재단과	인문학협동조합	공동	주최로	열린	'신경숙	표절	사태와	한국문학의	미래	토론회에서였다.
NP_AJT	NP_CNJ	NP	NP	NP_AJT	VP_MOD	NP	NP	NP_CNJ	NP_MOD	NP	VNP
ARGM-TMP	ARGO			일.01							ARG1

작업자2

15일	문화재단과	인문학협동조합	공동	주최로	열린	'신경숙	표절	사태와	한국문학의	미래	토론회에서였다.
NP_AJT	NP_CNJ	NP	NP	NP_AJT	VP_MOD	NP	NP	NP_CNJ	NP_MOD	NP	VNP
				일.01							ARG1

[그림 21] 의미역 분석 검수 작업 화면

2.2.3. 작업 관리 기능

워크벤치는 [그림 22]와 같이 검증용 말뭉치 주석 진행 상황을 점검할 수 있는 관리자 페이지를 갖추었다. 주석자의 작업 시간, 작업량, 주석 단위 주석자 확인 등 층위별 검증용 말뭉치 구축 상황을 점검하여, 구축 작업 진행 상황에 대한 다양한 정보를 활용하였다.

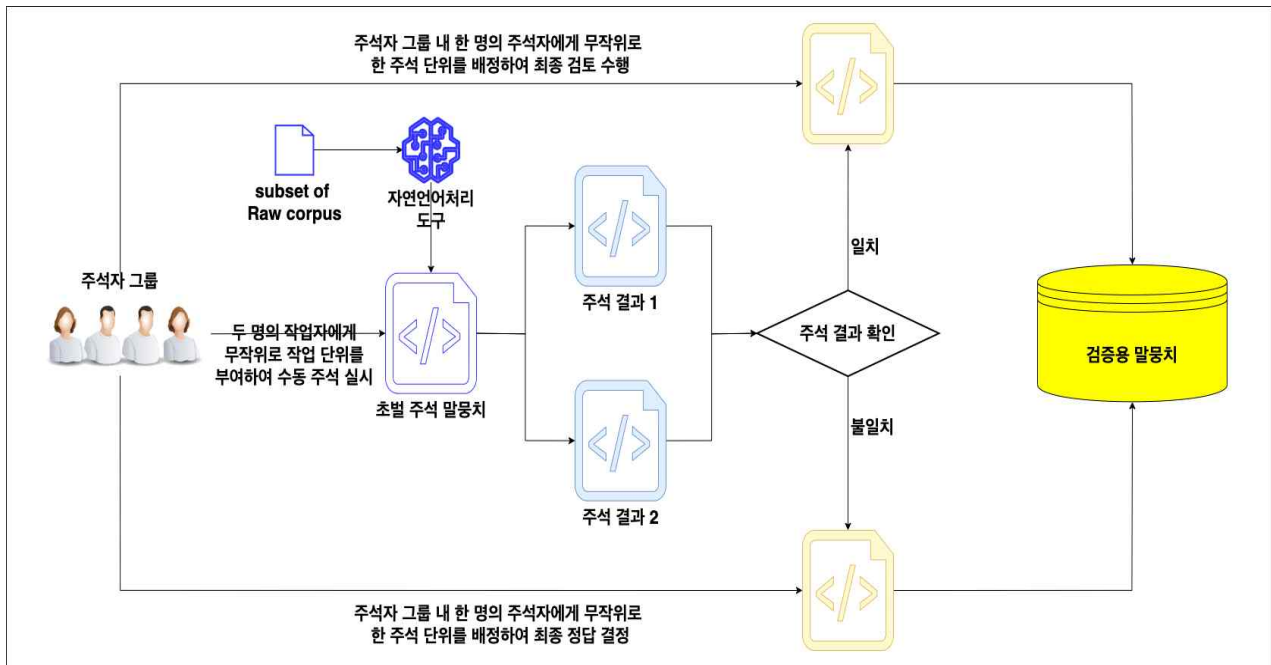
프로젝트 진행 상황													
#	프로젝트 이름	상세조회	총데이터수	시작일	완료예정일	작업				검수			
						작업	오브젝트	작업자	일평균 작업자	검수	오브젝트	검수자	일평균 검수자
1	카이스트 주석무형대용어 시범분석	남박말 조회	894	2019-10-22	2020-01-17	894	894	9	3	0	0	0	0
2	카이스트 말뭉치	남박말 조회	85,120	2019-08-26	2020-02-28	76,102	75,537	54	15	3,626	3,442	9	10

진행 확인													
번호	구분	프로젝트명	작업명	시작일	✓ 작업진행	반대	보류	검수 완료	재검수 대기	검수 대기	검수 진행	작업자	✓ 검수자
6157	작업	총위2_아미역분석_최종검증_검수	작업물가_다름	2020.02.04	1/1	0	0	0	0	1	0	1	0
6156	검수		검수count_다름	2020.02.04	2546/2546	0	0	0	0	2546	0	6	0
6155	검수		검수count_같은	2020.02.04	1016/1016	0	0	0	0	1016	0	6	0
6147	작업		검수count_다름	2020.02.03	314/314	0	0	0	0	314	0	7	0
6146	검수	총위2_아미역분석_구어_3_검수	검수count_같은	2020.02.03	198/198	0	0	0	0	198	0	6	0
6145	작업		작업물가_다름	2020.02.03	2/2	0	0	0	0	2	0	2	0
6144	작업		작업물가_같은	2020.02.03	4/4	0	0	0	0	4	0	4	0
6143	검수		검수count_다름	2020.02.03	169/169	0	0	0	0	169	0	5	0
6142	검수	총위2_아미역분석_구어_2_검수	검수count_같은	2020.02.03	46/46	0	0	0	0	46	0	5	0
6141	작업		작업물가_다름	2020.02.03	1/1	0	0	0	0	1	0	1	0

[그림 22] 워크벤치 관리자 페이지

2.3. 말뭉치 주석 및 검수

검증용 말뭉치는 층위별 초별 주석 말뭉치를 결과를 수정하거나, 그 결과를 활용하여 층위별로 새로운 정보를 주석하는 방법으로 진행하였다. 검증용 말뭉치 주석 과정은 [그림 23]과 같다.



[그림 23] 검증용 말뭉치 주석 과정

검증용 말뭉치 주석은 주석자 간 일치도를 고려하여 하나의 주석 단위를 두 명이 주석하는 다중 할당 방식으로 주석하였다. 층위별 주석 단위는 주석자가 한 번에 주석하는 분량을 뜻하는데, 층위별 주석 단위는 아래 <표 15>와 같다.

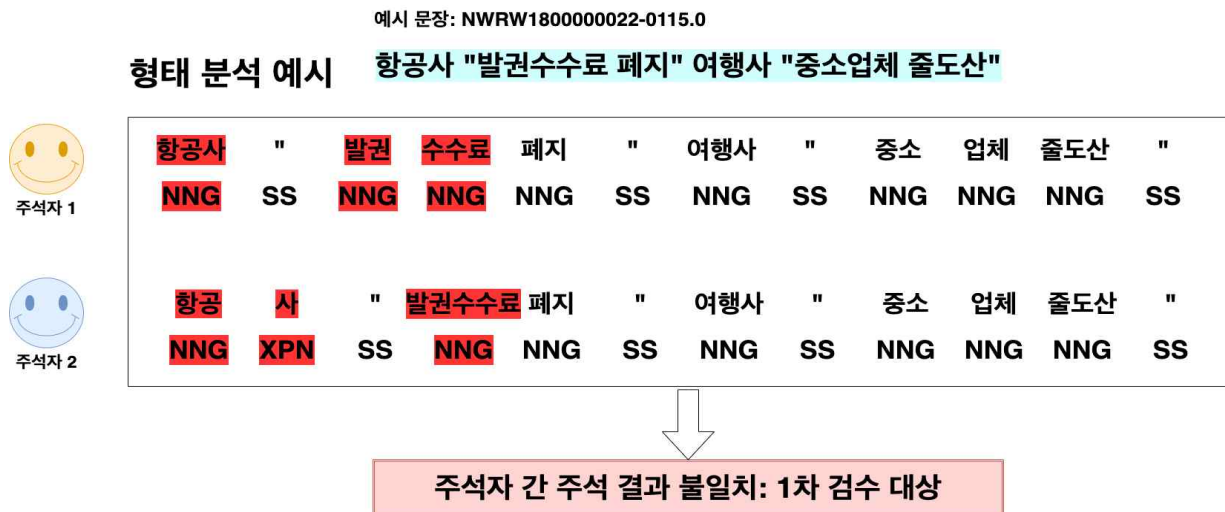
층위	형태 분석	어휘의미 분석	개체명 분석	주격 무형대용어 복원	상호참조 해결	구문 분석	의미역 분석
주석 단위	문단	문단	문단	문서	문서	문단	문장

<표 15> 층위별 주석 단위

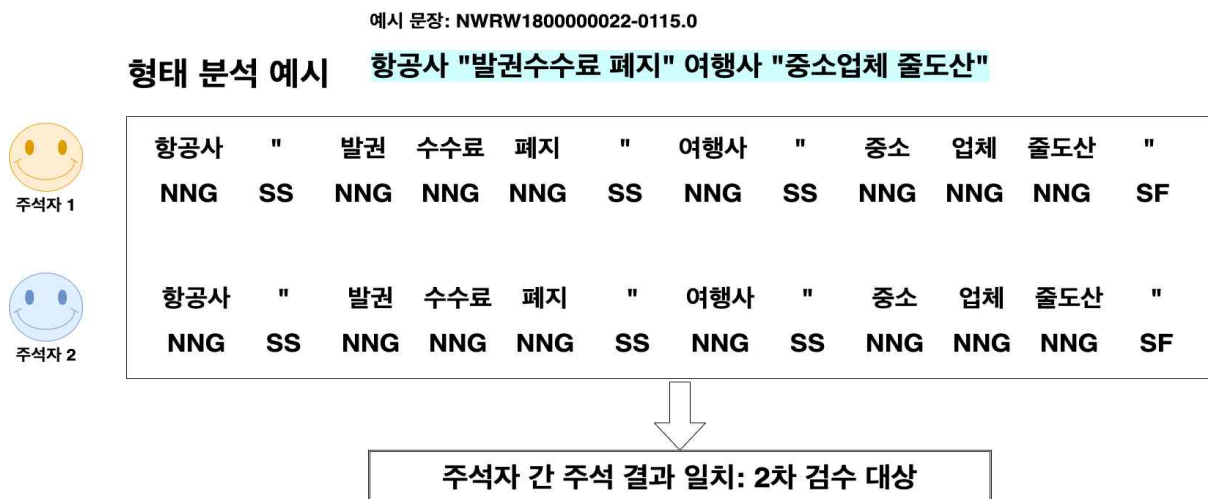
5) 모든 층위는 다중 할당 방식으로 계획하였으나, 개체명 말뭉치의 경우 구축사업단의 사업 일정 관계로 단일 할당 (한 주석 단위를 한 명이 주석)하는 방식으로 주석하였다.

문단을 주석 단위로 한 형태, 어휘의미, 개체명, 구문 분석 층위는 전, 후 문장의 문맥 정보를 활용하기 위해 문단을 한 분석 단위로 정하였다. 주격 무형대용어 복원과 상호참조 해결은 한 문서 내에서 주석을 수행하기 때문에 문서를 한 분석 단위로 정하였다. 의미역의 경우 한 문장 내에서 주석을 수행하기 때문에 문장을 한 분석 단위로 정하였다.

주석이 완료되면 두 주석자의 작업 결과에 따라 주석 완료 말뭉치를 두 가지로 분류하였다. 두 주석자가 한 주석 단위의 주석 내용 중 하나라도 불일치한 주석이 있는 경우를 1차 검수 대상으로 분류하고([그림 24]), 한 주석 단위의 주석 내용이 모두 일치한 경우 2차 검수 대상으로 분류하였다([그림 25]).



[그림 24] 1차 검수 대상 (형태 분석 예시)

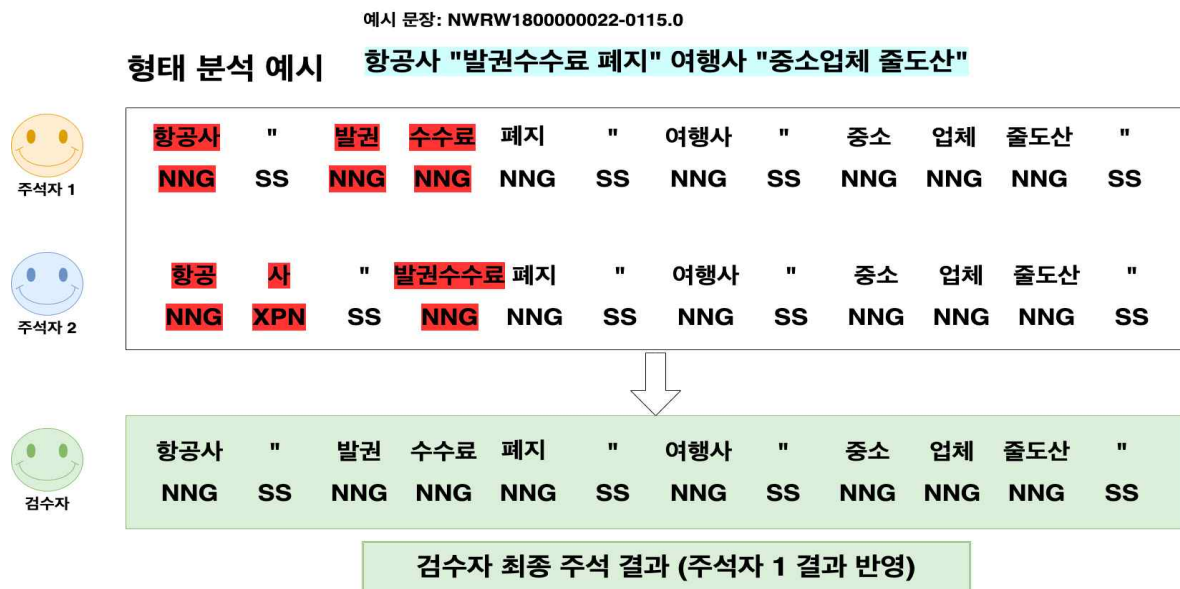


[그림 25] 2차 검수 대상 (형태 분석 예시)

검증용 말뭉치 검수 단계는 두 가지 검수 단계로 나누어서 진행하였다. 1차 검수는 서로 다른 두 주석 내용을 하나의 정답을 결정하기 위한 검수 과정이며([그림 26]), 2차 검수는 두 주석자가 같은 주석을 하였지만 제3의 답이 있는 경우를 검토하기 위한 검수 과정이다([그림 27]).

검증용 말뭉치 검수는 한 명의 검수자가 1차 검수와 2차 검수를 별도로 진행하였다. 검수 대상은 무작위로 할당되어 검수를 진행하였다. 층위별로 주석한 <표 8> ~ <표 14>의 내용을 검토하여 최종 주석 결과를 결정하였다.

검수 화면에서는 두 주석자의 주석 결과를 볼 수 있으며, 그 결과를 보고 검수자는 최종 주석 결과를 선택한다. 1차 검수의 경우 두 명의 주석자의 주석 결과가 다르므로 둘 중에 하나의 주석 결과를 선택하거나, 혹은 제3의 답을 선택할 수 있다. 2차 검수의 경우 두 명의 주석자가 서로 일치하기 때문에 제3의 답이 있는지 검토한다.



[그림 26] 1차 검수 예시 (형태 분석 예시)

형태 분석 예시 항공사 "발권수수료 폐지" 여행사 "중소업체 줄도산"



[그림 27] 2차 검수 예시 (형태 분석 예시)

일부 층위에 대해서는 지침의 변동 및 구축, 검수 작업 시에 오류로 인하여 검증용 말뭉치의 재검수를 수행하였다. 또한, 워크벤치 상에서 오류 혹은 기타 이유로 인한 검증용 말뭉치 누락분 또한 재작업을 통하여 보완하는 과정을 거쳐 검증 대상 분석 말뭉치의 7%에 해당하는 분량의 검증용 말뭉치를 구축하였다.

<표 16>, <표 17>은 분석 말뭉치 구축 규모 대비 검증용 말뭉치 구축 비율을 나타낸다. 어절 수를 기준으로 문어와 구어 모두 분석 말뭉치 구축 규모의 7%를 넘는 규모의 원시 말뭉치를 주석하였다.

	문서	문단	문장	어절
검증용 말뭉치 (A, 단위: 개)	520	4,438	10,400	140,633
분석 말뭉치 (B, 단위: 개)	7,265	62,440	150,085	2,000,215
비율 (A/B, 단위: %)	7.158	7.108	6.929	7.03

<표 16> 문어 분석 말뭉치 구축 규모 대비 검증용 말뭉치 구축 비율

	파일	발화	어절
검증용 말뭉치 합계 (A, 단위: 개)	32	17,551	70,744
분석 말뭉치 합계 (B, 단위: 개)	423	223,962	1,006,447
비율 (A/B, 단위: %)	7.601	7.878	7.03

<표 17> 구어 분석 말뭉치 구축 규모 대비 검증용 말뭉치 구축 비율

2.4. 검증용 말뭉치 수정

검증용 말뭉치는 검증 대상 말뭉치의 품질을 판단하는 비교 기준이 되는 말뭉치로 오류가 최소화되어야 한다. 따라서 이를 위해 검증용 말뭉치의 구축 및 검수 이후에 검증 대상 말뭉치와의 검증 과정에서 검증 대상 말뭉치 주석 결과와 비교 분석하여 검증용 말뭉치의 오류를 한 번 더 검토, 수정하는 단계를 거쳤다. 검증 결과에서 검증 대상 말뭉치와 검증용 말뭉치의 주석 결과가 다른 경우 검증 대상 말뭉치의 오류가 아닌 검증용 말뭉치가 오류가 아닌지 검토하여 검증용 말뭉치의 오류를 수정하였다.

① 형태 분석 검증용 말뭉치

형태 분석 검증용 말뭉치를 검토하기 위해 구축사업단이 구축한 형태 분석 검증 대상 말뭉치와 본 사업단이 구축한 검증용 말뭉치를 구축 단계별로 비교 분석하였다. 이 작업은 사전에 정의한 불일치 기준 3개 항목에 대하여 검토를 진행한 것으로, 불일치 기준이 되는 유형은 MORPHEME_LABEL_ERROR, MORPHEME_UNDERSPLIT, MORPHEME_OVERSPLIT이다.

우선 첫 번째 불일치 유형인 MORPHEME_LABEL_ERROR는 검증 대상 분석 말뭉치와 검증용 말뭉치에서 같은 형태에 대하여 서로 다른 표지를 주석한 경우를 말한다. 이 경우 필연적으로 두 말뭉치 중 최소한 한 개의 말뭉치는 잘못된 주석을 내리고 있는 것이기에 검증 말뭉치를 검토하기 위해 먼저 검토가 필요한 불일치 유형이다. 두 번째 불일치 유형인 MORPHEME_UNDERSPLIT은 같은 어절에 대하여 검증용 말뭉치보다 분석 말뭉치에서 더 많은 형태소를 분석한 것이며 MORPHEME_OVERSPLIT은 반대로 검증용 말뭉치에서 검증 대상 분석 말뭉치보다 더 많은 형태소를 분석한 것이다. 형태 분석은 기본적으로 하나의 어절을 몇 개의 형태로 분절하는지, 또 분절한 형태에 대하여 어떤 형태소를 주석하는지에 따라서 그 정확성을 알아볼 수 있다.

따라서 MORPHEME_LABEL_ERROR와 마찬가지로 MORPHEME_UNDERSPLIT, MORPHEME_OVERSPLIT 역시 검증용 말뭉치의 검토를 위하여 필수적으로 설정되어야 하는 불일치 유형이다.

본 사업단에서는 위 불일치 항목을 대상으로 수작업 검수를 진행하였으며 두 기준을 개별적으로 검수하기보다는 불일치가 발생한 문장을 분석하여 두 가지 불일치 사항을 통합적으로 검수하였다. 이 과정에서 검증 말뭉치의 오류로 판단되는 사항에 대해서는 지침에 따라 수정 사항을 반영하였다.

첫 번째 불일치 유형인 MORPHEME_LABEL_ERROR에서 나타나는 주요 오류 사례로는 먼저 ‘것’의 구어형인 ‘거’와 ‘이것, 그것, 저것’의 변이형인 ‘이거, 그거, 저거’의 주석 오류가 있다. 형태 분석 지침에서는 해당 단어들의 형태 분석에 대해서 아래 그림과 같이 제시하고 있다.

(3) ‘것’과 구어형 ‘거’의 분석	
‘거’의 형태를 그대로 인정하여 분석한다.	
[예시] 공부할 거를 준비해 왔니?	[거/NNB+를/JKO]
공부할 걸 가져왔니?	[거/NNB+ㄹ/JKO]
연습할 건 있니?	[거/NNB+ㄴ/JX]
먹을 게 모자라다.	[거/NNB+이/JKS]

[그림 28] 형태 분석 지침 ‘3.가.1).라).(3)’

다) 대명사의 이형태 분석	
(1) ‘이것, 그것, 저것: 이거, 그거, 저거’는 분석하지 않고 대명사로 인정한다. ‘~거’의 경우, ‘~거’의 형태를 그대로 인정하여 분석한다.	
[예시] 난 저거를 먹을래.	[저거/NP+를/JKO]
나는 여태 그걸 믿어 왔단다.	[그거/NP+ㄹ/JKO]

[그림 29] 형태 분석 지침 ‘3.가.2).다).(1)’

위 지침에 따라서 ‘거’, ‘이거, 그거, 저거’는 각각 ‘거/NNB’, ‘이거/NP, 그거/NP, 저거/NP’로 분석을 하여야 한다. 검증용 말뭉치에서 위 단어들을 틀리게 주석한 때도 있었기에 이를 아래의 그림과 같이 수정하였다.



[그림 30] 형태 분석 검증용 말뭉치 수정 - ‘거’

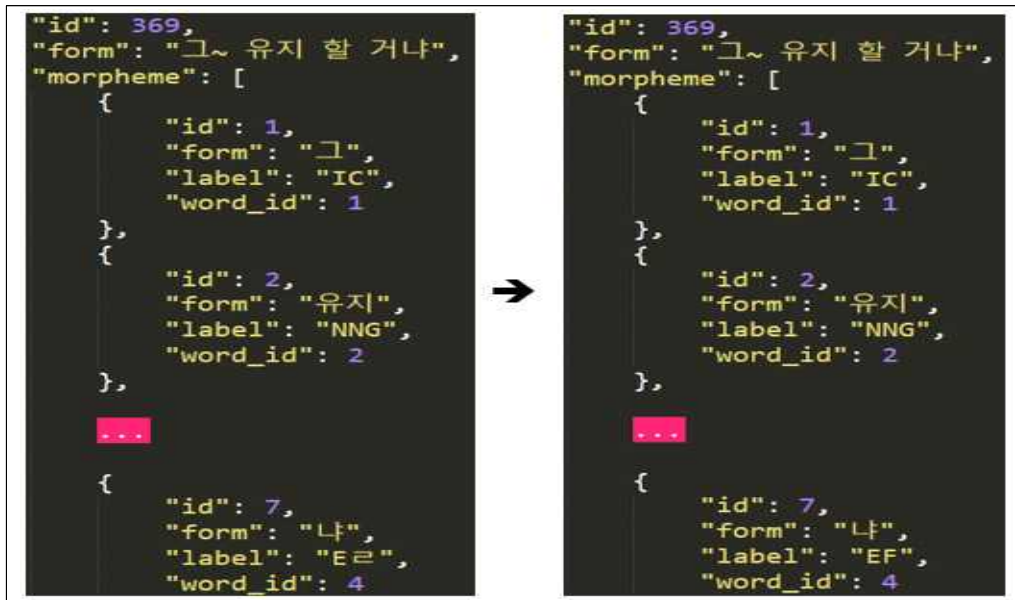


[그림 31] 형태 분석 검증용 말뭉치 수정 - ‘이거, 그거, 저거’

또 다른 오류 사례로는 지침에 규정되지 않은 주식 표지를 사용하는 때도 있었다. 해당 오류는 검증용 말뭉치 구축자가 작업 도중에 키보드의 ‘한/영’ 버튼을 잘못 누르는 등의 이유로 오타가 발생하여 나타난 것으로 보인다. ‘NN π ’, ‘E π ’, ‘두’ 등의 잘못 쓰인 표지들이 발견되었으며 아래의 그림과 같이 수정하였다.



[그림 32] 형태 분석 검증용 말뭉치 수정 - ‘NN π ’



[그림 33] 형태 분석 검증 말뭉치 수정 - ‘Eㄹ’

형태 분석 검증 대상 분석 말뭉치와 검증용 말뭉치에 대한 전반적인 검토 결과, 주요 불일치 유형은 두 말뭉치에서 같은 형태에 대하여 서로 다른 표지를 주석한 것과 같은 단어를 서로 다른 수의 형태로 분석한 것으로 나뉘었다. 이 중 같은 형태에 대하여 서로 다른 표지를 주석한 유형은 대부분 품사통용어를 대상으로 할 때 발생하였으며 같은 단어를 서로 다른 수의 형태로 분석한 유형은 접두사, 접미사나 어미 등을 대상으로 할 때 자주 나타났다.

② 어휘의미 분석 검증용 말뭉치

어휘의미 분석 검증용 말뭉치를 검토하기 위해 불일치 기준 3개 항목에 대하여 검토를 진행하였다. 불일치 기준이 되는 유형은 WSD_NUMBERRING_ERROR, WSD_FP_ERROR, WSD_FN_ERROR이다.

우선 첫 번째 불일치 유형인 WSD_NUMBERRING_ERROR는 검증 대상 분석 말뭉치와 검증용 말뭉치에서 같은 어휘에 대하여 서로 다른 의미 번호를 주석한 경우를 말한다. 이 경우 필연적으로 두 말뭉치 중 최소한 한 개의 말뭉치는 잘못된 주석을 내리고 있는 것이기에 검증용 말뭉치의 우선적인 검토가 필요한 불일치 유형이다. 두 번째 불일치 유형인 WSD_FP_ERROR는 검증용 말뭉치에서 의미 번호를 주석하지 않은 어휘에 대하여 검증 대상 분석 말뭉치에서 의미 번호를 주석한 경우이다. WSD_FN_ERROR는 이와 반대로 검증용 말뭉치에서 의미 번호를 주석한 어휘에 대하여 검증 대상 분석 말뭉치에서 의미 번호를 주석하지 않은 경

우이다. WSD_FP_ERROR와 WSD_FN_ERROR는 특정 고유명사나 합성어, 파생어의 어떤 형태를 어휘의미 분석의 대상으로 삼을지에 대한 구축사업단의 판단과 본 사업단의 판단이 달라서 발생한 것으로 보인다.

본 사업단에서는 위 불일치 항목을 대상으로 수작업 검수를 진행하였으며 세 기준을 개별적으로 검수하기보다는 불일치가 발생한 문장을 분석하여 두 가지 불일치 사항을 통합적으로 검수하였다. 이 과정에서 검증 말뭉치의 오류로 판단되는 사항에 대해서는 지침에 따라 수정 사항을 반영하였다.

첫 번째 불일치 유형인 WSD_NUMBERRING_ERROR에서 나타나는 주요 오류 사례로는 먼저 검증용 말뭉치에서 특정 어휘에 대하여 잘못된 의미 번호를 주석한 것이 있다. 그 예시는 아래의 그림과 같다.

```
"id": "NWRW1800000022-0021.1",
"form": "(전략) ... 내세워 시민들에 개방 꺼려 ... (후략)",
"WSD": [
  ...
  {
    "word": "개방",
    "sense_id": 5,
    "begin": null,
    "end": null,
    "word_id": 14
```

[그림 34] 어휘의미 분석 오류 - 의미 번호 오분석



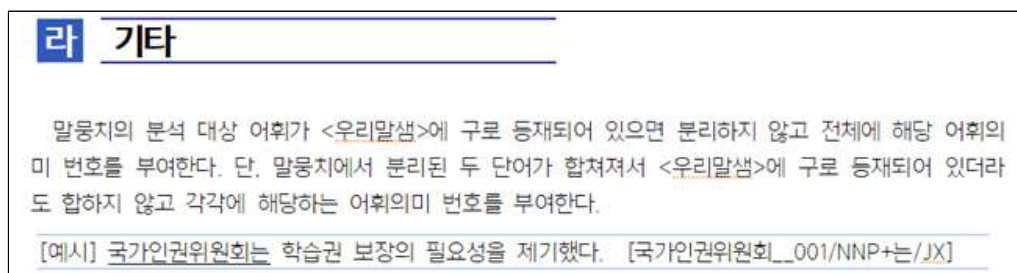
[그림 35] 어휘의미 분석 오류 - ‘개방’의 <우리말샘> 검색 결과

위 [그림 34]에서의 본문은 ‘이러저러한 이유로 옥상 공원을 시민에게 개방하기를 꺼리고 있다’ 라는 내용을 담고 있다. 따라서 <우리말샘>에서 찾아볼 수 있는 ‘개방’의 여러 가지 의미 중에서 ‘어떠한 공간 따위를 열어 자유롭게 드나들고 이용하게 한다’는 4번 의미가 ‘금하거나 경계하던 것을 푼다’라는 5번 의미보다 본문에 사용된 ‘개방’의 의미에 가깝다. 본 사업단에서는 이에 따라 해당 어휘들의 의미 번호를 수정하여 주석하였다.



[그림 36] 어휘의미 검증 말뭉치 수정 - ‘개방’

두 번째 불일치 유형인 WSD_FP_ERROR와 세 번째 불일치 유형인 WSD_FN_ERROR에서 나타나는 주요 오류 사례로는 <우리말샘> 구 등재어와 관련된 것이 있다. 어휘의미 분석 지침에서는 구로 등재된 어휘에 대하여 아래 그림과 같이 밝히고 있다.



[그림 37] 어휘의미 분석 지침 ‘3.라’

위의 지침과 달리 검증용 말뭉치에서는 <우리말샘>에 구로 등재된 일부 어휘를 각각 분리하여 의미 번호를 주석하고 있었기에 이를 수정하여 주석하였다.

• 헌법재판소「001」『법률』 법령의 위헌 여부를 일정한 소송 절차에 따라 심판하기 위하여 설치한 특별 재판소. 법원의 제청에 의한 법률의 위헌 여부, 탄핵, 정당의 해산, 국가 기관 상호 간 또는 국가 기관과 지방 자치 단체 간 및 지방 자치 단체 상호 간의 권한 쟁의, 헌법 소원에 관한 것을 심판한다.

[그림 38] ‘헌법재판소’의 <우리말샘> 검색 결과



[그림 39] 어휘의미 검증 말뭉치 수정 - 구 등재어

검증 대상 분석 말뭉치와 검증용 말뭉치에 대한 전반적인 검토 결과, 주요 불일치 유형은 두 말뭉치에서 같은 어휘에 대하여 서로 다른 의미 번호를 주석한 것과 한 말뭉치에서는 주석한 어휘를 다른 말뭉치에서는 주석하지 않은 것으로 나뉘었다. 이 중 같은 어휘에 대하여 서로 다른 의미 번호를 주석한 유형의 경우, 각 어휘의 의미 번호 판단 기준이 지침에 상세히 기술되어 있음에도 불구하고 말뭉치를 구축하는 작업자들의 직관에 따라서 어휘의미 주석 결과가 달라졌기 때문으로 보인다. 또한, 한 말뭉치에서 주석한 어휘를 다른 말뭉치에서는 주석하지 않은 불일치 유형의 경우, 작업자가 <우리말샘>에 등재된 구표제어를 제대로 검색하지 못하거나 합성어나 파생어, 고유명사에 대한 지침 이해도가 부족하여 잘못된 주석을 내린 경우가 많았다.

③ 개체명 분석 검증용 말뭉치

개체명 분석 검증용 말뭉치를 검토하기 위해 불일치 기준 4개 항목에 대하여 검토를 진행하였다. 불일치 기준이 되는 유형은 NER_BIO_SPAN_ERROR, NER_FN_ERROR, NER_FP_ERROR, NER_TYPE_ERROR이다.

검증용 말뭉치와 검증 대상 분석 말뭉치에서 개체의 범위를 다르게 주석한

NER_BIO_SPAN_ERROR의 경우, 많은 경우가 지침 해석의 차이에서 기인한 것으로 보인다. 최소 단위 태깅 원칙과 최장 개체명 태깅에서 서로 다른 기준을 적용한 경우이다. 이럴 때 개체의 분류를 명확히 결정하기 어려운 유형들에 대해서 다수 발견되었다.

수식어의 포함 여부에 따라 개체 태그가 달라지는 경우가 대표적이다. 즉, 과소 분할 또는 과대 분할에 따라 태그 부여 기준이 달라지는 경우가 많았다.

또한, 본 사업에서 사용한 구축지침은 개체명 주석 말뭉치 구축을 지향하기보다는 개체 주석 말뭉치 구축에 더 가깝다고 볼 수 있다. 이는 인명, 기관, 장소, 사건 등 일반적인 (First-order) 개체명뿐만 아니라 수량, 시간, 전문용어, 이론까지 광범위하게 포함하기 때문이다. 이 경우에 총칭(Generic) 등 일반적으로는 개체명으로 취급되지 않을 수 있는 일반 명사류 또한 주석 범위로 포함할 것인지 문제가 발생할 수 있는데, 이는 추후 지침에서 보완되어야 할 사항일 것이다. 이러한 문제는 백과사전적 개체의 주석 누락으로 이어질 수 있다.

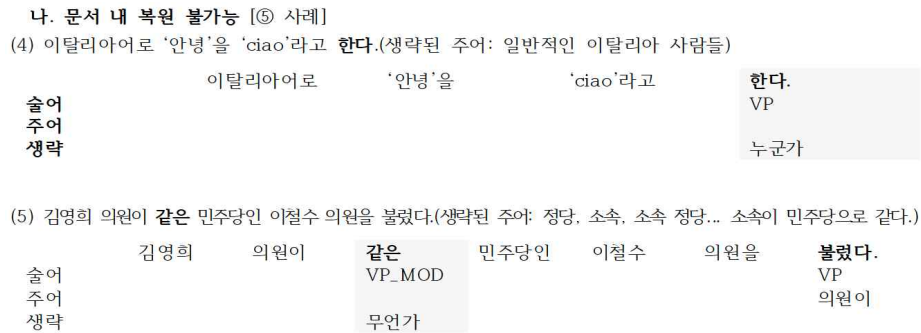
④ 주격 무형대용어 복원 검증용 말뭉치

주격 무형대용어 복원 검증용 말뭉치를 검토하기 위해 불일치 기준 3개 항목에 대하여 검토를 진행하였다. 불일치 기준이 되는 유형은 ZEROANAPHORA_ANTECEDENT_ERROR, ZEROANAPHORA_PREDICATE_OVER, ZEROANAPHORA_PREDICATE_MISSED 세 가지이다.

우선 첫 번째 불일치 유형인 ZEROANAPHORA_ANTECEDENT_ERROR는 검증 대상 분석 말뭉치와 검증용 말뭉치에서 동일 서술어에 서로 다른 선행사를 복원한 경우를 말한다. 이럴 때 필연적으로 두 말뭉치 중 최소한 한 개의 말뭉치는 잘못된 주석을 내리고 있는 것이기에 검증 말뭉치를 검토하기 위하여 먼저 필요한 불일치 유형이다. 두 번째 불일치 유형인 ZEROANAPHORA_PREDICATE_OVER는 분석 말뭉치에서 복원 대상이 아닌 술어에 주석한 경우이고 반대로 ZEROANAPHORA_PREDICATE_MISSED는 분석 말뭉치에서 복원 대상 술어를 제외한 것이다. 주격 무형대용어 복원은 기본적으로 TTA 의존 구문 분석 결과에서 지배하는 주어가 없는 술어 대상으로 생략된 주어를 복원하는 것으로 복원 대상인 술어를 판별하는 것이 중요하다. 따라서 불일치 유형 중에 ZEROANAPHORA_PREDICATE_OVER, ZEROANAPHORA_PREDICATE_MISSED는 검증 말뭉치의 검토를 위하여 필수적으로 설정되어야 하는 불일치 유형이다.

본 사업단에서는 위 불일치 항목을 대상으로 수작업 검토를 진행하였으며 두 기준을 개별적으로 검수하기보다는 불일치가 발생한 문장을 분석하여 두 가지 불일치 사항을 통합적으로 검수하였다. 이 과정에서 검증 말뭉치의 오류로 판단되는 사항에 대해서는 지침에 따라 수정 사항을 반영하였다.

첫 번째 불일치 유형인 ZEROANAPHORA_ANTECEDENT_ERROR에서 나타나는 주요 오류 사례로는 선행어에 해당하는 주어가 있지만 영주어로 복원한 오류가 있다. 특히 이러한 사례는 해당 문장에 선행어가 없지만 동일 문서 내에 주어가 있는 경우이다. 주격 무형대용어 분석 지침에서는 문서 내 복원 불가능한 주격을 복원하는 것을 아래 그림과 같이 제시하고 있다.



[그림 40] 주격 무형대용어 복원 지침 ‘2.나.(4), (5)’

위 지침에 따라서 문서 내 복원이 불가능한 것은 ‘누군가, 무언가’로 복원함을 파악할 수 있다. 검증 말뭉치에서 위 단어들을 틀리게 주석한 때도 있었기에 이를 아래의 그림과 같이 수정하였다.



[그림 41] 주격 무형대용어 분석 말뭉치 수정 ‘누군가’

두 번째 불일치 유형인 ZEROANAPHORA_PREDICATE_OVER에서 나타나는 주요 오류 사례로는 검증 대상 분석 말뭉치에서 주어와 서술어가 구문 분석에서 지배 관계를 가져 복원하지 않아도 되는 경우인데 복원한 경우이다. 검증용 말뭉치에서는 일부 오류 사례가

있었는데 복원 유형인 VP_MOD 중에서 주격을 복원하지 않은 사례가 있다. 주격 무형대용어 분석 지침에서는 서술어에 해당하는 VP_MOD를 복원하나 일부 서술어에 해당하지 않는 VP_MOD를 아래와 같이 정의한다.

ㄷ. 서술어에 해당하지 않는 VP(_MOD)(생략어 지침 7쪽):

- 관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고, 그러면서 등
모문과 분리되어 단독 문장을 이루지 못하는 술어

<예> 전기 요금이 오른 데에 이어 수도 요금이 올랐다. → *[전기 요금이 오른 데에 {있는다}이었다.]

[그림 42] 주격 무형대용어 복원 지침 ‘1.나.(ㄷ)’

위 지침에 따라서 서술어 대상이 아닌 것을 ‘관해, 대해, 의해, 향해, 통해, 따라, 아니라, 아니라, 불구하고, 그러면서’ 등으로 정의함을 파악할 수 있다. 검증용 말뭉치에서 해당 서술어지만 복원하지 않은 사례를 아래 그림과 같이 수정하였다.

# text = 재판부는 그러나 "PD수첩이	# text = 재판부는 그러나 "PD수첩이
0 재판부는 - - - -	0 재판부는 - - - -
1 그러나 - - - -	1 그러나 - - - -
2 "PD수첩이 - - - -	2 "PD수첩이 - - - -
3 촛불시위를 - - - -	3 촛불시위를 - - - -
4 유도했다는 -1 -1 -1 누군가	4 유도했다는 - - - -
5 증거가 - - - -	5 증거가 - - - -
6 없다"고 - - - -	6 없다"고 - - - -
7 밝혔다. - - - -	7 밝혔다. - - - -

[그림 43] 주격 무형대용어 분석 말뭉치 수정

세 번째 불일치 유형인 ZEROANAPHORA_PREDICATE_MISSED에서 나타나는 주요 오류 사례로는 검증 대상 분석 말뭉치에서 주어와 서술어가 구문 분석에서 지배 관계가 없기에 복원해야 하는데 복원하지 않은 경우이다. 검증용 말뭉치에서 일부 오류 사례가 있었는데 지배 관계가 있는데 주격을 복원한 사례가 있다. 주격 무형대용어 분석 지침에서는 복원 대상 서술어를 아래와 같이 정의한다.

1. 생략 술어 탐지 및 선행어 결정

가. 주어가 생략된 술어 탐지:

‘TTA 의존 구문 분석’ 결과, 지배하는 주어가 없는 VP를 대상으로 생략된 주어를 문서 내에서 복원

[그림 44] 주격 무형대용어 복원 지침 ‘1.가’

위 지침에 따라 TTA 의존 구문 분석을 바탕으로 생략 서술어가 아닌 것은 아래 그림과 같이 수정하였다. 아래 사례는 ‘PD수첩이’의 지배소는 ‘유도했다는’으로 지배하는 주어가 있기에 주격 복원은 필요하지 않다.

주격 무형대용어 분석 검증 대상 분석 말뭉치와 검증용 말뭉치에 대한 전반적인 검토 결과, 주요 불일치 유형은 두 말뭉치에서 주어 복원 대상이 아닌 술어와 복원 술어 대상을 제외한 것으로 나타났다. 두 가지 유형은 의존 구문 분석 오류에서 나타난다. 구문 분석 지침에 따라 주술 호응이 결정되는데 복문 해석이 중의성을 가질 때 다른 구문 분석 결과가 나타난다. 이는 복원 대상 서술어의 불일치를 기인한다. 이외에 분석 말뭉치와 검증 말뭉치의 복원한 주격이 다른 경우는 문서 내에 복원할 수 있는 주어를 영주어로 복원한 경우이다.

⑤ 상호참조 해결 검증용 말뭉치

상호참조 해결 검증용 말뭉치를 검토 작업은 검증용 말뭉치 내부에서 검출된 오류인 CR_DUPLICATE_ERROR에 대해서 검토를 수행하였다. CR_DUPLICATE_ERROR는 같은 개체를 서로 다른 상호참조 관계로 판단한 경우를 말한다. 이럴 때 잘못된 주석을 내리고 있는 것이기에 검증용 말뭉치에서 먼저 검토가 필요한 유형이다. 본 사업단에서는 위 불일치 항목을 대상으로 수작업 검수를 진행하였으며 중복되는 개체를 해당하는 상호참조 관계로 결정하였다. 아래는 해당 사례들이다.

➤ 일반명사 외의 복합명사, 대명사 등도 멘션으로 정의한다.

예1)

<예문> 노벨 평화상(etri저집 예시_29쪽)

<보기> [노벨](0), [노벨 평화상](0)

<오류>

<상호참조 태깅> {Ø}(예문 내 멘션들 중 상호참조 관계 없음)

예2) 대명사 멘션의 예

<예문> 이것은 여우입니다.(자체 생성 예시문)

<보기> [이것](0), [여우](0)

<오류>

<상호참조 태깅> {[이것], [여우]}(0)

[그림 45] 상호참조 해결 지침 ‘2.1.1’

위 지침에 따라서 일반명사 외의 복합명사, 대명사 등도 멘션으로 정의한 것을 알 수 있다. 검증용 말뭉치에서 대명사의 상호참조 관계가 중복하여 주석한 때도 있었기에 이를 수정하였다. 아래 그림에서 c_id 1의 word_id 22 ‘제’는 해당 상호참조 관계가 아니기에 삭제하였다. 해당 word_id는 맥락상 c_id 27에 해당한다. 대명사 ‘제’의 지칭하는 대상이 다르기에 맥락 파악이 필요하였다.

```

---Error---
말뭉치 SRBK00152 - 문서 SRBK00152에서 오류
CR_Duplicate_Mention 발생: (중복 멘션) sentence_id: 172,
word_ids: [0], "제"가 2개 발생
c_id: 1, m_id: 22
c_id: 27, m_id: 7

--- c_id: 1 ---
0: sentence_id: 12, word_ids: [3], "대표님"
1: sentence_id: 16, word_ids: [0], "저"
2: sentence_id: 18, word_ids: [1], "저"
3: sentence_id: 18, word_ids: [2], "제"
4: sentence_id: 46, word_ids: [2], "대표님"
5: sentence_id: 54, word_ids: [1], "저"
6: sentence_id: 54, word_ids: [2], "제"
7: sentence_id: 58, word_ids: [0], "내"
8: sentence_id: 60, word_ids: [0], "제"
9: sentence_id: 72, word_ids: [1], "저"
10: sentence_id: 75, word_ids: [1], "제"
11: sentence_id: 92, word_ids: [2], "손학규"
12: sentence_id: 92, word_ids: [2, 3], "손학규 대표님"
13: sentence_id: 96, word_ids: [5], "대표님"
14: sentence_id: 107, word_ids: [8], "대표님"
15: sentence_id: 110, word_ids: [7], "저"
16: sentence_id: 112, word_ids: [0], "대표님"
17: sentence_id: 114, word_ids: [0], "저"
18: sentence_id: 116, word_ids: [4], "저"
19: sentence_id: 125, word_ids: [2], "제"
20: sentence_id: 133, word_ids: [0], "제"
21: sentence_id: 143, word_ids: [4], "대표님"
22: sentence_id: 172, word_ids: [0], "제"

--- c_id: 27 ---
0: sentence_id: 1, word_ids: [1], "김"
1: sentence_id: 1, word_ids: [1, 2], "김 선생"
2: sentence_id: 2, word_ids: [1], "제"
3: sentence_id: 138, word_ids: [3, 4, 5, 6], "우리 김어
김어준 선생"
4: sentence_id: 138, word_ids: [5], "김어준"
5: sentence_id: 142, word_ids: [2], "저"
6: sentence_id: 149, word_ids: [1], "제"
7: sentence_id: 172, word_ids: [0], "제"
8: sentence_id: 172, word_ids: [1], "제"
9: sentence_id: 331, word_ids: [2, 3, 4, 5], "우리 저
김어준 선생"
10: sentence_id: 331, word_ids: [4], "김어준"
11: sentence_id: 580, word_ids: [0], "제"
12: sentence_id: 591, word_ids: [0, 1, 2], "우리 김어준
선생"
13: sentence_id: 591, word_ids: [1], "김어준"

```

[그림 46] 상호참조 해결 말뭉치 수정 ‘제’

두 번째 사례로 같은 개체를 다른 개체로 판단하여 서로 다른 군집에 포함한 경우로 아래와 같다. 상호참조 해결 지침에서 멘션의 의미 단위 정의를 아래와 같이 정의한다.

2.1.5. 멘션의 의미 단위 정의

- 멘션 의미의 최소단위는 개체명 정보, 개체명이 없는 경우 단어(형태)가 최소단위이다.

예1)

<예문> 아름다운 아메리카 플로리다주(etri 지권 예시_31 쪽)
 <보기> [아메리카](0), [아름다운 아메리카 플로리다주](0)
 <오류> [아름다운 아메리카](X)
 <상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 알프스 산맥(etri 지권 예시_31 쪽)
 <보기> [알프스](0), [알프스 산맥](0)
 <오류> [산맥](X)
 <상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

[그림 47] 상호참조 해결 지침 ‘2.1.5’

위 지침에 따라서 일반명사 ‘자신’은 개체명, 대명사도 아니나 단어(형태)가 최소 단위로 개체를 나타낸다. 이러한 개체 ‘자신’은 서로 다른 군집에 중복된 것으로 나타났기에 해당 군집에 속하도록 수정하였다.

```
---Error---
말뭉치 NWRW1800000032-0311 - 문서 NWRW1800000032-0311에서 오류
CR_DUPLICATE_MENTION 발생: (중복 멘션) sentence_id: 12,
word_ids: [0], "자신"가 2개 발생
c_id: 1, m_id: 10
c_id: 9, m_id: 4

--- c_id: 1 ---
0: sentence_id: 0, word_ids: [0, 1, 2, 3], "대표팀 첫 훈련한
문태중"
1: sentence_id: 2, word_ids: [0, 1, 2, 3, 4], "왼쪽 가슴에
태극마크를 단 문태중"
2: sentence_id: 3, word_ids: [0], "그"
3: sentence_id: 5, word_ids: [0, 1, 2, 3, 4, 5, 6, 7],
"최근 법무부로부터 체육분야 우수인재로 선정돼 특별귀화를 허가받은 문태중"
4: sentence_id: 6, word_ids: [0, 1, 2, 3, 4, 5, 6],
"한국인 어머니와 미국인 아버지 사이에서 태어난 그"
5: sentence_id: 8, word_ids: [0], "그"
6: sentence_id: 9, word_ids: [6], "내"
7: sentence_id: 9, word_ids: [14], "나"
8: sentence_id: 10, word_ids: [10], "문태중"
9: sentence_id: 11, word_ids: [0], "그"
10: sentence_id: 12, word_ids: [0], "자신"
11: sentence_id: 13, word_ids: [9], "문태중"
12: sentence_id: 16, word_ids: [0, 1, 2, 3], "첫 훈련을 마친
문태중"

--- c_id: 9 ---
0: sentence_id: 10, word_ids: [16, 17, 18, 19], "이번에
함께 귀화한 동생"
1: sentence_id: 10, word_ids: [16, 17, 18, 19, 20],
"이번에 함께 귀화한 동생 문태영"
2: sentence_id: 11, word_ids: [1], "동생"
3: sentence_id: 11, word_ids: [9], "동생"
4: sentence_id: 12, word_ids: [0], "자신"
```

[그림 48] 상호참조 해결 말뭉치 수정

구체적으로 위에서 c_id 9의 word_id 4 ‘자신’을 삭제하였다. 이는 문맥상 c_id 1 문태종의 ‘자신’이기 때문이다. c_id 9는 ‘문태종의 동생 문태영’의 군집이다. c_id 1의 word_id 10은 개체 유지를 하였다.

세 번째 사례는 검수 말뭉치에서 같은 개체들로 이루어진 똑같은 군집이 있는 오류 유형이다. 해당 유형은 똑같은 군집이기에 하나를 삭제해야 한다.

2.1.16. 숫자, 날짜 및 수량 표현(지침 추가 사항)

- 날짜, 금액, 수치 등 숫자 표현들을 멘션 추출 대상에 포함한다.

[그림 49] 주격 무형대용어 복원 지침 ‘2.1.16’

위 지침에 따라 숫자 표현도 멘션 추출 대상이 된다. 아래 사례는 ‘대기업의 내년 1월 업황 전망 지수’, ‘88’이 하나의 상호참조 관계이다. c_id 7, c_id 8이 중복된 군집으로 c_id 8 군집 자체를 삭제하였다.

```
---Error---
말뭉치 NWRW1800000032-0082 - 문서 NWRW1800000032-0082에서 오류
CR DUPLICATE_MENTION 발생: (중복 멘션) sentence_id: 7,
word_ids: [1, 2, 3, 4, 5, 6], "대기업의 내년 1월 업황 전망
지수"가 2개 발생
c_id: 7, m_id: 0
c_id: 8, m_id: 0

--- c_id: 7 ---
0: sentence_id: 7, word_ids: [1, 2, 3, 4, 5, 6], "대기업의
내년 1월 업황 전망 지수"
1: sentence_id: 7, word_ids: [7], "88"

--- c_id: 8 ---
0: sentence_id: 7, word_ids: [1, 2, 3, 4, 5, 6], "대기업의
내년 1월 업황 전망 지수"
1: sentence_id: 7, word_ids: [7], "88"
```

[그림 50] 상호참조 해결 말뭉치 수정

상호참조 해결의 검증 말뭉치에서 하나의 멘션이 다른 개체로 판단한 경우를 전반적으로 검토하였다. 특정 오류 중심으로 검토했지만, 검증용 말뭉치의 오류를 줄였다는 것에 의의를 가진다. 중복 오류가 나타난 사례를 분석해보니 멘션 간의 거리가 멀었던 것이 동일 멘션을 다른 군집에 포함하는 오류가 나타나는 것으로 나타났다. 이는 문어와 구어에 같은 특성으로 나타난다. 중복 사례 오류는 주식 작업 도중에 찾기 어려운 유형으로 판단되므로 검수 과정에서 중복 사례를 해결하는 절차가 필요하다.

⑥ 구문 분석 검증용 말뭉치

구문 분석 검증용 말뭉치를 검토하기 위해 불일치 기준 2개 항목을 바탕으로 분석 말뭉치와 검증 말뭉치에서의 단위 주석이 일치하지 않을 때 대해 중점적인 검토를 진행하였다. 불일치 기준이 되는 유형은 DEPENDENCY_HEAD_ERROR, DEPENDENCY_LABEL_ERROR이다.

우선 첫 번째 불일치 기준인 DEPENDENCY_HEAD_ERROR는 분석 말뭉치와 검증 말뭉치의 동일 문장 주석을 비교했을 때 기준 어절의 지배소가 다르게 주석되어 그 분석이 일치하지 않는 경우를 말한다. 두 번째 불일치 기준 DEPENDENCY_LABEL_ERROR는 기준 어절의 구문 또는 기능 태그(레이블)가 일치하지 않는 경우를 말한다. 이 두 기준은 각각 의존 구문 분석의 평가 기준인 LAS와 UAS에 대응한다.

이후 검출된 불일치 항목을 대상으로 수작업 검수를 진행하였다. 문장의 하위 단위에 대한 분석이 문장 전체의 수형도 구조를 결정하는 것과 동떨어질 수 없는 구문 분석의 특징을 고려하여, 두 기준을 개별적으로 검수하기보다는 불일치가 발생한 문장 전체를 검토하여 두 가지 불일치 사항을 통합적으로 검수 및 교정하였다. 이 과정에서 검증 말뭉치의 오류로 판단되는 사항에 대해서는 지침에 따라 그 수정 사항을 반영하였다.

주요 오류 사례로는 복문을 분석할 때 문장 첫머리에 나타나는 주어와 같이 문장 성분의 지배소를 인용절 또는 모문의 서술어로 잘못 할당하는 경우, 신문 기사 제목과 같이 축약된 표현이나 머리기사에서 주로 나타나는 단독 어절 기호 또는 칼럼 명 등을 분석하는 과정에서 문장의 직접 성분으로 볼 수 없는 어절의 지배소를 잘못 할당하는 경우, 서술성 명사에 대해 VP를 할당한 경우, 쉼표로 잇달아 열거되는 명사구 연쇄의 분석, 격조사가 생략된 명사구에 대한 기능 태그 주석 불일치 등으로 요약할 수 있다.

구문 분석 오류의 경우 문맥에 따라 분석이 달라질 수 있는 특성상, 검수 과정을 자동화할 수 없다는 어려움이 있었다. 이에 불일치가 발생한 개별 문장을 직접 검수하는 과정을 통해 검증용 말뭉치의 분석 품질을 개선하는 데에 작업의 초점을 두었다. 이 과정에서 일부 오류에 대한 교정은 지침을 근거로 아래와 같이 수정하였다. 아래 예시는 명사구 열거 분석 오류에 관한 교정 사례를 보인 것이다.

1	이날	2	NP
2	연주회에서	18	NP_AJT
3	남도	4	NP
4	설장구	5	NP
5	가락	6	NP_CNJ
6	대금과	7	NP_CNJ
7	피아노	8	NP
8	연주	9	NP_CNJ
9	무용	10	NP
10	금척무	11	NP_CNJ
11	가야금	12	NP
12	연주곡	13	NP
13	침향무	14	NP_CNJ
14	타악	15	NP
15	퍼포먼스	16	NP
16	아리랑	17	NP
17	등이	18	NP_SBJ
18	선보인다.	-1	VP

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

[그림 51] 구문 분석 말뭉치 수정 예시 - 명사구 열거 오분석

특히 신문 기사 제목에서 자주 나타나는 글머리 기호나 특수 표지의 경우 지침에서 명확한 처리 방법을 제시하지 않아 단순히 명사구 연쇄의 처리 지침에 따라 분석하거나, 문장의 루트에 할당하는 등의 분석 차이를 보인 경우가 많았다. 신문 기사에서 자주 나타나는 직접 인용절에 대해서도 인용절 바깥의 선행 성분을 인용절 내부의 성분에 연결하는 등 지침에서 처리 방법을 분명히 제시하지 않아 작업자마다 주관적인 분석을 적용하는 경우가 많았다.

# sent_id = NWRW1800000022-0321.1			
# text = [대구·경북] 우리문화를 아끼고 사랑하는			
1	[대구·경북]	2	NP
2	우리문화를	3	NP_OBJ
3	아끼고	4	VP
4	사랑하는	5	VP_MOD
5	사람들이	6	NP_SBJ
6	모였다; 국제로타리	7	NP
7	3700지구	8	NP
8	내	9	NP_MOD
9	아리랑로타리클럽	-1	NP

→

# sent_id = NWRW1800000022-0321.1			
# text = [대구·경북] 우리문화를 아끼고 사랑하는			
1	[대구·경북]	9	NP
2	우리문화를	3	NP_OBJ
3	아끼고	4	VP
4	사랑하는	5	VP_MOD
5	사람들이	6	NP_SBJ
6	모였다; 국제로타리	7	VP
7	3700지구	8	NP
8	내	9	NP_MOD
9	아리랑로타리클럽	-1	NP

[그림 52] 구문 분석 말뭉치 수정 예시 - 지배소 할당 및 구문 태그 할당 오류

전반적인 검토 결과 구문 분석에서의 주요 불일치 유형으로는 복문 구조 또는 신문 기사의 제목이나 머리 기사를 분석할 때 단위 어절의 지배소를 잘못 할당하는 사례가 자주 나타났다. 이들 오류는 지침이 그 분석 방법을 명확히 제시하지 못하고 있어 작업자의 분석이 일관적이지 못한 경향을 보였다. 또한, 문장 내에서 명사가 연쇄될 경우 명부류 또는 명관류 등 품사통용어를 비롯한 특정 성분에서 격조사가 생략된 경우에도 기능 태그 주석

에 일부 차이를 보였는데, 이는 신문 기사 특유의 압축적 서술에 기인한 오류로 볼 수 있다. 한 문장 내에 문법 형태소 등이 광범위하게 빠진 한편 복문 구조가 복잡해지면 복잡해질수록 통사적 중의성을 명확하게 해소하기 어려워 주석 간의 불일치가 발생하는 경향을 보였다.

⑦ 의미역 분석 검증용 말뭉치

의미역 분석 검증용 말뭉치를 검토하기 위해 불일치 기준 2개 항목을 바탕으로 두 말뭉치에서의 주석 단위가 불일치한 경우에 대해 중점적인 검토를 진행하였다. 불일치 기준이 되는 유형은 SRL_PI_ERROR와 SRL_AI_ERROR이다.

우선 첫 번째 불일치 기준인 SRL_PI_ERROR는 분석 말뭉치와 검증 말뭉치의 동일 문장 주석을 비교했을 때 프레임 또는 프레임의 의미 번호 분석이 일치하지 않는 경우를 말한다. 대부분은 의사보조용언 구성, 부정 술어, 문장 부사에 대해 격틀을 부여한 경우였고, 의미 번호의 경우에는 주석 기준 사전이 되는 Korean PropBank, ETRI PropBank, U-PropBank, 우리말샘에서의 의미번호가 잘못 주석된 경우에 해당했다. 두 번째 불일치 기준 SRL_AI_ERROR는 해당 프레임의 의미역 주석이 잘못된 경우로, 필수격(ARGx)의 격 번호가 일치하지 않거나, 부가격에 대한 주석이 일치하지 않는 경우에 해당했다.

특히 서술어에 대한 의미 번호 결정 문제에서 발생한 불일치 사항은 후속되는 격틀 논항 분석의 불일치로 광범위하게 이어지는 모습을 보였는데, 어떠한 용언이 격틀 주석의 대상인가 아닌가에 따라 그 모든 논항의 주석 여부가 종속되기 때문이다. 말뭉치 오류 검증 작업에서는 앞서 검증 말뭉치 구축 당시 지침 변동 사항을 수시로 반영한 점을 고려하여, 아래 표에 보이는 바와 같이 문법적 기능만을 담당할 것으로 볼 수 있는 대상 격틀을 기준으로 집중적인 검토 작업을 수행하였다.

의하.01, 관하.01, 통하.01, 대하.01, 위하.01, 비하.01, 따르.02, 인하.02, 불구하.01, 비롯하.01, 더불.01, 말미암.01, 그러.01, 이러하.01, 그리하.01, 이렇.01, 어떻.01, 그러하.01, 어떠하.01, 앓.01, 그렇.01, 있.01, 보.0X, 하.0X, 되.0X, 있.0X, 가.0X, 오.0X, 두.0X, 같.01, 잇.01, 없.01 등

<표 18> 의미역 분석 말뭉치에서의 집중 검토 대상 서술어 목록

이들 목록은 단독 어절을 구성할 수 없는 ‘~에 대하여’, ‘~과 관하여’, ‘~에 따르면’, ‘~으로 인해’, ‘~에도 불구하고’, ‘~을 비롯하여’, ‘~과 더불어’, ‘~에 말미암아’ 등에 대하여 프레임을 주석할 때 자주 나타나는 대상이다. 또한 ‘이렇듯’, ‘그러나’, ‘그리고’ 와 같이 문장 부사로 사용된 성분에 대해 격틀 주석을 한 경우를 검출하기 위해 이러하.01, 그리하.01, 이렇.01, 어땡.01, 그리하.01, 어떠하.01, 앓.01, 그렁.01 등의 격틀 주석을 집중 검토하였다. 그 밖에 ‘-르(을) 수/리 있/없다’ 나 ‘-르(을)/ㄴ(은) 것 같다’ 와 같은 보조 용언 및 의사 보조 용언 구성에 자주 사용되는 앓.01이나 없.01 등의 부정 술어, 있.01, 보.0X, 하.0X, 되.0X, 가.0X, 오.0X, 두.0X, 같.01 등의 격틀도 집중 검토 대상에 포함하였다.

철도노조의	반대에도	불구하고	철도	민영화	사업을	강행하기로
		ARGM-CND			ARG1	강행_4444401
	ARG1	불구하고_4444401				

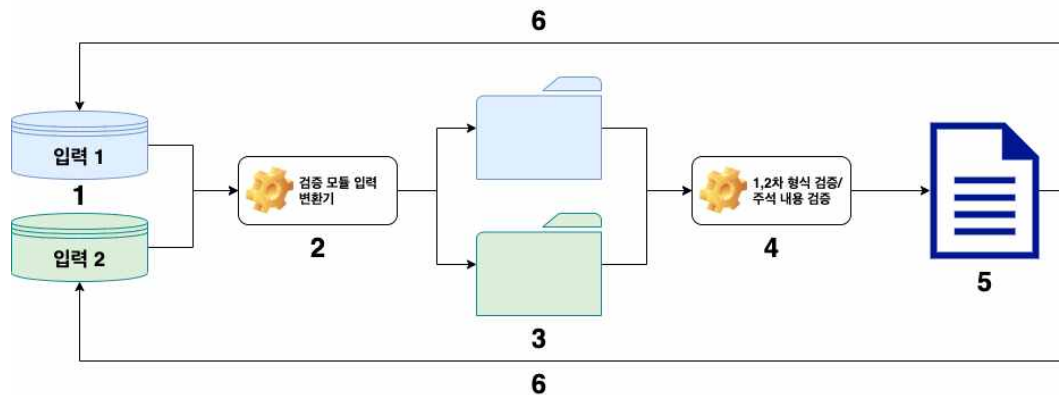
그렇다고	제가	일방적으로
그렇_4444401		
	ARG0	ARGM-PRD

〈표 19〉 의미역 분석 말뭉치에서의 주요 불일치 사례

전반적인 검토 결과 의미역에서의 주요 불일치 사항은 대상 술어에 대한 격틀 주석 여부에 따라서 귀속된 모든 논항의 주석 사항이 폭넓게 달라지는 문제에서 기인하였다. 대다수의 불일치 사례는 지침이 주석 대상 범위를 뚜렷하게 제시하지 못하고 있어 발생한 것인데, 앞서 검증 말뭉치의 구축 과정에서 말뭉치 전반에 대한 검토가 수행되었고, 이로 인해 프레임의 의미 번호와 논항 주석 결과가 폭넓게 검수가 진행된 점을 고려하여 최종 교정 단계에서는 문법적 기능만을 수행하는 동사 및 형용사에 대해 재검수 및 교정하는 작업에 집중하였다.

3. 형식 및 내용 검증

형식 검증 및 주식 내용 검증은 [그림 52]와 같이 진행된다. (1) 먼저 검증하고자 하는 두 말뭉치(검증용 말뭉치, 검증 대상 분석 말뭉치)를 저장한 후 (2) 검증 모듈 입력 변환기에 입력하면 검증 모듈 입력 형식으로 변환된다. (3) 이후 두 말뭉치의 검증 모듈 입력 파일이 저장된 두 폴더를 비교하여 1, 2차 형식 검증 및 주식 내용을 검증한다. (4) 검증 모듈 입력 파일이 담긴 두 폴더 내에서 문서(문어), 파일(구어) 일련번호가 일치한 파일들은 1, 2차 형식 검증과 주식 내용 검증을 모두 거치게 되며, 검증 대상 분석 말뭉치에만 있는 문서(문어), 파일(구어) 파일은 1차 형식 검증만 수행한다. (5) 1, 2차 형식 검증 및 주식 내용 결과는 결과 로그 파일로 출력된다. (6) 출력을 참고하여 검증용 말뭉치와 검증 대상 분석 말뭉치를 수정, 보완한다.



[그림 52] 형식 검증 및 주식 내용 검증 과정

형식과 주식 내용 검증 모듈의 입력은 표준 주식 형식에 기반을 둔다. 검증 모듈 입력이 표준 주식 형식과 호환되기 위해서는 말뭉치 내 문서(문어), 파일(구어) 단위로 나누는 과정이 필요하다. 그 이유는 검증용 말뭉치와 검증 대상 분석 말뭉치의 일련번호 체계가 문서(문어), 파일(구어) 단위부터 공유하게 되어 있기 때문이다. 따라서 검증용 말뭉치와 검증 대상 분석 말뭉치를 검증 모듈 입력 단위로 나누는 변환기⁶⁾를 만들어 두 말뭉치를 각각 문서(문어), 파일(구어) 단위로 바꾸어 검증 모듈 입력으로 사용한다.

6) koreanvalidationpublic/converter/spilt_as_document.py

3.1. 검증 실시 현황

7개 층위 분석 말뭉치 검증은 층위별로 총 3회에 나누어 검증이 예정되어 있었으나, 사업 기간 중 검증 일정 조정을 통하여 문, 구어 시범 검증 1회를 포함하여 최소 2회에서 최대 4회까지 층위별로 실시하였다. 층위별 실시 현황은 <표 20>과 같다.

층위	구분	시범 검증	1차 검증	2차 검증	3차 검증
형태 분석	문어	실시	실시	실시	-
	구어	실시	-	-	실시
어휘의미 분석	문어	실시	-	-	실시
	구어	실시	-	-	실시
개체명	문어	실시	-	-	실시
	구어	-	-	-	실시
주격 무형대용어 복원	문어	실시	실시	실시	-
	구어	실시	-	-	실시
상호참조 해결	문어	실시	-	-	실시
	구어	실시	-	-	실시
구문 분석	문어	실시	실시	실시	실시
의미역 분석	문어	실시	-	-	실시

<표 20> 층위별 분석 말뭉치 검증 현황

형태 분석 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1차 검증에서는 시범 검증 수정본을 포함하여 약 90만 어절을 검증하였다. 2차 검증에서는 시범 검증 및 1차 검증 수정본을 포함하여 약 110만 어절을 검증하였다. 따라서 문어 말뭉치는 총 3회에 걸쳐 약 200만 어절을 검증하였다. 구어 말뭉치의 경우 시범 검증에서 1만 어절을 검증하였고, 3차 검증에서 시범 검증 수정 분량 포함 총 100만 어절을 검증하였다.

어휘의미 분석 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1, 2차 검증은 생략하고, 3차 검증에서는 시범 검증 수정본을 포함하여 약 200만 어절을 모두 검증하였다. 따라서 문어 말뭉치는 총 2회에 걸쳐 약 200만 어절을 검증하였다. 구어 말뭉치의 경우 시범 검증에서 1만 어절을 검증하였고, 3차 검증에서 시범 검증 수정 분량 포함 총 100만 어절을 검증하였다.

개체명 분석 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1, 2차

검증은 생략하고, 3차 검증에서는 시범 검증 수정본을 포함하여 약 200만 어절을 모두 검증하였다. 따라서 문어 말뭉치는 총 2회에 걸쳐 약 200만 어절을 검증하였다. 구어 말뭉치의 경우 사업 일정상 시범 검증 분량에 해당하는 말뭉치를 국립국어원에서 우선 배포하여 시범 검증을 하지 않았다. 3차 검증에서 시범 검증 수정 분량 포함 총 100만 어절을 검증하였다.

주격 무형대용어 해결 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1차 검증에서는 시범 검증 수정본을 포함하여 약 94만 어절을 검증하였다. 2차 검증에서는 시범 검증 및 1차 검증 수정본을 포함하여 약 108만 어절을 검증하였다. 따라서 문어 말뭉치는 총 3회에 걸쳐 약 200만 어절을 검증하였다. 구어 말뭉치의 경우 시범 검증에서 1만 어절을 검증하였고, 3차 검증에서 시범 검증 수정 분량 포함 총 100만 어절을 검증하였다.

상호참조 해결 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1, 2차 검증은 생략하고, 3차 검증에서는 시범 검증 수정본을 포함하여 약 200만 어절을 모두 검증하였다. 따라서 문어 말뭉치는 총 2회에 걸쳐 약 200만 어절을 검증하였다. 구어 말뭉치의 경우 시범 검증에서 1만 어절을 검증하였고, 3차 검증에서 시범 검증 수정 분량 포함 총 100만 어절을 검증하였다.

구문 분석 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1차 검증에서는 시범 검증 수정본을 포함하여 약 58만 어절을 검증하였다. 2차 검증에서는 시범 검증 및 1차 검증 수정본을 포함하여 약 60만 어절을 검증하였다. 3차 검증에서는 시범 검증, 1차 검증 수정본을 포함하여 약 80만 어절을 검증하였다. 따라서 문어 말뭉치는 총 4회에 걸쳐 약 200만 어절을 검증하였다.

의미역 층위 문어 말뭉치의 경우 시범 검증에서 2만 어절을 검증하였다. 1, 2차 검증은 생략하고, 3차 검증에서는 시범 검증 수정본을 포함하여 약 200만 어절을 모두 검증하였다. 따라서 문어 말뭉치는 총 2회에 걸쳐 약 200만 어절을 검증하였다.

각 층위별 형식 검증 및 주석 내용 검증 대상 통계는 <표 21>와 같다.

층위	차수	검증 대상 분석 말뭉치 (단위: 어절)	검증용 말뭉치 (단위: 어절)	주석 내용 검증 비율 ⁷⁾ (단위: %)
형태 분석	문어 시범	20,405	20,405	100
	구어 시범	10742	10742	100
	1차 문어	902591	63202	4.34
	2차 문어	1097622	70380	6.41
	3차 구어	1011774	70761	6.99
어휘의미 분석	문어 시범	20,405	20,405	100
	구어 시범	10742	10742	100
	3차 문어	2000215	140633	7.03
	3차 구어	1011774	70761	6.99
개체명 분석	문어 시범	20,405	20,405	100
	3차 문어	2000215	140633	7.03
	3차 구어	1011774	70761	6.99
주격 무형 대용어 복원	문어 시범	20,405	20,405	100
	구어 시범	10742	10742	100
	1차 문어	939739	49992	5.32
	2차 문어	1079583	90641	8.40
	3차 구어	1011774	70761	6.99
상호참조 해결	문어 시범	20,405	20,405	100
	구어 시범	10742	10742	100
	3차 문어	2000215	140633	7.03
	3차 구어	1011774	70761	6.99
구문 분석	문어 시범	20,405	20,405	100
	1차 문어	581513	10133	1.74
	2차 문어	600077	24769	4.13
	3차 문어	798221	70650	8.85
의미역 분석	문어 시범	20,405	20,405	100
	3차 문어	1011774	70761	6.99

〈표 21〉 형식 검증 및 주석 내용 검증 대상 통계

7) 해당 검증 차수에서 주석 내용 검증 수행 비율이다. 이 증 대상 분석 말뭉치와 검증용 말뭉치가 일치하는 범위에 해당한다.

3.2. 형식 검증 항목 및 결과

1차 형식 검증은 모든 층위의 검증 대상 분석 말뭉치를 대상으로 주석 표준 형식 준수 여부를 검사하는 과정이다. 따라서 주석 표준 형식에 맞지 않는 말뭉치 형식이 검증 모듈의 입력될 경우 검사를 수행할 수 없다. 1차 형식 검증은 2차 형식 검증, 주석 내용 검증 모듈⁸⁾에 포함되어 있으며, 가장 먼저 수행된다. 세부 내용은 <표 22>와 같으며, 1차 형식 검증 모듈의 자료 구조는 <표 23>과 같다.

1차 형식 검증에서는 주석 표준 형식의 1수준 ‘id’와 ‘document’ 구성 유효성 여부와 해당 key의 2수준 구성 유효성 여부를 검사한다. 1차 형식 검증 오류 코드는 <표 24>와 같다.

모듈 함수	검증 수행 내용
SingletonFormatErrorCheckerFn	<ul style="list-style-type: none"> - 주석 표준 형식의 key 값 체크 - 말뭉치, 문서, 문단, 문장 일련번호 등 필수 일련번호 존재 여부 파악
FormatErrorCheckerFn	<ul style="list-style-type: none"> - 동일 검증 모듈 입력 내에서 같은 key에 대한 value 일치 검사 - 본 단계에서 검증용 말뭉치의 SingletonFormatErrorChecker도 함께 수행 - 원문 일치 여부 검사

<표 22> 1 차 형식 검증 세부 내용

8) koreanvalidationpublic/ValidationScript.py

Class	Variables	주요 Method
ValidationResult	<ul style="list-style-type: none"> - errors: List[CorpusError]: 불일치 내용 리스트 - scores: List[Score]: 검증 점수 리스트 	<ul style="list-style-type: none"> - result_str: errors 요소의 개수가 0이면 ‘통과’ 출력, 0이 아니면 불일치 내용 출력 - scores 요소 개수가 0이면 ‘통과’ 출력, 0이 아니면 검증 점수 출력 - score_str: 개별 검증 모듈 입력의 검증 점수 반환
CorpusError	<ul style="list-style-type: none"> - corpus_id: String (말뭉치 일련번호) - error_tag: String (1차 형식 오류 유형) - gold: String (검증용 말뭉치의 주석 내용) - pred: String (검증 대상 분석 말뭉치의 주석 내용) 	<ul style="list-style-type: none"> - to_str: 오류 내용을 종합하여 String으로 출력
Score(ABC)	<ul style="list-style-type: none"> - key: String (검증 점수 종류) 	

〈표 23〉 1차 형식 오류 모듈 자료 구조

형식 오류 코드	오류 내용
FORMAT_ERROR_NO_CORPUS_ID	입력 말뭉치의 'id' key가 없는 경우
FORMAT_ERROR_ID_MISMATCH	검증 모듈 입력의 'id' 가 불일치한 경우
FORMAT_ERROR_NO_DOCUMENT	입력 말뭉치의 'document' key가 없는 경우
FORMAT_ERROR_NO_DOCUMENT_ID	'document-id' key가 없는 경우
FORMAT_ERROR_DOCUMENT_ID_MISMATCH	입력 말뭉치들의 'document-id' 가 불일치한 경우
FORMAT_ERROR_DOCUMENT_LENGTH_MISMATCH	입력 말뭉치들의 'document' 의 객체 개수가 불일치한 경우
FORMAT_ERROR_NO_SENTENCE	'document-sentence' key가 없는 경우
FORMAT_ERROR_NO_SENTENCE_ID	'document-sentence-id' key가 없는 경우
FORMAT_ERROR_SENTENCE_FORM_MISMATCH	'document-sentence-form' 이 원시 말뭉치와 불일치한 경우
FORMAT_ERROR_SENTENCE_LENGTH_MISMATCH	'document-sentence' 의 객체 개수가 불일치한 경우

〈표 24〉 1차 형식 오류 코드

2차 형식 검증은 1차 형식 검증 이후 진행되는 형식 검증이다. 본 단계에서 검증하는 내용은 총위 공통 주식 표준 형식뿐만 아니라 총위별 주식 표준 형식 중 분석 총위별로 정의된 key, 자료형, value 유효성 검사를 수행한다. 2차 형식 검증 모듈의 자료 구조는 〈표 25〉와 같다.

층위	Class	설명
형태 분석	MorphKey CheckerFn	- 'sentence-morpheme' key 구성 여부 확인
	MorphLabel CheckerFn	- 'morpheme' 내 'label' value 유효 범위 검사 - 유효 범위 (태그 세트 49개): NNG, NNP, NNB, NP, NR, VV, VA, VX, VCP, VCN, MMA, MMD, MMN, MAG, MAJ, IC, JKS, JKC, JKG, JKO, JKB, JKV, JKQ, JX, JC, EP, EF, EC, ETN, ETM, XPN, XSN, XSV, XSA, XR, SF, SP, SS, SE, SO, SW, SL, SH, SN, NA, NF, NV, NAP
	MorphWord IDCheckerFn	- word_id 일치 확인
어휘의 미 분석	WsdForm CheckerFn	- 'sentence-WSD' key 구성 여부 확인 - 'sense_id' 의 자료형이 정수형(int)인지 확인 - 'sense_id' 가 정수형일 때, 어휘의미 번호 유효 범위 검사 - 어휘의미 번호 유효 범위: 1~999
개체명 분석	NEForm CheckerFn	- 'sentence-NE' key 구성 여부 확인 - 'NE' 내 'id', 'form', 'label', 'begin', 'end' 구성 여부 확인
	NELabel CheckerFn	- 'NE' 내 'label' value 유효 범위 검사 - 유효 범위 (태그 세트 15개): PS, FD, TR, AF, OG, LC, CV, DT, TI, QT, EV, AM, PT, MT, TM
주격 무형대 용어 복원	ZAKey CheckerFn	- 'document-ZA' key 구성 여부 확인 - 'ZA' 내 'predicate', 'antecedent' key 구성 여부 확인 - 'predicate' 내 'form', 'sentence_id', 'begin', 'end' 구성 여부 확인 - 'antecedent' 내 'type', 'form', 'sentence_id', 'begin', 'end' 구성 여부 확인
	ZAFormat CheckerFn	- 'predicate' 와 'antecedent' 의 'begin', 'end' value 유효 범위 검사 - 'begin' value 유효 범위 검사 - 'end' value 유효 범위 검사
상호 참조 해결	CRKey CheckerFn	- 'document-CR' key 구성 여부 확인 - 'CR' 내 'mention' key 구성 여부 확인 - 'mention' 내 'form', 'sentence_id', 'word_ids' 구성 여부 확인
	Duplicate Mention CheckerFn	- 서로 다른 'mention' 에서 동일한 'form', 'sentence_id', 'word_ids' 조합 발생 여부 확인
구문	DepKey	- 'sentence-DP' key 구성 여부 확인

분석	CheckerFn	<ul style="list-style-type: none"> - ‘DP’ 내 ‘word_id’, ‘word_form’, ‘head’, ‘label’ 구성 여부 확인
	DepFormatCheckerFn	<ul style="list-style-type: none"> - ‘DP’ 내 ‘head’ 와 ‘label’ value 유효 범위 검사 - ‘head’ value 유효 범위: -1 (최상위 지배소), $1 \leq head \leq \max(wordId)$ (지배소) - ‘label’ value 유효 범위: 구문 태그, 기능 태그 유효 범위를 모두 충족 - 구문 태그 유효 범위 (태그 9개): NP, VP, AP, VNP, DP, IP, X, L, R - 기능 태그 유효 범위 (태그 6개): none, SBJ, OBJ, MOD, AJT, CMP, CNJ
의미역 분석	SRLKeyFormatCheckerFn	<ul style="list-style-type: none"> - ‘sentence-SRL’ key 구성 여부 확인 - ‘SRL’ 내 ‘predicate’, ‘argument’ key 구성 여부 확인 - ‘predicate’ 내 ‘form’, ‘begin’, ‘end’, ‘lemma’, ‘sense_id’ key 구성 여부 확인 - ‘argument’ 내 ‘form’, ‘begin’, ‘end’, ‘label’ key 구성 여부 확인 - ‘label’ value 유효 범위: 필수역, 부가역 유효 범위를 모두 충족 - 필수역 유효 범위 (태그 세트 4개): ‘ARG0’, ‘ARG1’, ‘ARG2’, ‘ARG3’ - 부가역 유효 범위 (태그 세트 15개): ‘ARGM-LOC’, ‘ARGM-DIR’, ‘ARGM-CND’, ‘ARGM-MNR’, ‘ARGM-TMP’, ‘ARGM-EXT’, ‘ARGM-PRD’, ‘ARGM-PRP’, ‘ARGM-CAU’, ‘ARGM-DIS’, ‘ARGM-ADV’, ‘ARGM-NEG’, ‘ARGM-INS’, ‘AUX’, ‘ARGA’

〈표 25〉 2차 형식 오류 검증 모듈 자료 구조

2차 형식 검증에서는 주석 표준 형식의 필수 말뭉치 유형, 필수 분석 층위의 유효성을 검사한다. 따라서 층위 공통 형식 검증과 층위별 형식 검증을 동시에 수행한다. 2차 형식 오류 코드는 〈표 26〉과 같다.

층위	형식 오류 코드	설명
공통	FORMAT_ERROR_SENTENCE_MISMATCH	- 'sentence' 내 'id' value가 유효하지 않은 경우
	FORMAT_ERROR_SENTENCE_FORM_MISMATCH	- 'sentence' 내 'form' 이 원시 말뭉치와 일치하지 않은 경우
	FORMAT_ERROR_WORD_INDEX_ERROR	- 'sentence' 내 'word' 가 원시 말뭉치와 일치하지 않은 경우
형태 분석	FORMAT_ERROR_MORPHEME_NO_MORPH_KEY	- 'sentence' 내 'morpheme' key가 없는 경우
	FORMAT_ERROR_MORPHEME_WORD_ID_MISMATCH	- 'word' 내 'id' 가 원시 말뭉치의 어절 수와 일치하지 않은 경우
	FORMAT_ERROR_MORPHEME_ILLEGAL_MORPH_LABEL	- 'morpheme' 내 'label' value가 유효하지 않은 경우
어휘의 미 분석	FORMAT_ERROR_WSD_NO_WSD_KEY	- 'sentence' 내 'WSD' key가 없는 경우
	FORMAT_ERROR_WSD_SENSE_ID_TYPE	- 'WSD' 내 'sense_id' 의 자료형이 유효하지 않은 경우
	FORMAT_ERROR_WSD_SENSE_ID_OUT_OF_RANGE	- 'WSD' 내 'sense_id' 가 정수형일 때 유효 범위를 벗어난 경우
개체명 분석	NER_NO_NE	- 'sentence' 내 'NE' key가 없는 경우
	NER_NO_NE_ID	- 'NE' 내 'form' key가 없는 경우
	NER_FORM_INDEX_MISMATCH	- 'NE' 내 'begin' 과 'end' value를 활용하여 원시 말뭉치에서 얻은 텍스트와 'NE' 내 'form' 이 불일치한 경우 - 실제 형식 오류는 아니며, 국립국어원의 요청으로 추가한 경고 코드
	NER_NO_NE_LABEL	- 'NE' 내 'label' key가 없는 경우
	NER_NE_ILLEGAL_LABEL	- 'NE' 내 'label' value가 유효하지 않은 경우
	NER_NO_BEGIN_INDEX	- 'NE' 내 'begin' key가 없는 경우
	NER_NO_END_INDEX	- 'NE' 내 'end' key가 없는 경우
주격 무형대 용어 복원	FORMAT_ERROR_ZEROANAPHORA_NO_ZA_KEY	- 'document' 내 'ZA' key가 없는 경우
	FORMAT_ERROR_ZEROANAPHORA_NO_PREDICATE	- 'ZA' 내 'predicate' key가 없는 경우
	FORMAT_ERROR_ZEROANAPHORA_NO_ANTECEDENT	- 'ZA' 내 'antecedent' key가 없는 경우
	FORMAT_ERROR_ZEROANAPHORA_PREDICATE_FORM	- 'predicate' , 'antecedent' 내 'form'

	APHORA_NO_FORM	key가 없는 경우
	FORMAT_ERROR_ZEROAN APHORA_NO_BEGIN	- ‘predicate’, ‘antecedent’ 내 ‘begin’ key가 없는 경우
	FORMAT_ERROR_ZEROAN APHORA_WRONG_BEGIN	- ‘begin’ 값이 유효 범위를 벗어난 경우. 단, 영주어의 ‘begin’ 은 -1
	FORMAT_ERROR_ZEROAN APHORA_WRONG_ORDER	- ‘begin’ 값이 ‘end’ 값보다 큰 경우
	FORMAT_ERROR_ZEROAN APHORA_NO_END	- ‘predicate’, ‘antecedent’ 내 ‘end’ key가 없는 경우
	FORMAT_ERROR_ZEROAN APHORA_NO_TYPE	- ‘antecedent’ 내 ‘type’ key가 없는 경 우
	FORMAT_ERROR_ZEROAN APHORA_NULL_ ANTECEDENT	- 주석 대상 ‘predicate’ 의 ‘antecedent’ 가 없는 경우
상호 참조 해결	FORMAT_ERROR_ZEROAN APHORA_ ANTECEDENT_DIFFERENT	- ‘antecedent’ 에 해당하는 ‘predicate’ 가 없는 경우
	FORMAT_ERROR_CR_NO_ CR_KEY	- ‘document’ 내 ‘CR’ key가 없는 경우
	FORMAT_ERROR_CR_NO_ MENTION_KEY	- ‘CR’ 내 ‘mention’ key가 없는 경우
	CR_DUPLICATE_MENTION	- ‘mention’ 이 중복될 경우
	FORMAT_ERROR_CR_NO_ MENTION_FORM_KEY	- ‘mention’ 내 ‘form’ key가 없는 경우
구문 분석	FORMAT_ERROR_CR_NO_ MENTION_SID_KEY	- ‘mention’ 내 ‘sentence_id’ key가 없는 경우
	FORMAT_ERROR_DEPEND ENCY_NO_DP_KEY	- ‘sentence’ 내 ‘DP’ key가 없는 경우
	FORMAT_ERROR_DEPEND ENCY_NO_WORD_ID	- ‘DP’ 내 ‘word_id’ key가 없는 경우
	FORMAT_ERROR_DEPEND ENCY_NO_WORD_FORM	- ‘DP’ 내 ‘word_form’ key가 없는 경우
	FORMAT_ERROR_DEPEND ENCY_NO_HEAD	- ‘DP’ 내 ‘head’ key가 없는 경우
	FORMAT_ERROR_DEPEND ENCY_NO_LABEL	- ‘DP’ 내 ‘label’ key가 없는 경우
	FORMAT_ERROR_DEPEND ENCY_ILLEGAL_SYNTAX_	- ‘label’ 의 구문 태그의 유효 범위를 벗어 난 경우

	LABEL	
	FORMAT_ERROR_DEPENDENCY_ILLEGAL_FUNCTION_LABEL	- ‘label’ 의 기능 태그의 유효 범위를 벗어난 경우
의미역 분석	FORMAT_ERROR_SRL_NO_SRL_KEY	- ‘sentence’ 내 ‘SRL’ key가 없는 경우
	FORMAT_ERROR_SRL_PREDICATE_NO_KEY_{{유효 범위}}	- ‘predicate’ 내 하위 key 유효 범위를 벗어날 경우
	FORMAT_ERROR_SRL_ARGUMENT_NO_KEY_{{유효 범위}}	- ‘argument’ 내 하위 key 유효 범위를 벗어날 경우
	FORMAT_ERROR_SRL_ARGUMENT_LABEL_{{유효 범위}}	- ‘label’ 내 하위 key 유효 범위를 벗어날 경우
	FORMAT_ERROR_SRL_SENSE_ID_LENGTH_6	- ‘predicate’ 의 sense_id의 유효 범위를 벗어난 경우

〈표 26〉 2차 형식 오류 코드

형태 분석 층위 형식 검증 결과는 〈표 27〉과 같다. 형식 오류가 발생한 분석 단위는 주석 내용 오류를 검증하지 않고, 구축사업단에 수정, 보완을 요청하였다. 주격 무형 대용어 복원 층위와 구문 분석 층위에서는 형식 오류가 발견되지 않았다.

층위	차수	오류 유형	빈도 (단위: 개)	비율 ⁹⁾ (단위: %)
형태 분석	문어 시범	FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	3	0.45
	구어 시범	FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	46	5.83
		FORMAT_ERROR_MORPHEME_WORD_ID_ MISMATCH	24	3.04
	1차 문어	FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	3	0.50
	2차 문어	FORMAT_ERROR_MORPHEME_WORD_ID_ MISMATCH	1	0.02
	3차 구어	FORMAT_ERROR_MORPHEME_WORD_ID_ MISMATCH	6	0.13
		MORPHEME_WORD_LENGTH_MISMATCH	55	1.22
어휘의 미 분석	문어 시범	FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	1	0.07
	구어 시범	-	-	-
	3차 문어	FORMAT_ERROR_WSD_SENSE_ID_ OUT_OF_RANGE	1	0.00
	3차 구어	-	-	-
개체명 분석	문어 시범	FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	1	0.07
	3차 문어	-	-	-
	3차 구어	FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	10	1.56
상호참 조 해결	문어 시범	-	-	-
	구어 시범	-	-	-
	3차 문어	CR_DUPLICATE_MENTION	120	6.89
	3차 구어	FORMAT_ERROR_SENTENCE_LENGTH_ MISMATCH	1	0.23
		FORMAT_ERROR_SENTENCE_FORM_ MISMATCH	1	0.23
의미역	시범	-	-	-

분석	3차	FORMAT_ERROR_SRL_SENSE_ID_LENGTH_6	3	0.00
		FORMAT_ERROR_SRL_ARGUMENT_LABEL_ ARGM	2	0.00
		FORMAT_ERROR_SRL_ARGUMENT_LABEL_ ARGM-LOC(장소)	1	0.00
		FORMAT_ERROR_SRL_ARGUMENT_LABEL_ >>>	1	0.00
		FORMAT_ERROR_SRL_ARGUMENT_LABEL_ ARR2	1	0.00
		FORMAT_ERROR_SRL_ARGUMENT_LABEL_ ARG0-TMP	1	0.00

〈표 27〉 층위별 분석 말뭉치 형식 검증 오류 유형 및 통계

9) 해당 검증 차수의 형식 오류 및 내용 오류의 합에서 형식 오류 비율. 소수 셋째 자리에서 반올림함.

3.3. 내용 검증 지표 및 결과

검증 대상 분석 말뭉치가 형식 검증을 통과하면 검증 대상 분석 말뭉치 표본(7%)을 검증용 말뭉치와 주석 내용을 비교하여 검증한다. 층위별 내용 오류 검증 항목 및 검증 대상은 층위별 구축지침을 준용하여 결정하였으며, 검증 지표는 해당 검증 항목을 평가하기 위한 지표를 선정하여 국립국어원의 승인을 얻었다.

내용 검증 결과인 검증용 말뭉치와 검증 대상 분석 말뭉치의 불일치 부분이 기록된 결과 로그 파일은 먼저 본 사업단 내 검토를 한 후 국립국어원으로 보고하였다. 이후 국립국어원의 불일치 주석 단위 정오 판별 결과를 본 사업단과 층위별 분석 말뭉치 구축사업단에 통보하여 수정, 보완을 요청하였다. 이후 수정, 보완된 말뭉치를 다시 형식 검증부터 수행하여 불일치 부분이 기록된 새로운 결과 로그 파일을 검토하고 국립국어원에 보고하는 형식의 순환 과정을 거쳤다.

3.3.1. 형태 분석

형태 분석 주석 내용 오류 검증 항목은 <표 28>과 같다. 형태 분석 주석 내용 오류 검증 항목은 어절 내 형태소 분절 일치(A) 여부 검사와 형태소 분절이 일치한 어절에 대한 형태소 태그 부여 일치(B)를 검사한다. 검증 지표는 어절 내 형태소 분절 일치도의 정확률과 형태소 태그 부여 일치의 F1 점수를 곱하여 최종 주석 내용 일치도를 산출한다.

검증 항목	검증 대상	검증 지표
어절 내 형태소 분절 일치 (A)	모든 어절	정확률(A) * F1 점수(B)
형태소 태그 부여 일치 (B)	형태소 분절이 일치한 어절	

<표 28> 형태 분석 내용 오류 검증 항목

형태 분석 층위 주석 내용 오류 검증 모듈의 자료 구조는 <표 29>와 같다. 또한, 주석 내용 오류 코드는 <표 30>과 같다.

Class	입력	출력	설명
MorphSplit CheckerFn	주석 표준 형식 형태 분석 말뭉치	결과 로그 파일 (정밀도, 재현율, F1 점수)	<ul style="list-style-type: none"> - 두 말뭉치의 'word_id' 의 일치 확인 - 동일 'word_id' 내의 형태소 개수 비교 후 과소, 과대 분절 판별 - 같은 개수로 분절된 어절에 대하여 - 'form' 과 'label' 일치 확인

〈표 29〉 형태 분석 내용 오류 검출 모듈 자료 구조

형식 오류 코드	설명
MORPHEME_UNDERSPLIT	- 어절 내 형태소 과소 분절 오류
MORPHEME_OVERSPLIT	- 어절 내 형태소 과대 분절 오류
MORPHEME_SPLIT_ERROR	- 어절 내 형태소 분절 불일치 오류
MORPHEME_LABEL_ERROR	- 분절 일치 어절 내 형태소 태그 불일치

〈표 30〉 형태 분석 내용 검증 오류 코드

형태 분석 검증 대상 분석 말뭉치는 문어 3회, 구어 2회 등 총 5회에 걸쳐 검증을 수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 ¹⁰⁾ (단위: %)
형태 분석	문어 시범	어절	20,405	20,405	100
	구어 시범	어절	10742	10742	100
	1차 문어	어절	902591	63202	4.34
	2차 문어	어절	1097622	70380	6.41
	3차 구어	어절	1011774	70761	6.99

〈표 31〉 형태 분석 말뭉치 형식 및 주석 내용 검증 대상 통계

형태 분석 층위 주석 내용 검증 오류 유형 및 통계는 〈표 32〉와 같다. 형태 분석 층위에서는 어절 내 분절이 일치하는 어절에 대해서 형태소 태그 부착 검사까지 수행하였다. 어

10) 해당 검증 차수에서 주석 내용 검증 수행 비율이다. 이 중 대상 분석 말뭉치와 검증용 말뭉치가 일치하는 범위에 해당한다.

절 내에서 하나 이상의 형태소에서 MORPHEME_OVERSPLIT, MORPHEME_UNDERSPLIT, MORPHEME_SPLIT_ERROR가 발생한 경우, 형태소 태그 부착 검사를 수행하지 않았다. 따라서 MORPHEME_LABEL_ERROR는 어절 내 형태소 분절이 모두 일치한 어절 중에서 형태소 태그가 다른 경우이다.

층위	차수	오류 유형	빈도	비율 ¹¹⁾
형태 분석	문어 시범	MORPHEME_OVERSPLIT	183	27.23
		MORPHEME_UNDERSPLIT	47	6.99
		MORPHEME_SPLIT_ERROR	34	5.06
		MORPHEME_LABEL_ERROR	404	60.12
	구어 시범	MORPHEME_OVERSPLIT	71	9.00
		MORPHEME_UNDERSPLIT	233	29.53
		MORPHEME_SPLIT_ERROR	41	5.20
		MORPHEME_LABEL_ERROR	374	47.40
	1차 문어	MORPHEME_OVERSPLIT	110	18.49
		MORPHEME_UNDERSPLIT	70	11.76
		MORPHEME_SPLIT_ERROR	57	9.58
		MORPHEME_LABEL_ERROR	343	57.65
	2차 문어	MORPHEME_OVERSPLIT	304	6.96
		MORPHEME_UNDERSPLIT	1453	33.26
		MORPHEME_SPLIT_ERROR	1058	24.22
		MORPHEME_LABEL_ERROR	1552	35.53
	3차 구어	MORPHEME_OVERSPLIT	460	10.20
		MORPHEME_UNDERSPLIT	1342	29.75
		MORPHEME_SPLIT_ERROR	532	11.79
		MORPHEME_LABEL_ERROR	2110	46.77

〈표 32〉 형태 분석 말뭉치 내용 검증 오류 유형 및 통계

형태 분석 층위 주석 내용 검증 결과 일치도 점수는 〈표 33〉과 같다.

층위	차수	분절 일치도	태그 일치도	전체 일치도	비고
형태 분석	문어 시범	98.73	99.11	97.85	
	구어 시범	96.74	98.17	94.97	
	1차 문어	99.41	99.61	99.02	
	2차 문어	96.10	99.01	95.14	
	3차 구어	96.24	98.08	94.39	

〈표 33〉 형태 분석 말뭉치 주석 내용 검증 점수

11) 해당 검증 차수의 형식 오류 및 내용 오류의 합에서 각 오류 유형의 비율. 소수 셋째 자리에서 반올림함.

3.3.2. 어휘의미 분석

어휘의미 분석 주석 내용 오류 검증 항목은 <표 34>와 같다. 어휘의미 분석 주석 내용 오류 검증 항목은 문서 내 어휘의미 주석 대상 탐지 일치 여부와 탐지한 어휘의 어휘의미 번호 일치 여부를 검사한다. 어휘의미 분석 주석 내용 검증은 어휘의미 번호 주석 대상이 되는 모든 대상에 대해 주석을 제대로 했는지 재현율(Recall)로, 어휘의미 번호 주석 일치 여부를 정밀도(Precision)로 산출하고, 최종 주석 일치도는 이 두 지표의 조화 평균인 F1 점수로 산출한다.

검증 항목	검증 대상	검증 지표
어휘의미 주석 대상 탐지 일치	모든 어절	F1 점수
어휘의미 부여 일치	어휘의미 주석 대상 탐지 일치 어휘	

<표 34> 어휘의미 분석 주석 내용 오류 검증 항목

Class	입력	출력	설명
WsdScoreCheckerFn	주석 표준 형식 어휘의미 분석 말뭉치	결과 로그 파일 (정밀도, 재현율, F1 점수)	<ul style="list-style-type: none"> - 주석 대상 단위를 파악하기 위해서 'word' 와 'word_id' 가 모두 일치 확인 - 같은 주석 대상을 주석한 경우 'sense_id' 일치 여부 확인

<표 35> 어휘의미 분석 내용 오류 검출 모듈 자료 구조

형식 오류 코드	설명
WSD_FP_ERROR	- 어휘의미 주석 대상이 아닌 어휘에 주석
WSD_FN_ERROR	- 어휘의미 주석 대상인 어휘를 주석하지 않음
WSD_NUMBERRING_ERROR	- 어휘의미 주석 대상 일치 어휘의 어휘의미 번호 불일치

<표 36> 층위별 주석 내용 검증 오류 코드

어휘의미 분석 검증 대상 분석 말뭉치는 문어 2회, 구어 2회 등 총 4회에 걸쳐 검증을

수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 (단위: %)
어휘의미 분석	문어 시범	어절	20,405	20,405	100
	구어 시범	어절	10742	10742	100
	3차 문어	어절	2000215	140633	7.03
	3차 구어	어절	1011774	70761	6.99

〈표 37〉 어휘의미 분석 말뭉치 형식 및 주석 내용 검증 대상 통계

어휘의미 분석 층위 주석 내용 검증 오류 유형 및 통계는 〈표 38〉과 같다. 어휘의미 분석 층위에서는 어절 내 분절이 일치하는 어절에 대해서 어휘의미 번호 부착 검사까지 수행하였다. 주석 단위 내에서 WSD_FP_ERROR, WSD_FN_ERROR가 발생한 경우 어휘의미 번호 일치 검사를 수행하지 않았다. 따라서 WSD_NUMBERING_ERROR는 주석 단위 내에서 어휘의미 부착 대상 어휘 주석 및 범위가 일치한 것 중에 어휘의미 번호를 다르게 주석한 경우이다.

층위	차수	오류 유형	빈도	비율
어휘의 미 분석	문어 시범	WSD_FP_ERROR	224	16.62
		WSD_FN_ERROR	209	15.50
		WSD_NUMBERING_ERROR	914	67.80
	구어 시범	WSD_FP_ERROR	162	22.28
		WSD_FN_ERROR	136	18.71
		WSD_NUMBERING_ERROR	429	59.01
	3차 문어	WSD_FP_ERROR	8149	46.56
		WSD_FN_ERROR	5906	33.75
		WSD_NUMBERING_ERROR	3445	19.69
	3차 구어	WSD_FP_ERROR	1269	27.41
		WSD_FN_ERROR	925	19.98
		WSD_NUMBERING_ERROR	2453	52.60

〈표 38〉 어휘의미 분석 말뭉치 주석 내용 검증 결과

어휘의미 분석 층위 주석 내용 검증 점수는 〈표 39〉와 같다.

층위	차수	정밀도	재현율	F1 점수	비고
어휘의미 분석	문어 시범	93.11	98.73	95.84	
	구어 시범	88.73	96.10	92.27	
	3차 문어	90.46	94.89	92.47	
	3차 구어	88.34	97.06	92.49	

〈표 39〉 어휘의미 분석 말뭉치 주석 내용 검증 점수

3.3.3. 개체명 분석

개체명 분석 주석 내용 오류 검증 항목은 〈표 40〉과 같다. 개체명 분석 주석 내용 오류 검증 항목은 문서 내 개체 범위 일치 여부 검사와 개체로 주석된 어휘에 부여된 개체명 태그 일치 여부를 검사한다. 개체명 분석의 개체 범위 일치는 정밀도(Precision)로, 개체 태그 부여 일치는 재현율(Recall)을 산출하고, 최종 주석 일치도는 이 두 지표의 조화 평균인 F1 점수로 산출¹²⁾한다.

검증 항목	검증 대상	검증 지표
개체 범위 일치	모든 어절	F1 점수
개체 태그 부여 일치	개체 범위 일치 어휘	

〈표 40〉 개체명 분석 주석 내용 오류 검증 항목

Class	입력	출력	설명
WarningFn	주석 표준 형식 개체명 분석 말뭉치	결과 로그 파일 (알림 출력)	<ul style="list-style-type: none"> - 국립국어원의 요청으로 추가 - ‘NE’ 내 ‘begin’, ‘end’ 를 통해 문장에서 추출한 ‘form’ 과 ‘NE’ 의 ‘form’ 이 같은지 확인 - 다를 경우 ‘NER_FORM_INDEX_MISMATCH’ 알림 출력
ScoringFn		결과 로그 파일 (F1 점수)	<ul style="list-style-type: none"> - 개체 범위 및 라벨 부여 일치 확인

〈표 41〉 주석 내용 오류 검출 모듈 자료 구조

12) 개체명 분석 최종 주석 일치도를 산출하기 위해 사용한 라이브러리는 seqeval를 사용하였다.
(<http://github.com/chakki-works/seqeval>)

형식 오류 코드	설명
NER_FP_ERROR	- 개체가 아닌 어휘에 주석
NER_FN_ERROR	- 개체인 어휘를 주석하지 않음
NER_BIO_ERROR	- 개체 범위 불일치
NER_TYPE_ERROR	- 개체 태그 불일치

〈표 42〉 층위별 주석 내용 검증 오류 코드

개체명 분석 검증 대상 분석 말뭉치는 문어 2회, 구어 1회 등 총 3회에 걸쳐 검증을 수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 (단위: %)
개체명 분석	문어 시범	어절	20,405	20,405	100
	3차 문어	어절	2000215	140633	7.03
	3차 구어	어절	1011774	70761	6.99

〈표 43〉 개체명 분석 말뭉치 형식 및 주석 내용 검증 대상 통계

개체명 분석 층위 주석 내용 검증 오류 유형 및 통계는 〈표 44〉와 같다. 개체명 분석 층위에서는 개체 범위가 일치하는 개체에 대하여 개체 태그 부여 일치 검사까지 수행하였다. 주석 단위 내에서 NE_FP_ERROR, NE_FN_ERROR, NER_BIO_ERROR가 발생한 경우 개체 태그 부여 일치 검사를 수행하지 않았다. 따라서 NER_TYPE_ERROR는 주석 단위 내에서 개체 범위가 일치한 것 중에 개체 태그를 다르게 주석한 경우이다.

층위	차수	오류 유형	빈도	비율
개체명 분석	문어 시범	NER_FP_ERROR	440	41.31
		NER_FN_ERROR	141	13.24
		NER_BIO_ERROR	333	31.27
		NER_TYPE_ERROR	136	12.77
	3차 문어	NER_FP_ERROR	2351	30.01
		NER_FN_ERROR	2350	29.98
		NER_BIO_ERROR	1920	24.48
		NER_TYPE_ERROR	1214	15.48
	3차 구어	NER_FP_ERROR	214	33.33
		NER_FN_ERROR	227	35.36
		NER_BIO_ERROR	100	15.58
		NER_TYPE_ERROR	68	10.59

〈표 44〉 개체명 분석 말뭉치 주석 내용 검증 오류 유형 및 통계

개체명 분석 층위 주석 내용 검증 점수는 <표 45>와 같다.

층위	차수	F1 점수	비고
개체명 분석	문어 시범	87.02	
	3차 문어	86.02	
	3차 구어	94.48	

<표 45> 개체명 분석 말뭉치 주석 내용 검증 점수

3.3.4. 주격 무형대용어 복원

주격 무형대용어 복원 주석 내용 오류 검증 항목은 <표 46>과 같다. 주격 무형대용어 복원 주석 내용 오류 검증 항목은 문서 내 복원 대상 서술어 주석 일치 여부(A)와 복원 대상 서술어의 생략어 복원 일치 여부(B)를 검사한다. 검증 지표는 복원 대상 서술어 주석 일치 여부의 F1 점수와 생략어 복원 일치 여부의 정확률을 곱하여 최종 주석 내용 일치도를 산출한다.

검증 항목	검증 대상	검증 지표
복원 대상 서술어 주석 일치(A)	문서 내 모든 문장	F1 점수(A) * 정확률(B)
생략어복원 일치 (B)	복원 대상 서술어 주석이 일치한 서술어에 대한 복원된 생략어	

<표 46> 주격 무형대용어 복원 주석 내용 오류 검증 항목

Class	입력	출력	설명
ZAContent CheckerFn	주석 표준 형식 주격 무형대용어 복원 말뭉치	결과 로그 파일 (predicate F1 점수, antecedent 정확률)	- ‘predicate’ 와 ‘antecedent’ 의 일 치 기준은 두 말뭉치의 주석 결과 ‘sentence_id’ 가 일치하고, ‘begin’ , ‘end’ 가 겹치면 일치한 것으로 판단

<표 47> 주석 내용 오류 검출 모듈 자료 구조

형식 오류 코드	설명
ZEROANAPHORA_PREDICATE_MISSED	- 복원 대상 서술어를 주석하지 않은 경우
ZEROANAPHORA_PREDICATE_OVER	- 복원 대상 서술어가 아닌 서술어를 복원 대상으로 지정한 경우
ZEROANAPHORA_ANTECEDENT_DIFFERENT	- 복원 대상 서술어의 복원된 생략어가 불일치

〈표 48〉 층위별 주석 내용 검증 오류 코드

주격 무형대용어 복원 검증 대상 분석 말뭉치는 문어 3회, 구어 2회 등 총 5회에 걸쳐 검증을 수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 (단위: %)
주격 무형대용어 복원	문어 시범	어절	20,405	20,405	100
	구어 시범	어절	10742	10742	100
	1차 문어	어절	939739	49992	5.32
	2차 문어	어절	1079583	90641	8.40
	3차 구어	어절	1011774	70761	6.99

〈표 49〉 주격 무형대용어 복원 말뭉치 형식 및 주석 내용 검증 대상 통계

주격 무형대용어 복원 층위 주석 내용 검증 오류 유형 및 통계는 〈표 70〉과 같다. 주격 무형대용어 복원 층위에서는 복원 대상 서술어에 대해서 생략어 복원 검사까지 수행하였다. 주석 단위 내에서 ZEROANAPHORA_PREDICATE_OVER, ZEROANAPHORA_PREDICATE_OVER가 발생한 경우 생략어 복원 일치 검사를 수행하지 않았다. 따라서 ZEROANAPHORA_ANTECEDENT_DIFFERENT는 주석 단위 내에서 복원 대상 서술어가 일치한 것 중에 생략어 복원을 다르게 주석한 경우이다.

층위	차수	오류 유형	빈도	비율 ¹³⁾
주격 무형대 용어 복원	문어 시범	ZEROANAPHORA_PREDICATE_OVER	259	30.54
		ZEROANAPHORA_PREDICATE_OVER	201	23.70
		ZEROANAPHORA_ANTECEDENT_ DIFFERENT	388	45.75
	구어 시범	ZEROANAPHORA_PREDICATE_OVER	222	24.00
		ZEROANAPHORA_PREDICATE_MISSED	121	13.08
		ZEROANAPHORA_ANTECEDENT_ DIFFERENT	582	62.92
	1차 문어	ZEROANAPHORA_PREDICATE_OVER	841	37.36
		ZEROANAPHORA_PREDICATE_MISSED	491	21.81
		ZEROANAPHORA_ANTECEDENT_ DIFFERENT	919	40.83
	2차 문어	ZEROANAPHORA_PREDICATE_OVER	1244	26.91
		ZEROANAPHORA_PREDICATE_MISSED	1168	25.27
		ZEROANAPHORA_ANTECEDENT_ DIFFERENT	2210	47.81
	3차 구어	ZEROANAPHORA_PREDICATE_OVER	1645	28.95
		ZEROANAPHORA_PREDICATE_MISSED	861	15.15
		ZEROANAPHORA_ANTECEDENT_ DIFFERENT	3176	55.90

<표 50> 주격 무형대용어 복원 분석 말뭉치 주석 내용 검증 오류 유형 및 통계

주격 무형대용어 복원 층위 주석 내용 검증 점수는 <표 51>과 같다.

층위	차수	주어 복원 ¹⁴⁾	생략어복원 ¹⁵⁾	종합 점수 ¹⁶⁾	비고
주격 무형대용 어 복원	문어 시범	90.56	81.41	73.73	
	구어 시범	89.27	59.33	52.96	
	1차 문어	88.39	81.89	72.39	
	2차 문어	89.21	77.82	69.42	
	3차 구어	88.36	66.67	58.92	

<표 51> 주격 무형대용어 복원 말뭉치 주석 내용 검증 점수

13) 해당 검증 차수의 형식 오류 및 내용 오류의 합에서 형식 오류 비율. 소수 셋째 자리에서 반올림하였다.

14) F1 점수

15) 정확률

16) 주어 복원, 정확률의 곱

3.3.5. 상호참조 해결

상호참조 해결 주석 내용 오류 검증 항목은 <표 52>와 같다. 상호참조 해결 주석 내용 오류 검증 항목은 문서 내 모든 문장의 공지시 관계 군집 형성 일치 여부를 검사한다. 검증 지표는 공지시 관계 군집 형성 여부의 F1 점수로 최종 주석 일치도를 산출한다.

MUC F1 점수는 Vilain et al. (1995)¹⁷⁾에서 제안된 점수 산출 방식이다. MUC 점수 산출 방법은 공지시 관계 개체의 집합을 개체 간의 이음 집합으로 보고, 이음을 얼마나 잘 주석했는지를 평가하는 집합이다. MUC 점수의 재현율을 검증 대상 분석 말뭉치의 집합이 상위 집합이 되는 데 필요로 하는 이음의 개수로 점수를 산출한다.

검증 항목	검증 대상	검증 지표
공지시 관계 군집 형성 일치	문서 내 모든 문장	MUC F1 점수

<표 52> 상호참조 해결 주석 내용 오류 검증 항목

17) Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995, November). A model-theoretic coreference scoring scheme. In Proceedings of the 6th conference on Message understanding (pp. 45-52). Association for Computational Linguistics.

Class	입력	출력	설명
Mention CheckerFn	주식 표준 형식 상호참조 해결 말뭉치	결과 로그 파일 (Mention 정밀도, 재현율, F1 점수)	<ul style="list-style-type: none"> - Mention 정밀도: 검증 대상 분석 말뭉치의 주식 개체들이 검증용 말뭉치의 주식 개체를 기준 일치 여부 확인 - Mention 재현율: 검증 대상 분석 말뭉치의 주식 개체들이 검증용 말뭉치의 주식 개체 기준 재현율 확인 - Mention F1 점수: Mention 정밀도와 Mention 재현율의 조화 평균
MUCFn		결과 로그 파일 (MUC 정밀도, 재현율, F1 점수)	<ul style="list-style-type: none"> - MUC 정밀도: 검증용 말뭉치의 공지시 관계 집합이 검증 대상 분석 말뭉치의 공지시 관계 집합의 상위 집합이 되는 데 필요한 이음의 개수를 평가 - MUC 재현율: 검증 대상 분석 말뭉치의 공지시 관계 집합이 검증용 말뭉치의 공지시 관계 집합의 상위 집합이 필요한 이음의 개수를 평가 - MUC F1 점수: MUC 정밀도와 MUC 재현율의 조화 평균 산출

〈표 53〉 주식 내용 오류 검출 모듈 자료 구조

형식 오류 코드	설명
CR_ONLY_GOLD_MENTION	- 검증용 말뭉치에 주석된 개체가 검증 대상 분석 말뭉치에서 주석되지 않은 경우
CR_ONLY_PRED_MENTION	- 검증용 말뭉치에 주석되지 않은 개체를 검증 대상 분석 말뭉치에서 주석한 경우
CR_WRONG_SPLIT	- 공지시 관계의 개체를 같은 공지시 관계 군집에 포함하지 않은 경우
CR_WRONG_MERGE	- 공지시 관계가 아닌 개체를 같은 공지시 관계 군집에 포함한 경우

〈표 54〉 층위별 주식 내용 검증 오류 코드

상호참조 해결 검증 대상 분석 말뭉치는 문어 2회, 구어 2회 등 총 4회에 걸쳐 검증을 수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 (단위: %)
상호참조 해결	문어 시범	어절	20,405	20,405	100
	구어 시범	어절	10742	10742	100
	3차 문어	어절	2000215	140633	7.03
	3차 구어	어절	1011774	70761	6.99

〈표 55〉 상호참조 해결 말뭉치 형식 및 주석 내용 검증 대상 통계

상호참조 해결 층위 주석 내용 검증 오류 유형 및 통계는 〈표 56〉과 같다. 상호참조 해결 층위에서는 검증 대상 분석 말뭉치와 검증용 말뭉치의 멘션 주석을 비교하여, 검증용 말뭉치만 있는 멘션을 CR_ONLY_GOLD_MENTION, 검증 대상 분석 말뭉치에만 있는 멘션을 CR_ONLY_PRED_MENTION 오류로 분류하였다. 두 말뭉치에서 동일하게 주석된 멘션들 중 검증용 말뭉치 기준으로 군집의 개체 구성 일치 여부에 따라서 CR_WRONG_SPLIT, CR_WRONG_MERGE 오류로 분류하였다.

층위	차수	오류 유형	빈도	비율
상호참조 해결	문어 시범	CR_ONLY_GOLD_MENTION	1639	70.67
		CR_ONLY_PRED_MENTION	637	27.47
		CR_WRONG_SPLIT	24	1.03
		CR_WRONG_MERGE	19	0.81
	구어 시범	CR_ONLY_GOLD_MENTION	460	60.85
		CR_ONLY_PRED_MENTION	287	37.96
		CR_WRONG_SPLIT	7	0.93
		CR_WRONG_MERGE	2	0.26
	3차 문어	CR_ONLY_GOLD_MENTION	12107	69.50
		CR_ONLY_PRED_MENTION	4958	28.46
		CR_WRONG_SPLIT	143	0.82
		CR_WRONG_MERGE	92	0.53
	3차 구어	CR_ONLY_GOLD_MENTION	2685	61.87
		CR_ONLY_PRED_MENTION	1562	35.99
		CR_WRONG_SPLIT	53	1.22
		CR_WRONG_MERGE	38	0.88

〈표 56〉 상호참조 해결 분석 말뭉치 주석 내용 검증 오류 유형 및 통계

상호참조 해결 층위 주석 내용 검증 점수는 〈표 57〉과 같다.

층위	차수	멘션 F1 점수	MUC F1 점수	비고
상호참조 해결	문어 시범	71.70 ¹⁸⁾	68.33 ¹⁹⁾	
	구어 시범	65.56 ²⁰⁾	65.36 ²¹⁾	
	3차 문어	69.19 ²²⁾	68.20 ²³⁾	
	3차 구어	60.90 ²⁴⁾	59.44 ²⁵⁾	

〈표 57〉 상호참조 해결 말뭉치 주석 내용 검증 점수

3.3.6. 구문 분석

구문 분석 주석 내용 오류 검증 항목은 〈표 58〉과 같다. 구문 분석 주석 내용 오류 검증 항목은 문장 내 지배소 주석 일치 여부와 지배소 주석이 일치한 어절의 구문 및 기능 태그 부여 일치 여부를 검사한다. 검증 지표는 문장 내 지배소 주석과 태그 부여를 동시에 산출하는 LAS(Labeled Attachment Score) 점수로 최종 주석 일치도를 산출한다.

검증 항목	검증 대상	검증 지표
지배소 일치	모든 어절	LAS 점수
구문/기능 태그 일치	지배소 주석이 일치한 어절	

〈표 58〉 구문 분석 주석 내용 오류 검증 항목

Class	입력	출력	설명
DepContent CheckerFn	주석 표준 형식 구문 분석 말뭉치	결과 로그 파일 (LAS 점수)	<ul style="list-style-type: none"> - 검증용 말뭉치와 검증 대상 분석 말뭉치의 ‘DP’의 ‘word’ 순서대로 비교 - ‘word’ 간 비교는 ‘label’과 ‘head’에 대하여 실시 - 두 말뭉치의 ‘label’과 ‘head’이 모두 일치하는지 확인

〈표 59〉 주석 내용 오류 검출 모듈 자료 구조

18) 정밀도 81.90, 재현율 63.75

19) 정밀도 78.66, 재현율 60.40

20) 정밀도 71.24, 재현율 60.72

21) 정밀도 68.71, 재현율 62.33

22) 정밀도 79.44, 재현율 61.28

23) 정밀도 76.79, 재현율 61.34

24) 정밀도 67.93, 재현율 55.20

25) 정밀도 64.96, 재현율 54.78

형식 오류 코드	설명
DEPENDENCY_HEAD_ERROR	- 문장 내 지배소가 불일치한 경우
DEPENDENCY_LABEL_ERROR	- 문장 내 지배소가 일치한 어절의 구문/기능 태그가 불일치한 경우

〈표 60〉 층위별 주석 내용 검증 오류 코드

구문 분석 검증 대상 분석 말뭉치는 문어 4회로 총 4회에 걸쳐 검증을 수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 (단위: %)
구문 분석	문어 시범	어절	20,405	20,405	100
	1차 문어	어절	581513	10133	1.74
	2차 문어	어절	600077	24769	4.13
	3차 구어	어절	798221	70650	8.85

〈표 61〉 구문 분석 말뭉치 형식 및 주석 내용 검증 대상 통계

구문 분석 층위 주석 내용 검증 오류 유형 및 통계는 〈표 62〉와 같다. 구문 분석 층위에서는 문장 내 지배소 주석이 일치한 어절에 대해서 구문 태그 부착까지 수행하였다. 주석 단위 내에서 DEPENDENCY_HEAD_ERROR가 발생한 경우, 구문 태그 부착 검사를 수행하지 않았다. 따라서 DEPENDENCY_LABEL_ERROR는 주석 단위 내에서 지배소가 일치한 어절 중에서 구문 태그를 다르게 주석한 경우이다.

층위	차수	오류 유형	빈도	비율 ²⁶⁾
구문 분석	문어 시범	DEPENDENCY_HEAD_ERROR	1586	66.47
		DEPENDENCY_LABEL_ERROR	800	33.53
	1차 문어	DEPENDENCY_HEAD_ERROR	647	79.68
		DEPENDENCY_LABEL_ERROR	165	20.32
	2차 문어	DEPENDENCY_HEAD_ERROR	3190	85.34
		DEPENDENCY_LABEL_ERROR	548	14.66
	3차 문어	DEPENDENCY_HEAD_ERROR	8045	81.73
		DEPENDENCY_LABEL_ERROR	1789	18.27

〈표 62〉 구문 분석 말뭉치 주석 내용 검증 오류 유형 및 통계

26) 해당 검증 차수의 형식 오류 및 내용 오류의 합에서 형식 오류 비율이며, 소수 셋째 자리에서 반올림하였다.

구문 분석 층위 주석 내용 검증 점수는 <표 63>과 같다.

층위	차수	LAS 점수	비고
구문 분석	문어 시범	89.26	
	1차 문어	92.53	
	2차 문어	85.89	
	3차 구어	86.94	

<표 63> 구문 분석 말뭉치 주석 내용 검증 점수

3.3.7. 의미역 분석

의미역 분석 주석 내용 오류 검증 항목은 <표 64>와 같다. 의미역 분석 주석 내용 오류 검증 항목은 문장 내 의미역 부여 대상 서술어 주석 일치 여부와 주석이 일치한 서술어의 논항 주석 일치 여부를 검증한다. 검증 지표는 문장 내 서술어 주석 일치 여부와 논항 주석 일치 여부의 F1 점수로 최종 주석 일치도를 산출한다.

검증 항목	검증 대상	검증 지표
의미역 부여 대상 서술어 주석 일치	모든 어절	F1 점수
논항 주석 일치	서술어 주석이 일치한 문장 의 논항 주석 내용	

<표 64> 의미역 분석 주석 내용 오류 검증 항목

Class	입력	출력	설명
SRLContent CheckerFn	주석 표준 형식 의미역 분석 말뭉치	결과 로그 파일 (PI, AI, PIC 점수)	<ul style="list-style-type: none"> - PI(Predicate Identification): 검증용 말뭉치와 검증 대상 말뭉치의 문장별 ‘predicate’ 주석 일치 비교 - AI(Argument Identification): 문장 내 일치한 ‘predicate’ 마다 ‘argument’ 주석 일치 검사 실시 - PIC(Predicate Identification Classification): 서술어의 ‘sense_id’ 비교. 주석 내용 점수 산출에서는 제외.

<표 65> 주석 내용 오류 검출 모듈 자료 구조

형식 오류 코드	설명
SRL_PI_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 ‘predicate’ 가 불일치한 경우
SRL_AI(core)_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 동일 ‘predicate’ 의 필수역 주석이 불일치한 경우
SRL_PIC_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 ‘predicate’ 의 ‘sense_id’ 가 불일치한 경우
SRL_AI(all)_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 동일 ‘predicate’ 의 필수역과 부가역 주석이 불일치한 경우
SRL_AI(span)_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 동일 ‘predicate’ 의 범위가 불일치한 경우 (‘begin’, ‘end’ 기준)
SRL_AI(span-begin)_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 동일 ‘predicate’ 의 범위가 불일치한 경우 (‘begin’ 기준)
SRL_AI(span-end)_ERROR	- 검증용 말뭉치와 검증 대상 분석 말뭉치의 동일 문장 내 동일 ‘predicate’ 의 범위가 불일치한 경우 (‘end’ 기준)

〈표 66〉 층위별 주석 내용 검증 오류 코드

의미역 분석 검증 대상 분석 말뭉치는 문어 2회로 총 2회에 걸쳐 검증을 수행하였다.

층위	차수	단위	검증 대상 분석 말뭉치 (단위: 개)	검증용 말뭉치 (단위: 개)	주석 내용 검증 비율 (단위: %)
의미역 분석	문어 시범	어절	20,405	20,405	100
	문어 3차	어절	2000215	140633	7.03

〈표 67〉 의미역 분석 말뭉치 형식 및 주석 내용 검증 대상 통계

의미역 분석 층위 주석 내용 검증 오류 유형 및 통계는 〈표 68〉과 같다. 의미역 분석 층

위에서는 의미역 부여 대상 서술어가 일치한 문장에 대해서 논항 주석 일치 검사까지 수행하였다. 주석 단위 내에서 SRL_PI_ERROR가 발생한 경우, 해당 서술어의 논항 일치 검사는 수행하지 않았다. 따라서 SRL_AI_ERROR는 주석 단위 내에서 서술어 주석이 일치한 의미역 주석 중에서 논항 주석을 다르게 한 경우이다. SRL_AI(all)_ERROR의 경우 모든 논항을 대상으로 일치 검사를 수행한 결과이며, SRL_AI(core)_ERROR의 경우 필수역에 대상으로 일치 검사를 수행한 결과이다.

층위	차수	오류 유형 ²⁷⁾	빈도	비율 ²⁸⁾
의미역 분석	문어 시범	SRL_PI_ERROR	1302	3.34
		SRL_AI(all)_ERROR	3093	17.92
		SRL_AI(core)_ERROR	2288	13.26
	문어 3차	SRL_PI_ERROR	2297	2.18
		SRL_AI(all)_ERROR	16512	15.70
		SRL_AI(core)_ERROR	10132	9.64

〈표 68〉 의미역 분석 말뭉치 주석 내용 검증 오류 유형 및 통계

의미역 분석 층위 주석 내용 검증 점수는 〈표 69〉과 같다.

층위	차수	PI	AI(all)	AI(core)	비고
의미역 분석	문어 시범	87.47	60.61	70.33	
	문어 3차	91.70	64.74	75.63	

〈표 69〉 의미역 분석 말뭉치 주석 내용 검증 점수

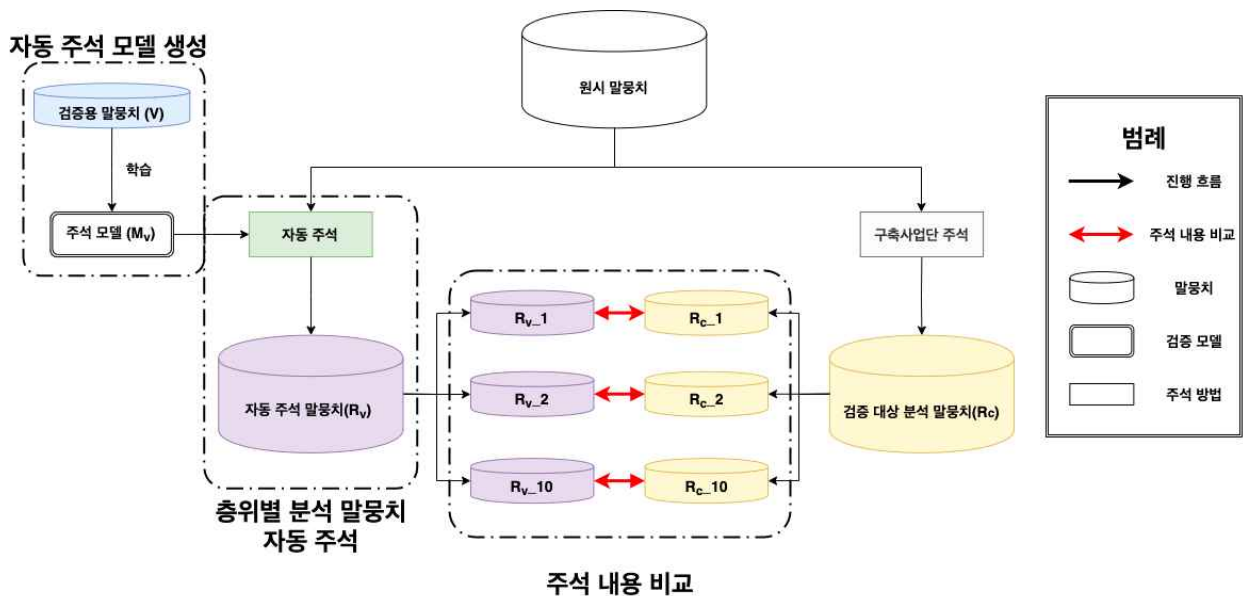
27) 의미역 분석 주석 내용 오류 유형 중 검증 대상 분석 말뭉치 평가 시에는 〈표 〉의 세 가지 오류 유형으로만 평가하였음. 세 가지 오류 유형 이외의 오류 유형은 국립국어원의 요청으로 검토 시 참고 자료로 추가한 오류 유형임. 모든 오류 정보는 오류 보고서를 참조하라.

28) 해당 검증 차수의 형식 오류 및 내용 오류의 합에서 형식 오류 비율이며 소수 셋째 자리에서 반올림하였다.

4. 일관성 검증

7개 층위 검증 대상 분석 말뭉치의 주석 일관성 검증은 검증용 말뭉치의 주석 경향성이 검증 대상 분석 말뭉치에도 유사하게 나타나는지 확인하여 간접적으로 평가하였다.

검증 대상 분석 말뭉치의 주석 일관성 검증을 위해 검증용 말뭉치를 학습 데이터로 사용한 층위별 자동 주석 모델(M_V)을 만들고, 이를 검증용 말뭉치 주석 범위가 아닌 원시 말뭉치 범위에 자동 주석을 수행한 새로운 말뭉치(이하 자동 주석 말뭉치)를 구축하였다. 그리고 자동 주석 말뭉치와 검증 대상 분석 말뭉치를 비교 검증하여 주석 내용 검증을 하였다([그림 53]).



[그림 53] 주석 일관성 검증 흐름도

자동 주석 말뭉치(R_V)와 검증 대상 분석 말뭉치(R_C)를 10개 구간(R_{V_1} , R_{V_2} , ..., R_{V_10} ; R_{C_1} , R_{C_2} , ..., R_{C_10})으로 나눈 하위말뭉치를 만들어 구간별 주석 내용 일치도를 평가하여 검증용 말뭉치의 주석 특징이 검증 대상 분석 말뭉치 전체에 걸쳐 나타나는지 주석 내용 일치도의 경향으로 평가하였다. 두 말뭉치의 10개 구간의 주석 일치도의 평균을 구하고, 각 구간의 관측값과 평균값을 비교하여 99% 신뢰 구간 (confidence alpha = 0.01)을 벗어날 경우, 해당 구간은 주석 일관성이 다른 구간에 비해 낮다고 평가하였다.

검증 대상 분석 말뭉치의 주석 일관성 검증을 위해 검증용 말뭉치를 학습한 모델 정보는 <표 69>와 같다.

층위	문/구어	검증 모델 설명	참고문헌
형태 분석	문어	- 어절의 분절을 고려하지 않고, 형태소에 주석된 형태소 태그의 연속적 조합을 고려한 모델(sequence labeling)	Ma and Hovy (2016)
	구어	- 통계 기반의 학습 모델로, 태그를 부착하고자 하는 형태소의 이전의 형태소에 부착된 가장 확률이 높은 태그를 부여 ²⁹⁾	-
어휘의미 분석	공통	- bi-gram을 활용한 통계 기반 학습 모델 - 어휘의미 부여 대상 단어에 선행한 형태소에 따라 어떤 어휘의미 번호를 부여했는지에 대한 통계 자료를 학습 데이터로 구축 - 어휘의미 부여 대상 단어가 속한 문장의 패턴 통계에 따라 가장 높은 확률을 가지는 sense_id를 부여 - 해당 패턴이 처음 등장하지 않은 경우, 그 단어에서 가장 많이 부여된 어휘의미 번호 부여	-
개체명 분석	공통	- 통계 기반의 학습 모델 - 검증용 말뭉치 내에서 찾은 개체명의 표층형과 개체 타입의 패턴을 학습	-
주격 무형대용어 복원	공통	- 생략어복원 대상 서술어와 생략어 사이에 같은 이음 일련번호(chain id)를 부여하여서 한 집단으로 학습 - 자동 주석 결과 생략어복원 대상 서술어와 서술어가 아닌 범위가 같은 이음 일련번호를 가지면 생략어를 복원하였다고 해석	Joshi et al. (2019)
상호참조 해결	공통	- 주격 무형대용어와 같은 모델을 사용 - 개체(mention)에 대한 단위를 어절 단위로 지정	Joshi et al. (2019)
구문 분석	문어	- 구문 분석 자동 분석기 UDPipe 사용	Straka et al. (2016)
의미역 분석	문어	- BERT 기반 SRL 모델 학습 - 논항 인식 F1 점수 산출	Lee et al. (2015); Bae et al. (2017)

<표 69> 층위별 주석 일관성 검증 모델

29) $P(now\ tag|now\ text, pre\ tag) \propto \frac{freq(now\ tag \cap now\ text)}{freq(now\ text)} * \frac{freq(now\ tag \cap pre\ text)}{freq(pre\ text)}$

7개 층위 주석 일관성 결과는 <표 70>과 같다. 자동 주석 말뭉치와 검증 대상 분석 말뭉치를 10개 구간으로 분할하여, 각 구간의 주석 내용 일치도를 산출하여 간접적으로 주석 일관성 검증을 수행하였다. 말뭉치 구간 문어 말뭉치는 문서 일련번호 기준, 구어 말뭉치는 파일 일련번호를 기준으로 순차적으로 동일한 분량으로 분할하였다.

층위		평가 지표	분할 구간										구간 평균	구간 길이
			1	2	3	4	5	6	7	8	9	10		
형태	문어	정확률	49.02	49.04	48.65	49.22	48.79	49.26	49.78	48.70	49.48	48.65	49.06	0.69
	구어	정확률	75.35	73.75	76.98	79.30	78.28	79.24	77.40	78.99	80.16	78.32	77.88	3.68
어휘의미	문어	정확률	87.64	86.81	87.40	87.45	87.81	87.99	87.47	87.95	87.41	88.14	87.61	0.70
	구어	정확률	77.72	79.31	79.51	79.86	78.89	79.20	79.66	78.39	79.09	80.75	79.24	1.50
개체명	문어	정확률	22.04	22.10	21.99	22.58	22.10	22.18	22.26	21.97	22.37	22.69	22.23	0.45
	구어	정확률	27.45	20.03	20.10	20.54	21.62	20.55	22.54	19.15	22.06	21.20	21.52	4.21
주격무형	문어	FI 접수	13.98	8.55	20.84	38.75	29.30	33.35	31.41	33.31	22.82	31.34	26.37	17.35
	구어	FI 접수	23.70	22.78	22.72	24.01	23.71	24.08	22.80	23.07	22.81	24.02	23.37	1.06
상호참조	문어	MUC FI접수	51.32	50.08	50.34	49.66	51.13	51.23	51.36	50.91	49.77	51.09	50.89	1.20
	구어	MUC FI접수	36.39	33.93	33.71	38.42	42.07	33.57	38.25	34.20	33.89	32.34	35.68	5.52
구문	문어	UAS	68.37	68.77	68.11	68.68	68.43	69.00	68.92	69.03	69.73	67.18	68.62	1.22
의미역	문어	FI접수	61.91	61.30	61.14	61.34	61.06	61.10	60.78	61.15	61.25	61.18	61.22	0.52

<표 70> 층위별 일관성 검증 결과

말뭉치 분할 구간 별 자동 주석 말뭉치와 검증 대상 분석 말뭉치의 주석 내용 일치도 평균값 대비 99% 신뢰구간(CI)은 문어 말뭉치³⁰⁾가 구어 말뭉치보다 짧았다(평균 신뢰 구간 길이: 문어 = 3.603, 구어 = 3.193). 이로 미루어 보아 문어 말뭉치가 대체적으로 구어

30) 문어 말뭉치의 경우, 주격 무형대용어 복원 1, 2구간이 상대적으로 낮은 일치도를 보이면서 평균 신뢰 구간의 길이를 증가시키는 요인으로 작용하였다.

말뭉치에 비해서 주석 일관성이 높다고 평가할 수 있다. 뿐만 아니라, 문어 말뭉치의 경우 신문 기사 장르로만 구성된 반면, 구어 말뭉치는 공적 독백, 공적 대화, 사적 대화 세 가지의 장르로 구성되었기 때문에, 장르에 따른 주석 일관성 차이도 나타날 수 있다고 평가하였다.

가장 짧은 신뢰구간을 보인 말뭉치는 개체명 분석 말뭉치로 각 분할 구간의 자동 분석 말뭉치와 검증 대상 분석 말뭉치의 주석 내용 일치도의 차이가 가장 적다(99% 신뢰 구간 길이 = $.422$ ($21.98 \leq CI \leq 22.43$, $confidence\alpha = .202$, $\sigma = .248$). 또한, 99% 신뢰 구간을 벗어나는 구간도 1개 구간으로, 개체명 분석 문어 말뭉치의 주석 일관성은 다른 말뭉치에 비해 비교적 높다고 평가하였다.

층위별 주석 내용 검증 결과와 비교해 볼 때, 검증용 말뭉치와 검증 대상 분석 말뭉치의 주석 내용 일치도가 80점 이상인 경우³¹⁾, 주격 무형대용어 문어 말뭉치와 형태 분석 구어 말뭉치를 제외하고 99% 신뢰구간의 길이가 모두 1.5 이하로, 대체로 높은 주석 일관성을 보였다. 따라서 검증용 말뭉치와 검증 대상 분석 말뭉치의 주석 내용 일치도가 높을 경우, 나머지 범위의 검증 대상 분석 말뭉치의 주석 일관성도 높을 것이라고 간접적으로 평가하였다.

31) 형태 분석(문어, 구어), 어휘의미 분석(문어, 구어), 개체명 분석(문어, 구어), 주격 무형대용어 복원(문어, 구어), 구문 분석, 의미역 분석 말뭉치가 이에 해당한다. 자세한 내용은 3장의 2.2절을 참고하라.

5. 통합 검증

통합 검증은 7개 층위 검증 대상 분석 말뭉치가 다층위 분석 말뭉치로 활용될 수 있는지 검증하는 단계이다. 검증 대상 분석 말뭉치의 원시 말뭉치 보존 여부 및 주석 표준 양식을 준수하였는지 등을 확인한다. 7개 층위의 분석 말뭉치를 하나로 합치기 위해 각 말뭉치의 문서, 문단, 문장, 어절 수의 원시 말뭉치 대비 기초 통계량을 비교하고, 주석 표준 양식에 정의된 일련번호 양식의 준수 여부 등을 검사한다.

통합 검증의 기준은 말뭉치 형식은 주석 표준 양식, 원시 말뭉치 기초 통계량은 국립국어원에서 배포한 문어 말뭉치 및 구어 말뭉치의 기초 통계량을 따른다. 기초 통계량은 <표 4>, <표 5>를 참조하라.

7개 층위 검증용 말뭉치의 통합 검증 결과는 <표 71>과 같다. 문어와 구어 원시 말뭉치를 기준으로 하여 7개 층위의 검증용 말뭉치가 동일한 원시 말뭉치에 주석을 수행하였으며, 동일한 주석 표준 형식을 따랐으므로 다층위 말뭉치로서 사용이 가능함을 확인하였다.

층위	구분	문서 ³²⁾ / 과일 ³³⁾	문장 ³⁴⁾ / 발화 ³⁵⁾	어절	검증 결과
형태 분석	문어	520	10400	140633	통과
	구어	32	17551	70744	통과
어휘의미 분석	문어	520	10400	140633	통과
	구어	32	17551	70744	통과
개체명 분석	문어	520	10400	140633	통과
	구어	32	17551	70744	통과
주격 무형대용어 복원	문어	520	10400	140633	통과
	구어	32	17551	70744	통과
상호참조 해결	문어	520	10400	140633	통과
	구어	32	17551	70744	통과
구문 분석	문어	520	10400	140633	통과
의미역 분석	문어	520	10400	140633	통과

<표 71> 층위별 검증용 말뭉치 통합 검증 결과

32) 문어 말뭉치

33) 구어 말뭉치

34) 문어 말뭉치

35) 구어 말뭉치

7개 층위 검증 대상 분석 말뭉치의 통합 검증 결과는 <표 72>와 같다. 문어와 구어의 원시 말뭉치를 기준으로 하여, 층위별 분석 말뭉치의 원시 말뭉치와의 일치 검사를 수행하여 다층위 말뭉치로서 사용 가능한 지 검증하였다.

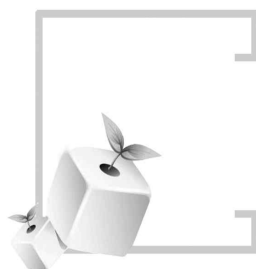
<표 72>에 따르면 검증 대상 분석 말뭉치 중 형태 분석 문어, 어휘의미 문어, 상호참조 문어 말뭉치에서 원시 말뭉치 대비 어절 수가 미달하였고, 문장 수가 초과하였다. 이는 원시 말뭉치에 문장 분할 오류로 인해 어절 수와 문장 수에 차이가 생긴 것으로 확인되어 이를 구축사업단에 수정을 요청하였다.

주격 무형대용어 복원 층위에서는 어절 수와 문장 수가 모두 초과하였는데, 이는 문어 말뭉치 주석 대상이 아닌 <byline> 태그 내부의 정보가 포함되면서 차이가 발생한 것으로 확인되어 이를 구축사업단에 수정을 요청하였다.

구문 분석 말뭉치 층위에서는 어절 수와 문장 수는 일치하였으나, 두 문장에서 어절 수와 실제 주석된 어절 수가 다른 것으로 확인되어 이를 구축사업단에 수정을 요청하였다.

층위	구분	문서/파일	문장/발화	어절	검증 결과
형태 분석	문어	7265	150085	2000213	어절 미달, 문장 초과
	구어	423	223962	1006447	통과
어휘의미 분석	문어	7265	150085	2000213	어절 미달, 문장 초과
	구어	423	223962	1006447	통과
개체명 분석	문어	7265	150082	2000215	통과
	구어	423	223962	1006447	통과
주격 무형대용어 복원	문어	7265	150338	2019322	문장, 어절 수 초과
	구어	423	223962	1006447	통과
상호참조 해결	문어	7265	150085	2000213	어절 미달, 문장 초과
	구어	423	223962	1006447	통과
구문 분석	문어	7265	150085	2000215	문장 당 어절 수 불일치
의미역 분석	문어	7265	150085	2000215	통과

<표 72> 층위별 검증 대상 분석 말뭉치 통합 검증 결과



제 3 장

원문 자료 검증 및 원시 말뭉치 구축



1. 원문 자료 검증

문어 및 신문 원문 말뭉치 구축사업단에서 수집한 결과물에 대한 원문 자료 검증은 다음의 세 가지 검증을 수행한다.

○ 원문 자료 검증

(1) 인코딩 검사

(2) XML 형식 검사

(3) 데이터 유효성 검사

(1) 인코딩 검사

원문 말뭉치의 문자 인코딩 규격은 UTF-8으로 명시되어 있다. 제출된 자료가 인코딩 규격을 준수하지 않는다면 이후의 검증과정에서 예기치 않은 부작용을 일으킬 수 있다. 인코딩 검사는 말뭉치 검증과정에서 가장 필수적인 요소이면서 가장 먼저 수행되어야 하는 절차다.

본 사업단에서는 ICU (International Components for Unicode)의 문자 집합 탐색 알고리즘에 기반을 둔 인코딩 추론 모듈을 사용하였다. 원문 자료 파일로부터 4,096바이트를 표본 추출하여 인코딩 추론을 수행하고 최상위 후보가 UTF-8인지 점검하였다.

(2) XML 형식 검사

원문 말뭉치의 제출물은 XML 형식을 따른다. 본 사업단에서는 W3C (World Wide Web Consortium)의 XML 명세 1.0 제5판에 근거하여 원문 자료의 XML 형식이 적형인지를 검사하였다.

XML 검사에서 적형이 아닌 것으로 평가된 자료의 경우 문서 내에서 문제가 되는 토큰과 해당 토큰이 위치하는 줄의 번호, 오류의 유형을 로그 파일에 기록하였다.

(3) 데이터 유효성 검사

원문 말뭉치는 문어와 신문 두 가지 유형으로 구분된다. 문어와 신문 말뭉치는 <header>가 요구하는 정보, <text> 요소의 속성 등에서 차이가 있다. 한편 문어 말뭉치는 장르 특성에 따라 네 가지의 하위 유형으로 분류되는데, 책-상상, 책-정보, 기타의 경우 서로 같은 데이터 구조를 공유하지만, 잡지의 경우 <subclass> 데이터에 명시적인 값이 반드시 없어야 하는 등 다른 장르와 구분되는 특징을 가지고 있다. 데이터 스키마 집합은 말뭉치별 특성에 따라 아래 <표 73>과 같이 세 종류로 구분하여 작성하였다.

말뭉치 유형	매체/장르 분류	분류 코드	스키마 코드
문어	책-상상	WAOR	WXOR
	책-정보	WBOR	
	기타	WZOR	
	잡지	WCOR	WCOR
신문	전국 종합지	NWOR	NXOR
	지역 종합지	NLOR	
	전문지	NPOR	
	인터넷 기반 신문	NIOR	
	기타	NZOR	

<표 73> 원문 말뭉치의 스키마 유형

스키마는 [그림 54]와 같이 JSON 형식으로 작성하였다. 스키마에는 해당 요소가 요구하는 하위 요소 및 속성의 종류가 기술된다. 또한, 자릿값의 유형을 열거하거나 패턴 제약 등을 통해 유효한 값의 범위를 지정한다. 데이터가 스키마를 위배하는 경우 데이터의 고유 주소 ‘Sid’ 와 함께 위배한 내용을 로그 파일에 기록하였다.

WXOR의 fileInfo 요소 정의	WXOR의 fileId 요소 정의
<pre> { "fileInfo": { "\$id": "#/SJML/header/fileInfo", "title": "The WXOR FileInfo Schema", "type": "object", "required": ["fileId", "annoLevel", "sampling", "class", "subclass"] } } </pre>	<pre> { "fileId": { "\$id": "#/SJML/header/fileInfo/fileId", "title": "The WXOR FileId Schema", "type": "string", "default": "", "examples": ["WBOR1900000014"], "pattern": "^(W[ABZ]OR19\\d{8})\$" } } </pre>

[그림 54] 원문 말뭉치의 스키마 예시

위 세 가지 하위 검사 결과에 따라 결과 유형을 <표 74>처럼 구분하고 각 유형에 해당하는 파일의 수량 및 비율을 점검하였다.

결과 유형	인코딩 검사	XML 형식 검사	데이터 유효성 검사
PPP	PASS	PASS	PASS
PP	PASS	PASS	FAIL
P	PASS	FAIL	NA
F	FAIL	NA	NA

<표 74> 하위 검사 적합 판정에 따른 검증 결과 유형

문어의 경우 <표 75>에서 확인할 수 있듯이 XML 검사를 통과하는 못하는 경우(58.59%)가 통과하는 경우보다 더 많았다. 형식 검사를 통과하지 못하는 원인으로서는 닫는 태그를 부착하지 않는 등 요소 및 속성을 기술하는 문법에 문제가 있는 경우도 있었지만 대부분의 경우는 CDATA와 PCDATA 영역을 구분하지 않아 발생하는 문제였다 태그 사이의 문자열은 어떤 명시적인 조치가 없는 한 기본적으로 PCDATA로 인식되는데 해당 영역에 ‘&’, ‘<’, ‘>’ 등 XML 파싱에 영향을 미치는 특수 문자가 출현하여 검증을 통과하지 못한 것이다.

신문의 경우 모든 파일이 XML의 형식을 잘 갖추고 있었다. 그러나 약 30%에 가까운 파일들은 데이터 유효성 검사를 통과하지 못하였다. 유효성 검사와 관련하여 신문 자료의 가장 큰 이슈는 공백 문자 처리였다. 연속된 공백 문자로 인해 유효성 검증에 걸리는 경우가 가장 많았다.

	F	P	PP	PPP	전체
문어	0 (0%)	11,748 (58.59%)	15 (0.08%)	8,289 (41.33%)	20,052 (100%)
신문	0 (0%)	0 (0%)	123 (30.38%)	282 (69.62%)	405 (100%)

〈표 75〉 원문 말뭉치 유형별 검증 결과: 파일 수 (단위: 개)

2. 원시 말뭉치 구축

문어 및 신문 원문 자료의 검증이 끝나면 원문 자료를 원시 말뭉치로 변환하는 작업을 진행하였다. 원시 말뭉치 구축의 전반적인 절차는 다음과 같다.

○ 원시 말뭉치 구축

- (1) 전처리 및 XML 디코딩
- (2) 데이터 변환
- (3) 어절 수 측정 및 XML 인코딩

(1) 전처리 및 XML 디코딩

전처리 작업은 XML을 디코딩하기 위하여 비적형의 XML 문서를 적형으로 변환하는 절차다. XML 형식 검사에서 발견되는 문제는 아래와 같이 PCDATA로 평가되는 영역에 ‘&’, ‘<’, ‘>’ 등과 같은 XML의 특수 기호들이 출현하여 발생하는 것이 대부분이다.

○ <author>홍길동 & 김철수</author>

이러한 문제에 대응하기 위해 문자열 값이 입력되는 말단 태그의 요소들에 대하여 아래와 같이 CDATA를 명시적으로 선언하도록 변경하였다. 전 처리된 자료들은 <XML> 구조에 대응하는 해시 테이블로 디코딩하였다.

○ <author> → <author><![CDATA[

○ </author> →]]></author>

(2) 데이터 변환

파싱된 원문 데이터는 원시 말뭉치의 구조와 스키마에 맞게 정보를 추가하고 구조를 변경하였다. 문어 원문 자료의 경우 문단 경계 정보를 추가하고 <p> 요소를 생성하였다. 신문의 경우 <text> 요소의 일부 속성 명을 변경하였다.

원시 말뭉치 지침은 <p> 데이터 내에 불필요한 공백 문자를 허용하지 않는다. 따라서 문자열 내의 탭 문자 및 문자열 양 끝의 공백을 제거하고 공백 문자가 연속되는 경우 하

나의 공백 문자로 치환하였다.

XML 특수 문자에 해당하는 ‘&’, ‘<’, ‘>’ 는 이스케이프 처리하였다. 경우에 따라 이미 이스케이프 처리가 되어 구축된 원문 자료가 있을 수 있으므로 이스케이프 표현에 대한 역변환을 먼저 수행하였다.

끝으로 <fileId> 및 <annoLevel>을 원시 데이터 형식으로 변환하였다.

(3) 어절 수 측정 및 XML 인코딩

변경 완료된 원시 자료에 대하여 어절 수를 측정하였다. 문어 및 신문 원시 말뭉치의 어절 수는 <p> 요소가 가진 토큰 수를 합산하여 산출하였다. 토큰은 공백 문자를 구분자로 하여 분리하였다. 최종 결과물 파일은 XML 형식과 UTF-8 인코딩을 준수하여 생성하였다. 파일명은 <fileId>를 따라 부여하였고, 확장자명은 .sjml로 통일하였다.

원시 말뭉치 구축 과정은 XML 특수 문자 처리, 문단 분리, 공백 문자 처리, 속성명 변경, 어절 수 측정, XML 파일 생성 등을 포함한다. 원시 구축에 대한 결과는 최종적으로 XML 형식의 원시 말뭉치 파일이 생성되었는지의 여부에 따라 PASS/FAIL로 구분하였다.

원문 자료의 원시 말뭉치 변환은 최종 결과 산출까지 순조롭게 작업되었다. 문어와 신문 모두 모든 원문 자료에 대해 원시 말뭉치 결과물을 얻을 수 있었다. 어절 규모는 문어가 6억 7천여 어절로, 신문이 10억 6천여 어절로 집계되었다.

	F	P	전체	어절 수
문어	0 (0%)	20,052 (100%)	20,052 (100%)	676,385,790
신문	0 (0%)	405 (100%)	405 (100%)	1,060,117,683

<표 76> 원시 말뭉치 구축 결과: 파일 수 (단위: 개)

3. 원시 말뭉치 검증

총 6종의 원시 말뭉치, 문어, 신문, 일상대화, 구어/준구어, 웹, 메신저 말뭉치에 대한 검증 작업을 수행하였다. 검증 절차는 다음과 같다.

○ 원시 말뭉치 검증

- (1) 인코딩 검사
- (2) XML 형식 검사
- (3) 데이터 유효성 검사
- (4) 발화 요소 오류 탐지
- (5) 검증 결과 로그 생성

(1) 인코딩 검사 및 (2) XML 형식 검사는 1장의 원문 검증의 절차와 크게 다르지 않다. 본 장에서는 원시 자료에 대한 (3) 데이터 유효성 검사와 (4) 발화 요소에 대한 오류 탐지 방법을 중심으로 기술하도록 하겠다. 이와 함께 1장에서 자세하게 다루지 않은 검증 결과 로그 방식을 소개한다.

(3) 데이터 유효성 검사

원시 말뭉치도 원문 말뭉치와 마찬가지로 말뭉치 유형별에 따라 데이터 구조에 차이가 있다. 일부 말뭉치들은 매체 및 장르 구분에 따라서도 데이터 구조가 구분된다. 이럴 때 스키마는 말뭉치의 하위 유형에 따라 달리 작성되었다. 원시 검증을 위해 총 8종의 스키마 집합을 작성하였다. <표 77>은 원시 말뭉치의 스키마 유형으로 말뭉치 및 하위말뭉치 유형과의 대응 관계를 확인할 수 있다.

말뭉치 유형	매체/장르 분류	분류 코드	스키마 코드
문어	책-상상	WARW	WXRW
	책-정보	WBRW	
	기타	WZRW	
	잡지	WCRW	WCRW
신문	전국 종합지	NWRW	NXRW
	지역 종합지	NLRW	
	전문지	NPRW	
	인터넷 기반 신문	NIRW	
	기타	NZRW	
일상대화	일상대화	SDRW	SDRW
구어/준구어	공적 독백	SARW	SXRW
	공적 대화	SBRW	
	대본	SERW	SERW
웹	누리소통망	ESRW	EXRW
	블로그	EBRW	
	게시글	EPRW	
	리뷰	ERRW	
메신저	2인대화	MDRW	MXRW
	다자대화	MMRW	

〈표 77〉 원시 말뭉치의 스키마 유형

스키마는 JSON 형식으로 기술하였다. 형식상으로는 원문 말뭉치와 차이가 없지만, 내용상으로는 더 세분된 명세를 하고 있다. 원시 말뭉치의 요소 및 속성이 원문 자료보다 다양하며 구조의 계층 또한 상대적으로 깊다. 또한, 문단 <p>, 발화 <u>를 비롯하여 일상대화의 <note>, 신문 말뭉치의 <text>, 웹 말뭉치의 <SGML>, 준구어의 <episode>, <scene>, <sp>, <stage> 등의 요소는 요소들이 복수로 출현하여 열거된다는 특징이 있다. 〈그림 32〉는 원시 말뭉치의 스키마 중 SERW의 p 요소에 대한 명세로, 고유 주소(\$id)를 보면 깊은 계층과 열거 가능성이 만들어 내는 구조적인 복잡성을 확인할 수 있다.

SERW의 p 요소 정의

```
{
  "p": {
    "$id": "#/SJML/text/episode/index/scene/index/sp/index/p",
    "title": "The SERW P Schema",
    "type": "string",
    "examples": [
      "양태야!!...길상아!!"
    ],
    "allof": [
      { "pattern": "^(.+)$" },
      { "not": { "pattern": "^(\\s)" } },
      { "not": { "pattern": "(\\s)$" } },
      { "not": { "pattern": "(\\s\\s+)" } },
      { "not": { "pattern": "(\\t)" } },
      { "not": { "pattern": "([<>])" } }
    ]
  }
}
```

[그림 55] 원시 말뭉치의 스키마 예시

(4) 발화 요소 오류 탐지

원시 말뭉치 중 일상대화, 구어/준구어, 메신저 대화는 발화 요소 <u>를 가진다. 발화 요소는 <unclear>와 같은 전사 마크업, <anon>과 같은 익명화 마크업 등 특수 마크업 포함하며, 특히 전사 마크업의 경우 전사자들의 수작업이 개입되므로 오류가 발생할 가능성이 크다. 또한, 오류는 전사 파일을 원시 말뭉치로 변환하는 과정에서도 생길 수 있다. 본 사업단에서는 모든 경우가 오류인 것은 아니지만 오류가 발생할 가능성을 가지는 특정 조건들을 탐색하였다. 발화 요소의 오류에 대해 탐색한 내용 중 몇 가지를 제시하면 다음과 같다.

○ 마크업 기호 안쪽으로 바로 공백이 있는 경우 (불필요한 공백 문자)

- 자 질문 궁금하신 점 한 가지 질문해<unclear>주십시오. </unclear>

○ 마크업 기호 바깥 양쪽으로 어절이 바로 인접한 경우 (어절 미분리 등 전사 오류)

- 저는 애니<anon type="name" n="2"/>션을 좋아했어요.

○ 첫 번째 발화 <u n="1">의 who 속성이 P1이 아닌 경우 (전사 파일 -> 원시 말뭉치 변환 오류)

- <u who="P5" n="1">박용우,남자,50대,전문의</u>

(5) 검증 결과 로그 생성

검증 결과는 스키마별로 구분하여 스프레드시트 형식으로 묶어 생성하였다. 첫 번째 시트 ‘process’에는 파일명과 함께 해당 파일의 인코딩 검사, XML 형식 검사, 데이터 유효성 검사의 결과가 PASS/FAIL로 기록된다. 두 번째 시트 ‘pathMap’에는 검사를 수행한 파일들의 경로 정보가 들어 있다. 세 번째 시트 ‘parseLog’에는 XML 형식 검사에서 발생한 로그가 기록되고 네 번째 시트 ‘validLog’에는 데이터 유효성 검사에서 검출된 스키마 위반 사항이 기록된다. 그 이후의 시트들은 ‘ex1’ ~ ‘exN’으로 차례로 연번이 매겨지며 말뭉치 유형에 따라 부가적으로 수행된 검사들의 로그가 기록된다. 결과 파일에서 가장 핵심적인 정보인 ‘process’ (<표 78>), ‘parseLog’ (<표 79>), ‘validLog’ (<표 80>) 시트의 모습은 아래와 같다.

process			
fileName	encoding	parsing	validating
SERW1900000001.sjml	pass	fail	fail
SERW1900000002.sjml	pass	pass	fail

<표 78> process 시트

parseLog			
fileName	line	type	log
SERW1900000001.sjml	5283	InvalidTag	Closing tag 'seaker' is expected in place of 'speaker'.
SERW1900000003.sjml	42013	InvalidTag	Tag '</stage' is an invalid name.

<표 79> parseLog 시트

validLog		
fileName	location	log
SERW1900000002.sjml	/SJML/text/episode/15/scene/42/sp/1/p	should NOT match pattern “(\\s)\$“
SERW1900000004.sjml	/SJML/header/sourceInfo/author	should be string

<표 80> validLog 시트

스프레드시트로 로그 리포트를 생성하면 하나의 파일에 결과를 모아 볼 수 있다는 장점이 있으나 로그의 양이 많아지면 결과 파일 생성 시 예외가 발생할 수 있다는 단점이 있다. 따라서 스프레드시트 리포트에는 말뭉치 파일당 하나의 로그만 기록하였다. 예를 들어 어떤 말뭉치 파일의 스키마 위반 사항이 여러 개 있더라도 ‘validLog’ 시트에는 가장 먼저 검출된 하나의 로그만 기록된다. 검사를 수행한 말뭉치들의 전체 위배 내용은 별도의 텍스트 파일을 통해 출력하였다.

원시 말뭉치에 대한 검증 작업 역시 파일 단위로 수행된다. 실제 검증 작업은 스키마 유형별로 파일을 묶어 진행하였으나 결과는 6종의 말뭉치 유형에 따라 제시하고자 한다. 원문 검증과 마찬가지로 하위 검사 항목의 성공 여부에 따라 PPP, PP, P, F 네 가지 유형으로 분류하고 각 유형별 수량과 비율을 제시한다.

	F	P	PP	PPP	전체
문어	0 (0%)	0 (0%)	10 (0.05%)	20,042 (99.95%)	20,052 (100%)
신문	0 (0%)	0 (0%)	19 (4.70%)	386 (95.30%)	405 (100%)
일상대화	0 (0%)	13 (0.65%)	1,480 (73.89%)	510 (25.46%)	2,003 (100%)
구어/준구어	0 (0%)	794 (3.99%)	8,206 (41.08%)	10,973 (54.94%)	19,974 (100%)
웹	0 (0%)	2,169 (71.57%)	862 (28.43%)	0 (0%)	3,032 (100%)
메신저 대화	0 (0%)	565 (7.64%)	2,998 (40.54%)	3,832 (51.82%)	7,395 (100%)

<표 81> 원시 말뭉치 유형별 검증 결과: 파일 수 (단위: 개)

인코딩 검사에서는 6종 원시 말뭉치의 모든 파일이 검사를 통과하였다. 검출 단계에서는 구어/준구어에서 1건 웹 말뭉치에서 1건 총 2개 파일이 인코딩 검사를 통과하지 못했지만 실제 파일을 열고 문자 세트를 확인했을 때 문제는 없었다. 인코딩 검사는 추론 모델을 사용하므로 늘 완벽한 결과를 내어주는 것은 아니다.

XML 형식 검사 결과를 보면 문어와 신문 말뭉치에는 문제가 있는 파일이 없었다. 신문의 경우 이미 원문 자료에서도 오류가 발견되지 않았었고, 문어의 경우 원시 말뭉치로 변환하는 과정에서 오류가 전부 해결되었다. 일상대화, 구어/준구어, 메신저대화는 형식 검사를 통과하는 비율이 그렇지 않은 경우보다 높았지만 형식 수정 작업이 일부 필요해 보

인다. 웹 말뭉치는 형식 검사를 통과하지 못한 파일이 더 많았다.

문어와 신문 말뭉치는 유효성 검사에서도 긍정적인 결과를 얻었다. 일부 문제가 발견되기는 했지만 문어는 99.95%, 신문은 95.30%로 높은 통과율을 보였다. 나머지 말뭉치들은 대체로 결과가 좋지 못하며 위반 사항에 대한 대대적인 수정이 필요해 보인다.

XML 형식 검사에서 발견된 오류는 원문 자료 검증에서와 마찬가지로 PCDATA 영역에 ‘&’, ‘<’, ‘>’ 등의 특수 문자가 출현해서 발생하는 문제가 가장 많았다. 특히 웹 말뭉치의 경우 <url> 요소를 가지고 있는데, url에서는 ‘&’ 문자가 자주 출현하기 마련이다. 또한 블로그, 게시글, 리뷰 등에서 나타나는 ‘>_<’와 같은 이모티콘의 처리가 필요해 보인다.

일상대화, 구어(공적대화, 공적독백)와 같이 발화 요소 <u>를 가지는 말뭉치들에서는 내부 마크업 형식에 문제가 있는 경우가 많았다. 자주 출현하는 오류는 다음과 같다.

○ 여는 태그와 닫는 태그가 교대로 중첩되는 경우

- 제가 <unclear><trunc>잘</unclear></trunc>

○ 태그 형식이 잘못된 경우

- <note></not>

유효성 검사에서는 공백 문자 처리와 관련된 문제가 가장 많았다. 공백 문자가 불필요하게 연속되거나, 문자열 값의 시작이나 끝 부분에 공백 문자가 있거나, 탭문자나 개행 문자가 포함된 경우이다.

통과율이 매우 떨어지는 일상대화 말뭉치와 웹 말뭉치에서는 일관된 오류가 발견되었다. 일상대화 말뭉치에서는 <vocal> 태그가 <voice> 태그로 일괄적으로 잘못 달려 있었다. 웹 말뭉치에서는 모든 파일의 <fileId>가 잘못되어 있었는데, 연도 코드 ‘19’와 8자리 일련번호 사이에 ‘W’가 삽입되어 있었다.

○ 일상대화: 정의되지 않은 태그 <voice>를 사용함

- <voice desc=“목청가다듬는소리“/>

○ 웹: 아이디 부여방식에 맞지 않는 <fileId>

- <fileId>EBRW19W00000208</fileId>



제 4 장

결 론 및 제 언

1. 결론

국립국어원의 말뭉치 구축 사업은 4차 산업혁명에 대비한 우리말 빅데이터(말뭉치)를 구축하는 사업이다. 이는 향후 4차 산업혁명 시대의 인공지능 서비스 개발 및 기술 혁신을 위한 중요 자료로 활용하기 위한 기반을 다지는 사업으로서의 의미가 매우 큰 사업이라고 할 수 있다. 국어 빅데이터(말뭉치) 구축 사업의 일환으로 추진된 본 사업은 다양한 분야의 책, 잡지, 보고서 등 문어 자료를 모아 말뭉치로 구축하여 국어 인공지능 개발 산업과 국어 연구 등에서 공공 자료로 자유롭게 활용할 수 있도록 하는 데 그 목적이 있다.

7개 층위 분석 말뭉치 검증의 경우, 검증용 말뭉치와 검증 대상 분석 말뭉치를 비교하여 검증하기 위해서 검증용 말뭉치의 품질 보증을 위해서 많은 노력을 기울였다. 한국어의 교착어 특성상 하나의 어절이 여러 의미로 해석될 수 있는 가능성이 있기 때문에, 문맥 정보를 고려한 신중한 주석이 필요했다. 주석자들이 주석 지침을 준수하여 신중하게 주석하였다 하더라도 의미적 중의성과 다양성, 모호성으로 인한 여러 가지 주석의 가능성이 존재하였다.

이러한 가능성을 줄이기 위해 검증용 말뭉치 구축 과정에서 주석자 간 일치도를 고려하여 주석 과정을 설계하였고, 주석 내용에 대한 교차 검증을 수행하여 검증용 말뭉치의 주석 품질을 높이려고 하였다. 또한, 본 사업단에서는 주석 체계가 완전하지 않은 상태에서, 주석 내용의 결함이 있을 수 있다는 것을 감안하고 그러한 결함을 주석 단계에서 보완하는 방식을 택하여 짧은 시간 내에 다층위의 말뭉치를 구축하는 데에 적합한 방식을 취했다고 평가한다.

검증용 말뭉치를 구축하기 위해서 취한 병렬적인 구축 방법은 기존의 직렬적인 구축 방식과는 차이가 있었다. 직렬 구조의 다층위 말뭉치 구축은 하나의 완성된 분석 말뭉치 위에 새로운 내용을 주석하여 앞 단계의 주석 속도와 결과에 영향을 받을 수밖에 없다. 그러나 본 사업단에서는 병렬적으로 여러 층위의 분석 말뭉치를 구축한 후에 하나로 합치는 방법을 취하였다. 이는 분석 대상의 원시 말뭉치가 동일한 경우에 빠른 시간 내에 다양한 층위의 분석 말뭉치를 구축하는 효율적인 방법이었다고 평가한다.

비교적 단기간에 검증용 말뭉치를 구축하고, 이를 활용하여 검증 대상 분석 말뭉치를 검증하기 위해서 검증용 말뭉치의 품질과 권위가 중요하다. 그러나 환경적인 제약으로 인해 이를 완전히 달성하지 못할 수 있다. 그렇기 때문에 두 말뭉치의 비교 결과에서 이견이 발생하였을 경우, 전체 사업의 목적에 따라 두 말뭉치의 주석 정답을 유연하게 결정하는 것이 필요했다. 국립국어원에서는 이러한 분석 단위에 대해서 예시와 해석을 제공하였다. 국립국어원의 해석을 토대로 검증용 말뭉치의 내용을 수정하고 보완함으로써 검증용 말뭉치의 품질을 보완하였다.

그럼에도 불구하고 검증용 말뭉치의 품질은 완전하지 않을 수 있다. 주석 오류를 포함한 검증용 말뭉치로 검증 대상 분석 말뭉치를 평가한다는 것은 검증의 의미를 퇴색시킬 수도 있다. 그러나 검증용 말뭉치와 검증 대상 분석 말뭉치의 주석 불일치 내용을 보고함으로써, 두 주석 집단이 다시 한 번 해당 주석 단위에 대해서 숙고하고, 수정 및 보완할 수 있는 기회를 제공하였다는 데에서 검증의 기능을 하였다고 평가한다.

구축한 검증용 말뭉치를 학습한 모델을 사용한 주석 일관성을 검증한 것은 원시 말뭉치의 구성이 균형적이므로 언어 현상을 대표한다는 가정 하에 제안된 방법이다. 문어 원시 말뭉치의 경우, 신문 기사 장르로만 이루어져 있기 때문에 세 장르(공적 독백, 공적 대화, 사적 대화)로 이루어진 구어 원시 말뭉치에 비해 언어 현상을 고르게 대표한다고 평가할 수 있었다.

균형적인 구성의 원시 말뭉치에 자동 주석을 수행하여 만든 자동 주석 말뭉치의 품질은 주석 일관성 검증 결과에 큰 영향을 주지 않았다고 할 수 있다. 이는 주석 일관성 검증의 목표가 검증 대상 분석 말뭉치 전반에 걸쳐 검증용 말뭉치가 가지고 있는 주석 특성을 확인하여 주석 일관성을 간접적으로 평가하는 과정이다. 따라서 검증용 말뭉치가 제대로 학습된 모델이 생성되었다면, 자동 주석 말뭉치와 검증 대상 분석 말뭉치의 일치도는 주석 일관성 검증에 큰 문제가 되지 않는다고 평가한다. 그렇지만 주석 일관성 검증의 기준을 99% 신뢰구간과, 10개의 분할로 한 것은 비교적 관대한 검증 기준을 적용했다고 할 수 있다.

원문 자료 수집 자료와 이를 활용하여 구축한 원시 말뭉치, 그리고 새로운 종류의 구축된 원시 말뭉치를 검증함으로써 향후 이를 활용한 분석 말뭉치 구축 및 활용 시에 발생할 수 있는 오류를 사전에 예방하였다.

종합적으로, 본 사업의 말뭉치 검증 과정을 통하여 분석 말뭉치 및 원시 말뭉치의 품질을 평가함으로써 양질의 공공자원을 구축할 수 있었다는 데에 본 사업의 의의를 둘 수 있다.

2. 제언

본 사업은 대규모 언어 자원의 품질을 검증하고 일원화된 검증 체계를 확보함으로써 대규모 언어 자원의 구축 과정을 체계적으로 지속적으로 관리, 정제하는 작업이었다. 이러한 작업이 국가 차원에서 대규모로 진행되었다는 것은 그동안의 국어학계와 전산학계가 쌓아온 연구 결과가 실용적 성과를 내고 있음을 보여주는 사업이라고 할 수 있다. 하지만 사업 진행 과정에서 드러난 몇 가지 아쉬운 점을 지적하면 다음과 같다. 본 사업의 경험을 교훈으로 삼아 추후 말뭉치 구축 사업이 성공적이길 바란다.

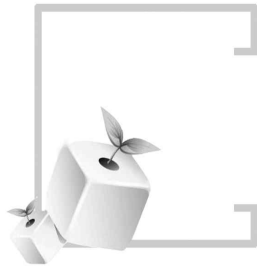
첫째, 주식 지침은 계약서에 준하는 효력을 지닌 것으로, 반드시 국립국어원과 검증사업단, 그리고 구축사업단이 모두 공유할 수 있는 공식적인 경로를 통해 세 주체가 모두 확인한 지침과 개정판만이 유효한 지침으로 적용되어야 한다. 본 사업의 경우 층위별 분석 말뭉치 구축 사업과 동시에 진행되는 과정에서 주식 지침을 완비하지 못하고, 사업 진행 과정에서 몇 차례 개정이 이루어졌다. 이로 인해 층위별 구축사업단에서 지침을 변경한 경우에 본 사업단이 즉각 반영해야 하는 데에 어려움이 있었다. 향후 사업에서는 가능한 한 사전에 주식 지침을 완비하고, 부득이하게 지침이 변경되는 경우에는 각 작업 주체가 동시에 공유하는 공식적인 경로로 전달되어 유효한 변경 사항에 대해서만 작업에 반영될 수 있기를 바란다. 주식 지침을 포함한 주식 체계가 확립될 때 어느 누가 주식 작업을 하더라도 동일한 주식 결과를 낼 수 있는 작업 환경이 될 것이다.

둘째, 검증용 말뭉치 주식자들의 분석 말뭉치 구축 경험을 쌓을 수 있는 충분한 여건이 조성되지 못한 점도 아쉬운 점이다. 본 사업에 주식자들은 국어학 혹은 언어학을 전공한 전문 인력으로 주식의 성격을 파악하고 이해하는 데에 어려움이 없었다. 그러나 실제로 주식하면서 예상하지 못했던 언어 현상들에 대한 분석을 주식 작업 중에 해결해 나가는 것들은 전문 인력들이라 하더라도 쉬운 일이 아니었다. 그뿐만 아니라, 본 사업은 참조할 만한 과거 사례가 없었으므로, 주식 작업을 설계하고 그 결과물을 활용하여 검증을 수행해야 하는 본 사업단 입장에서도 사업 수행 환경에서 발생한 예상치 못한 문제들을 보완하는 데 많은 시간과 노력이 필요했다. 본 사업을 기반으로 삼아 향후 과제를 수행할 때에는 사업 동력을 보다 과제 자체에 집중하게 될 수 있게 되기를 바란다.

넷째, 향후 국어 빅데이터 구축 사업의 각 과제들이 분야에 따라 성취 정도 및 수준이 다를 수 있음을 고려하여, 과제의 성격에 맞게 사업을 세분화하고, 연 단위로 성취해야 할 과업의 목표를 명확하게 하였으면 한다. 주어진 기간 내에 수행 가능한 과업의 목표를 설정하여야 양질의 연구 결과물을 얻을 수 있을 것이다.

향후의 국어 빅데이터 구축 사업이 더욱 성공적으로 진행되기 위해서는 본 사업을 비롯한 2019년 국립국어원의 국어 빅데이터 구축 사업을 반면교사로 삼아 사업 계획 단계

에서부터 치밀하고 세부적인 준비가 필요할 것이다. 사업의 시행착오도 사업 성공을 위한 귀중한 경험임을 너그러이 받아들일 수 있을 때 학문과 연구의 성과가 더욱 적극적으로 실용적으로 활용될 수 있을 것이다.



부록

<부록 1> 층위별 초별 주석 말뭉치 JSON 구조

층위	구분	제공 정보 (keys)		설명
형태 분석	공통	text		문단 단위의 원문 텍스트
		document_id		원시 말뭉치 문서 ID
		paragraph_id		같은 문서 내 문단 ID
		morps ³⁶⁾	par_lemma_id	문단 내 형태소 ID
			par_word_id	문단 내 어절 ID
			text	형태소 텍스트
			tag	형태소 태그
	구어	part_id		구어 원시 말뭉치 파일을 10개 발화 단위로 나누어 만든 부분 파일의 ID
		morps ³⁷⁾	speaker	구어 화자 정보
			part_lemma_id	부분 파일 내 형태소 ID
			part_word_id	부분 파일 내 어절 ID
어휘 의미 분석	공통	text		문단 단위의 원문 텍스트
		document_id		원시 말뭉치 문서 ID
		paragraph_id		같은 문서 내 문단 ID
		wsds ³⁸⁾	par_lemma_id	문단 내 형태소 ID
			par_word_id	문단 내 어절 ID
			text	형태소 텍스트
			tag	형태소 태그
			is_target	어휘 의미 주석 대상 표시
	구어	part_id		구어 원시 말뭉치 파일을 10개 발화 단위로 나누어 만든 부분 파일의 ID
		wsds ³⁹⁾	speaker	구어 화자 정보
			part_lemma_id	부분 파일 내 형태소 ID
			part_word_id	부분 파일 내 어절 ID
개체 명 분석	공통	text		문단 단위의 원문 텍스트
		document_id		원시 말뭉치 문서 ID
		paragraph_id		같은 문서 내 문단 ID
		ners	par_word_id	문단 내 어절 ID
			text	어절 텍스트
			morps ⁴⁰⁾	어절 내 형태소 정보 ⁴¹⁾
	구어	ners	morps ⁴²⁾	어절 내 형태소 정보 ⁴³⁾
주격 무형 대용어 복원	공통	document_id		원시 말뭉치 문서 ID
		za	paragraph_id	같은 문서 내 문단 ID
			sentence_id	문단 내 문장 ID
			text	문장 텍스트
			word_id	문장 내 어절 ID
			text	어절 텍스트
			tag	구문 분석 태그
			head	지배소
			is_target	복원 대상 서술어 표시
	구어	za	speaker	구어 화자 정보
상호	공통	document_id		원시 말뭉치 문서 ID

참조 해결		corefs	paragraph_id		같은 문서 내 문단 ID
			sentence_id		문단 내 문장 ID
			text		문장 텍스트
			NEs ⁴⁵⁾	lemma_id	문장 내 형태소 ID
				word_id	문장 내 어절 ID
				text	형태소 텍스트
				NE	개체 태그
				NE_id	개체 ID
	구어	corefs	speaker		구어 화자 정보
구문 분석	문어	deps	par_word_id		문단 내 어절 ID
			text		어절 텍스트
			tag		구문 분석 태그
			morps		어절 내 형태소 정보 ⁴⁶⁾
의미 역 분석	문어	srls	document_id		원시 말뭉치 문서 ID
			paragraph_id		같은 문서 내 문단 ID
			sentence_id		문단 내 문장 ID
			text		문장 텍스트
			deps ⁴⁷⁾	word_id	문장 내 어절 ID
				text	어절 텍스트
				tag	구문 분석 태그
				head	지배소
				morps	어절 내 형태소 정보 ⁴⁸⁾

36) 엑소브레인 언어분석 툴킷 V3.0 형태 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

37) 엑소브레인 언어분석 툴킷 V3.0 형태 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

38) 본 사업단에서 구축한 형태 분석 검증용 말뭉치 결과를 활용하여 초별 주석 말뭉치를 구축함

39) 본 사업단에서 구축한 형태 분석 검증용 말뭉치 결과를 활용하여 초별 주석 말뭉치를 구축함

40) 엑소브레인 언어분석 툴킷 V3.0 형태 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

41) 형태 분석 초별 주석 말뭉치에 포함된 공통 morps key와 같음

42) 엑소브레인 언어분석 툴킷 V3.0 형태 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

43) 형태 분석 초별 주석 말뭉치에 포함된 구어 morps key와 같음

44) 엑소브레인 언어분석 툴킷 V3.0 구문 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

45) 엑소브레인 언어분석 툴킷 V3.0 개체명 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

46) 형태 분석 초별 주석 말뭉치에 포함된 공통 morps key와 같음

47) 엑소브레인 언어분석 툴킷 V3.0 구문 분석 결과를 활용하여 초별 주석 말뭉치를 구축함

48) 형태 분석 초별 주석 말뭉치에 포함된 공통 morps key와 같음

〈부록 2〉 주석 워크벤치 입력 및 출력 (JSON)

층위	입력 ⁴⁹⁾	출력 ⁵⁰⁾	추가 정보 (key) ⁵¹⁾		설명
형태 분석	형태 초별 주석 말뭉치	주석자에 의해 수정된 형태 분석 말뭉치	isHold		작업 보류 여부
			problem_yn		문단 단위 작업 불가 여부
			problem_reason		문단 단위 작업 불가 이유
어휘 의미 분석	어휘 의미 분석 초별 말뭉치	주석자에 의해 주석된 어휘 의미 말뭉치	is_problem		어절 단위의 작업 불가 여부
			cw_comment		어절 단위 주석자 메모
			problem_yn		문단 단위 작업 불가 여부
			problem_reason		문단 단위 작업 불가 이유
개체명 분석	개체명 초별 분석 말뭉치	주석자에 의해 주석된 개체명 말뭉치	data	lemma_id	개체가 포함된 형태소 ID
				text	형태소 텍스트
				word_id	어절 ID
			entities	startPosition	개체 시작 범위
				endPosition	개체 끝 범위
				text	개체 텍스트
				type	개체명 태그
			problem_yn		문단 단위 작업 불가 여부
			problem_reason		문단 단위 작업 불가 이유
주격 무형 대용어 복원	주격 무형 대용어 복원 초별 말뭉치	주석자에 의해 주석된 주격 무형 대용어 복원 말뭉치	is_anaphora		문장 내 주어 복원 대상 서술어 포함 여부
			dependencies	text	복원 주어
				p_idx	복원 주어가 포함된 문단의 인덱스
				word_id	복원 주어의 어절 ID
				sentence_id	복원 주어가 포함된 문장 ID
				paragraph_id	복원 주어가 포함된 문단 ID
				real_target	실제 주어 복원 여부 ⁵²⁾
			problem_yn		문서 단위 작업 불가 여부
			problem_reason		문서 단위 작업 불가 이유
상호참조 해결	상호참조 해결 초별 말뭉치	주석자에 의해 주석된 상호참조 해결 말뭉치	groups	group_id	상호참조 군집 ID
				lemma_ids	공지시 관계 개체 형태소 ID 범위
				paragraph_ids	공지시 관계 개체 문단 위치 ID
				sentence_id	공지시 관계 문장 ID
				word_ids	공지시 관계 어절 ID
				text	공지시 관계 개체 (대표 개체)
			item	group_id	상호참조 군집 ID
				lemma_ids	공지시 관계 개체 형태소 ID 범위

				paragraph_ids		공지시 관계 개체 문단 위치 ID	
				sentence_id		공지시 관계 문장 ID	
				word_ids		공지시 관계 어절 ID	
				text		공지시 관계 개체	
			problem_yn		문서 단위 작업 불가 여부		
			problem_reason		문서 단위 작업 불가 이유		
구문 분석	구문 분석 초별 말뭉치	주석자에 의해 주석된 구문 분석 말뭉치	deps	head		지배소	
				function_tag		기능 태그	
			problem_yn		문단 단위 작업 불가 여부		
			problem_reason		문단 단위 작업 불가 이유		
의미역 분석	의미역 분석 초별 말뭉치	주석자에 의해 주석된 의미역 분석 말뭉치	srls	srl_ par sed	1수 준	리스트	문장 내 의미역 주석 결과 ⁵³⁾
					2수 준	리스트1	문장 내 어절
						리스트2	서술어 주석 결과
						리스트3	논항 주석 결과
			problem_yn		문서 단위 작업 불가 여부		
			problem_reason		문서 단위 작업 불가 이유		

49) <표 >의 층위별 초별 주석 말뭉치 JSON 구조 참조

50) 초별 주석 말뭉치가 주석자에 의해서 수정된 JSON 형식의 말뭉치

51) <표 >의 층위별 초별 주석 말뭉치 JSON 구조 기준 추가된 JSON key

52) <표 >의 주격 무형 대용어 복원 is_target 대비 실제 주어 복원 여부

53) 문장 내 의미역 주석 결과 집합

〈부록 3〉 층위별 분석 말뭉치 구축 지침 검토 의견

① 형태 분석 말뭉치 구축 지침 검토 의견

지침 위치	지침 내용	검토 의견
V8 - 38쪽	③ ‘-조’는 축약형을 그대로 태깅한다. [예시] 어서 출근하죠. [출근/NNG+하/XSV+조/EF+./SF]	- 예외 규정 추가 의견: 일부 종결어미 ‘조’는 축약형 종결어미로 볼 수 없으므로 분석하여 처리한다. (예: ‘해죠, 봐죠’ [‘주/VX+오/EF’])
V8 - 39쪽	연결어미(EC): 본 지침에서는 우리말샘에 따라 연결어미를 구분하는 것을 원칙으로 한다.	- 축약어 규정 추가 의견: 연결어미 축약형의 경우 지침의 축약어 규정을 적용하여 전체를 한꺼번에 연결어미(EC)로 처리한다. (예: ‘-다는데(EC), -라는데(EC), -자는데(EC)’)
V8 - 47쪽	다) 동사파생접미사(XSV) 목록 표	- 목록 추가 의견: 동사파생접미사 목록에 ‘드리’ 추가 (예: 말씀드리다) - 근거 : <우리말샘> 드리다 품사「접사」 「011」((몇몇 명사 뒤에 붙어)) ‘공손한 행위’의 뜻을 더하고 동사를 만드는 접미사임 (예: 공양드리다, 불공드리다, 말씀드리다.)

② 어휘의미 분석 말뭉치 구축 지침 검토 의견

지침 위치	지침 내용	검토 의견
3쪽	붙임. 접사처리(통합) 과정에서 제외된 접사 목록 중 생1(生) 예) 갑자생	- ‘연년생(年年生)’에서 ‘-생’의 경우 지침에서 언급한 ‘-생1(生)’에 해당되는지에 대한 판단이 모호함 - ‘연년생’의 ‘-생’은 ‘갑자생’의 ‘-생’과 다소 차이가 있는 것으로 볼 수 있으므로, 붙임으로 <‘연년생’은 접사 처리(통합)에서 제외함>이라는 항목을 추가할 필요가 있음
3쪽	다. <우리말샘>에 같은 한글 배열로 이루어진 형태가 등재되지 않은 어휘의 경우, 어휘의미 번호는	- 구어 말뭉치 자료에서 특정 인명이 텍스트에 나타날 때, 개인 정보 보호를 위해 인명을 ‘name1’과 같은 형식으로 전사된 때도 있으나, 이에 대한 어휘의미 주석 지

	'777'로 한다.	<p>침이 없음</p> <ul style="list-style-type: none"> - 이에 대한 지침 마련이 필요하며, 검증용 말뭉치 주석 작업 시에는 이러한 항목을 777로 처리함
5쪽	<p>마. 말뭉치 원어절의 오타, 탈자 등의 오류로 의미 분석이 불가능한 경우, 어휘의미 번호는 '999'로 한다.</p> <p>예) <u>고깃</u>을 부리다.</p> <p>단, <우리말샘>에 등재된 표준어에 대한 일반적인 비표준어로 판단되는 경우, 향후 등재 가능성을 고려하여 '777'을 부여한다.</p>	<ul style="list-style-type: none"> - 문어 말뭉치 자료에서는 이 지침의 적용에 큰 문제가 없으나, 구어 말뭉치 자료에서는 이 지침의 적용에 문제가 있을 수 있음 - 원 구어 발화를 전사한 자료인 구어 말뭉치 자료의 특성상 '999'에 해당하는 것으로 볼 여지가 있는 대상은 “원 구어 자료 발화자의 발화 실수에 해당하는 것”과 “전사자의 전사 오류에 해당하는 것”이 있을 수 있는데, 이를 구별하는 방안과 그에 따라 주석하는 방안을 검토하여 마련할 필요가 있음 - 예) ‘<u>이병</u>을 연기한다고 해’ → 이 예에서 ‘이병’은 문맥상 ‘입영’에 해당한다. 이는 전사자가 전사하는 과정에서 발생한 오류일 가능성이 큼 - 예) “그리고 엄마도 name4이 군대 간다고 그 <u>영상</u>을 받고 힘들어하는데, 내가 거기다 대고 강아지 그렇다고 할 수가 없어서” → 이 예에서 ‘영상’은 문맥상 ‘영장’으로 추측되는데, 이는 발화자가 발화 실수를 한 것인지, 전사자가 오류를 범한 것인지를 판단하기가 쉽지 않으므로 이와 같은 예들에 대한 보다 구체적인 처리 방안을 지침으로 마련할 필요가 있음

③ 개체명 분석 말뭉치 구축 지침(Ver. 2.5.) 검토 의견

지침 위치	지침 내용	검토 의견
19쪽	PS_NAME	- ‘관음보살’, ‘선재동자’의 경우 PS로 주석하지만, 문화재인 경우 AF로 태깅함. 구체적인 태깅 기준이 필요함.
24쪽	FD_ART	- ‘전통무용’과 같이 ‘전통OO’의 경우 장르로

	예술 관련 학문 분야 외에도 예술의 분류, 장르가 태깅 대상에 포함된다...예술의 '장르'로 나왔을 때 한정하며, 예술 활동의 '결과물(작품)'은 태깅하지 않는다.	판단하는지 여부. '모던발레', '민속예술'과 같이 세부 장르에 대한 지침이 마련되어야 함. - '케이팝', '애니메이션'과 같이 장르이자 결과물로 해석되는 경우 태깅 기준이 모호함.
53쪽	OGG_SPORTS 스포츠 기관/단체 특정한 스포츠 유형이 들어가야 개체명 대상이 됨. - '잉글랜드 3부리그', '1부(리그)와 같이 구체화한 경우에만 OGG_SPORTS로 할당 예) K리그, 메이저리그, 프리미어리그	- U-17 대표팀이 유럽 강호들을 상대로 대패했다. - 남자 국가대표팀과 쿤룬 레드스타의 2연전으로... 예시와 같이 특정 스포츠가 직접 나타나지 않더라도 축구나 아이스하키 대표팀을 지칭하는 것을 문맥에서 알 수 있는 경우 태깅에 대한 지침이 마련되어야 함. - 1군, 2군의 경우 1군/QT, 2군/QT로 태깅하나 1부, 2부, 3부 리그는 OG로 태깅함. 1군이 공식 경기에 참여하는 그룹을 나타내므로 일관된 지침을 제공할 필요가 있음.
79쪽	CV_FOOD 음식/곡물 명칭, 음식 재료 음식 유형(국, 밥, 무침, 향신료, 안주) 포함	- 음식 유형과 음식 재료에 대한 기준이 모호함. 가) '국물', '육수', '소스', '소금', '후추', '고춧가루' 등 - 예시 (가)와 같이 일반적인 용어에는 태깅하지 않고 예시 (나)와 같이 특정성, 유일성을 지니는 음식 재료나 음식 명칭에만 태깅하는 것이 바람직함. 나) 서리태콩, 취나물, 케첩, 마요네즈 등 - 음식명에 대한 태깅 여부 고려해야 함. 다) 전주비빔밥, 평양냉면, 짜파구리 등
81쪽	CV_CLOTHING '바지'는 태깅하지 않는다.	- '바지'가 상위 층위에 해당하여 주석하지 않는 것은 이해되나 계열어 및 상하위어의 체계성을 고려한 태깅 기준이 필요함. 의복의 상위 유형에는 태깅하지 않고 의복 종류나 특정 옷에만 태깅하는 것이 바람직

		<p>함.</p> <p>예) ‘치마’, ‘셔츠’, ‘스웨터’, ‘점퍼’, ‘재킷’, ‘코트’, ‘양말’, ‘레깅스’ (X)</p> <p>‘청바지’, ‘코듀로이 바지’, ‘면바지’, ‘주름치마’ (O)</p>
81-83쪽	<p>CV_POSITION</p> <p>직위/직책 명칭, 스포츠 포지션, 사람이 가질 수 있는 특정한 역할의 경우</p> <p>CV_OCCUPATION</p> <p>직업 명칭</p>	<p>- ‘시청자’ CV로 볼 수 있음. 특정한 역할이라는 기술이 부적절하며 CV_POSITION, CV_OCCUPATION의 구체적인 기준이 필요함.</p> <p>예) 위안부, 장애인, 비정규직, 취업자, 피해자, 태극전사, 최고위원단, 상임대표단, 보좌진, 친박계 등</p> <p>- CV 태깅 범위에 대한 구체적인 예시가 추가되어야 함.</p> <p>예) 담임+[선생님]CV, [7급]CV+[공무원]CV, [고위공무원]CV, [부장]CV+[판사]CV, 대학+[교수]CV, [바둑]+[9단], [교통경찰]CV, [신문기자]CV, [의학전문기자]CV, [인디뮤지션]CV, [가정주부]CV 등</p>
99쪽	<p>QT_COUNT</p> <p>개수, 빈도</p> <p>‘수관형사+단위명사, 수관형사 + 일반명사’ 모두 통합하여 태깅하는 것으로 수정. 일반명사가 비개체명일 경우도 통합하여 태깅한다.</p> <p>두 마리/QT, 두 여자/QT</p>	<p>- 의미상으로 수량에 해당하지 않을 때 대해 고려 필요함. 예외적일 때 설명이 추가 보완되어야 함.</p> <p>예) 두 눈에서 불이 번쩍 나다.</p> <p>- 한자어 접사나 한자어 수관형사가 붙으면 수량 태깅 범위에 포함할 것인지 고려 필요함.</p> <p>예) 양팔, 양손, 쌍수, 양 갈래</p>
105쪽	<p>QT_CHANNEL</p> <p>TV/라디오 채널 번호</p> <p>명확한 채널 번호에 해당하는 경우에만 태깅한다.</p> <p>CJ홈쇼핑/AF, 코미디 TV/AF,</p> <p>MBCsports+/AF</p> <p>채널명이 아닌 방송 기관</p>	<p>- 채널의 다양성을 고려하여 구체적인 예를 추가할 필요가 있음.</p> <p>예) 아프리카TV, 팟캐스트, 유튜브 등</p>

	으로 판단될 경우 OG로 태깅할 수 있다. EBS/OG	
128쪽	TM_DISEASE 증상/증세/질병	<ul style="list-style-type: none"> - 예제에 ‘기침’을 포함하고 있는데 ‘감기’, ‘고열’, ‘코막힘’, ‘체함’, ‘위경련’, ‘설사’, ‘토’ 등의 증세에 대해서도 개체명 태깅 대상이 되는지 모호함. - ‘스트레스’, ‘성장 장애’, ‘알콜 중독’, ‘우울증’, ‘불안증’, ‘치매’ 등의 증세에 대해서도 개체명 태깅 대상이 되는지 기준이 필요함.
129쪽	TM_HW 하드웨어 용어, 전자기기 에 해당하는 제품 예) 컴퓨터, 팩스, 핸드폰, 전화(기), 네비게이션 등	<ul style="list-style-type: none"> - 문맥에 따라 ‘전화하다’, ‘전화를 걸다’, ‘팩스를 보내다’, ‘컴퓨터를 하다’와 같이 기기를 지칭하기보다는 행위의 의미를 지니는 경우, 일반적인 의미로 사용될 경우에도 태깅하는지 여부에 대한 설명이 추가되어야 함.
134쪽	TM_SPORTS 스포츠/레저 용어(기술, 규칙 이름 등) 방향성 정 보+땅볼/안타/파울/홈런 등 묶어서 TM으로 할당 함 1,2루, 1볼넷 TM으로 태 깅, 6이닝 QT로 태깅	<ul style="list-style-type: none"> - 스포츠 규칙, 용어에 대해 종목별로 구체 적인 예시 추가 필요함. <p>1볼넷/TM으로 되어 있으나 ‘2볼넷’, ‘볼넷 2개’는 ‘30도루’와 같이 QT로 해석될 수 있음. 골프의 경우, ‘파3/QT’, ‘6오버파/QT’, ‘오버파/TM’로 적용됨. ‘3안타 2볼넷’은 ‘3안타/QT + 2볼넷/QT’로 태깅할지 통합하여 태깅할지 여부도 논의가 필요함.</p>

④ 주격 무형대용어 복원 말뭉치 구축 지침(Ver. 2.5.) 검토 의견

지침 위치	지침 내용	검토 의견
‘3. 주격무 형대용어복 원 말뭉치_ 구축 지침 (2019)(1912 02).hwp’ 의 1쪽	주어 복원 대상이 아닌 술어 배제 ㄱ. 보조 용언 ㄴ. 의사 보조 용언 구 성	<p><지침의 수정></p> <ul style="list-style-type: none"> - 주어 복원 대상이 아닌 술어의 목록을 현 재보다 확대해야 함. 특히 보조 용언, 의 사 보조 용언 구성 등이 주어 복원 대상 이 아닌 술어의 목록에 추가되어야 함 - 보조 용언에 대한 지침 내용은 표준국어 대사전의 기술을 근거로 하여 보조 용언 목록을 정하고 있으나 사전에는 기술되어 있지 않지만 실제로 보조 용언의 지위를

		<p>가지는 표현들이 다수</p> <ul style="list-style-type: none"> - 예: ‘철수는 학교에 가야 한다.’의 ‘하다’는 현재 보조 용언으로 취급하여 주어 복원 대상 아니므로 처리하는데, ‘철수는 학교에 가야 된다.’의 ‘되다’는 본용언으로 취급하여 주어 복원 대상으로 처리하고 있으며, 이러한 처리는 실제 언어 현상과 괴리가 있을 수 있음 - 의사 보조 용언 구성에 대한 지침 수정 - 예: ‘지각을 해서는 안 된다.’의 ‘-어서는 안 되다’는 의사 보조 용언 구성 목록에 포함하지만, ‘지각을 하면 안 된다.’의 ‘-면 안 되다’는 포함하지 않음. 즉, 전자의 ‘되다’는 주어 복원 대상이 아니고 후자의 ‘되다’는 주어 복원 대상이므로 주석 일관성에 문제가 발생할 수 있음
<p>‘3. 주격무형대용어복원 말뭉치_구축지침(2019)(191202).hwp’의 1쪽</p>	<p>주어 복원 대상이 아닌 술어 배제</p> <p>ㄷ. 서술어에 해당하지 않는 VP(_MOD)</p> <ul style="list-style-type: none"> - 관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고, 그러면서 등 모문과 분리되어 단독 문장을 이루지 못하는 술어 	<p><지침의 추가 혹은 수정></p> <ul style="list-style-type: none"> - 지침에는 ‘관해, 대해’ 등의 술어가 주어 복원 대상이 아니라고 명시 - 국립국어원에서는 이러한 술어들에 대해 구문분석 지침에서 별도로 기술된 바 없다는 점에서 이들이 주어를 가질 수 있다고 판단하며, 표층형에 주어가 없는 경우 이를 복원하지 않음 - 이러한 구문 분석 결과가 주격 무형대용어 복원 주석에 영향을 줄 경우, 지침에 명시해 줄 필요가 있음. 주격 무형대용어 복원의 입장에서는 어떤 술어들이 지침에 주어 복원 대상이 아니라고 기술되어 있는 경우, 이들이 애초에 주어를 가지지 않는 것으로 이해할 가능성이 있음 - 또한, 어떤 술어가 주어를 가지기는 하지만 그것이 생략된 경우에 복원하지 않는다는 점 자체에 문제가 있는 것은 아닌지 검토가 필요함.

		<ul style="list-style-type: none"> - 주어를 가지는 술어는 그것이 실제 문장에서 생략되었을 경우 그것을 복원하고, ‘-에 대해’의 ‘대해’와 같이 술어에서 유래하였으나 현재 불변화사의 지위를 가지는 것은 주어를 가지지 않는 것으로, 따라서 주어 복원 대상도 아니라고 보는 것이 더욱 주석 일관성을 담보할 수 있음
<p>‘3. 주격무형대용어복원 말뭉치 - 구 축 지 침 (2019)(191202).hwp’의 1쪽</p>	<p>다. 선행어 = 전방조응 + 후방조응</p>	<p><지침의 추가></p> <ul style="list-style-type: none"> - 생략된 주어의 선행어를 결정하는 순서가 제시되어 있는데, 모호한 점이 있음 - 주어가 문서 내에서 복원 가능한 경우, 그 순서는 ‘문장 내 가능 → 문장 내 불가능’이고, ‘문장 내 가능’에서는 ‘① 용언 관형어의 수식 대상 → ②같은 주격을 갖는 공지시 표현 → ③그 외 같은 문장 내 표현’ - 그러나 ③으로 판단되는 경우에도 같은 자격을 갖춘 선행어 후보가 여러 개여서 그 중에 어떤 것을 선행어로 판정해야 할지 알 수 없는 때도 있어 ③번 조건에 대하여 더 세분화하여 지침을 제시할 필요가 있음
<p>‘3. 주격무형대용어복원 말뭉치 - 구 축 지 침 (2019)(191202).hwp’의 5쪽</p>	<p>바. ‘-에서’ 주격 조사, 부사격 조사 여부 유의</p>	<p><지침의 구체화></p> <ul style="list-style-type: none"> - 지침에 ‘-에서’는 주격 조사로만 분석 가능할 때는 해당 성분을 주어로 분석하고, 주격과 부사격 조사로 모두 분석 가능할 때는 해당 성분을 부사어로 분석하는 것이 원칙이나, 실제 분석에서는 이 두 경우 중 어디에 해당하는지 판단하기가 쉽지 않음 - 더 다양한 용례를 대상으로 자세히 설명해주는 것이 필요
<p>‘3. 주격무형대용어복원 말뭉치 - 구 축 지 침 (2019)(191202).hwp’</p>	<p>사. 인칭</p>	<p><지침의 구체화></p> <ul style="list-style-type: none"> - 생략된 주어의 선행어를 선택할 때 인칭이 같아야 한다는 조건이나 설명이나 예시 보완 필요 - 예: 현재 제시된 예는 일상적 사용역에 나타난 1인칭의 경우인데, 공적 사용역에서

의 5쪽		는 대명사가 아니라 일반 명사구로 1인칭을 표현하기도 하는 등 변이가 있음
‘의존구문 분석_가이드라인.pdf (ver. 2015/12/16)’의 10-11쪽	(2) 관형절 내포문 분석 방법 (2.3) 모문과 내포문의 주어가 같고, 서술어가 다른 경우	<p><지침의 추가></p> <ul style="list-style-type: none"> - 모문과 내포문의 주어가 같은 경우, 내포문의 종류에 따라 분석 방식이 다름. 특히, 관형절의 처리 방식에 주의가 필요 - 그러나 이 점이 현재 구문분석 지침에만, 그것도 다소 불명확하게 그리고 부분적으로 기술되어 있을 뿐이므로 주격 무형대용어 지침에도 자세한 기술이 필요 - 특히, 3개 이상의 절이 내포되어 있고, 그 절들에 공통적인 주어가 앞에 1회만 나타날 때 대한 상세한 기술이 필요 (먼저 나오는 절 2개가 모두 관형절인 경우, 뒤에 나타난 관형절을 상위문처럼 취급하여 거기에 문면의 주어가 의존하는 것으로 분석한다는 것)
(없음) (‘3. 주격 무형대용어 복원말뭉치_구축지침 (2019)(191202).hwp’와 관련)	내포문의 유형 판단 문제	<p><지침의 추가></p> <ul style="list-style-type: none"> - 내포문의 유형 판단에 관한 기술 추가 필요 - 주어진 술어 활용형만을 가지고 그 내포문이 어떠한 유형의 것인지 판단하면 안 되고, 해당 활용형이 뒤에 (의사) 보조 용언으로 연결되어 있으면 그 구성 전체를 기준으로 내포문의 유형을 판단해야 함 - 예: ‘철수는 영희에게 빵을 줄 수 있으며 꽃은 이미 주었다.’에서 ‘줄’은 그 자체로는 관형사형이지만 뒤에 ‘-을 수 있-’이라는 의사 보조 용언 구성과 결합하였으므로 해당 내포문은 관형절이 아니라 부사절로 판단. 이에 따라 문장에 나타난 ‘철수는’은 ‘주었다’가 아니라 ‘줄’에 의존하는 것으로 보고, ‘주었다’의 주어를 복원해야 함. - 내포문의 유형은 주어 복원에 직접 영향을 미치므로, 내포문 유형의 판단에 (의사) 보

		조 용언 구성까지 고려해야 한다는 점을 지침에 명시할 필요가 있음.
(없음) (‘3. 주격 무형대용어 복원말뭉치 _구축지침 (2019)(1912 02).hwp’ 와 관련)	형태 축약형의 분석	<p><지침의 추가></p> <ul style="list-style-type: none"> - 주어 복원과 밀접히 관련된 형태 축약형의 분석 방법 추가 필요 - 예: ‘철수는 “너무 기뻐다” 라며 “춤을 추고 싶다” 라고 하였다.’ 에서 ‘라며’ 는 ‘이라고 하며’ 의 축약형으로, ‘철수는’ 이 이에 의존하고 맨 뒤의 ‘하였다’ 는 주어 복원 대상 술어로 복원
‘3-1.주격 무형대용어 복원 구축 구어 세부 분석 _191204.hwp’ 의 1쪽	<p>1, 기본 원칙</p> <ul style="list-style-type: none"> - 문어 분석과 구어 분석의 큰 틀은 동일함. - <u></u>태그를 하나의 문장 단위로 보고 분석함. 	<p><지침의 구체화 혹은 수정></p> <ul style="list-style-type: none"> - 지침에는 기술되어 있지 않지만, 관형절과 그 피수식 명사(구)가 서로 다른 발화 단위(<u>)에 위치하고 그 명사가 관형절의 생략 주어로 인식될 경우, 그것을 선행어로 주석할 것을 추가해야 함 (국립국어원 해석)
(없음) (‘3-1.주격 무형대용어 복원 구축 구어 세부 분석 _191204.hwp’ 과 관련)	구어의 직접인용/간접인용 구별	<p><지침의 추가></p> <ul style="list-style-type: none"> - 구어 지침에 직접 인용과 간접인용을 어떠한 기준 필요 - 문어에서는 따옴표의 유무를 기준으로 직접 인용과 간접인용을 구별하고 있으나 (이 역시 현 지침에는 없으므로 명시해야 함) 구어에서는 어떠한 종류의 인용이든 따옴표 표시 자체가 없으므로, 이에 따라 구어에서 직접 인용과 간접인용을 어떻게 구별할지에 대한 지침을 마련할 필요가 있음. 어떠한 종류의 인용인지에 따라서 복원되는 주어의 형식(인칭)이 달라질 수 있음.
(없음) (‘3-1.주격 무형대용어 복원 구축	구어에서 상대방 발화에 맞장구치는 ‘그러하다/그렇다’, ‘맞다’ 등의 분석	<p><지침의 추가></p> <ul style="list-style-type: none"> - 현재 구어 지침에는 없는 내용이지만, 구어에서 상대방 발화에 맞장구치는 ‘그러하다/그렇다’, ‘맞다’ 등의 분석에 대

구어 세부 분석 _191204.hwp' 과 관련)		<p>하여 세부 지침을 정해야 함</p> <p>- 구어에는 ‘그렇습니다’, ‘그렇죠(그쵸)’, ‘그렇군요’ 등, 그리고 ‘맞아요’ 등의 맞장구 표현이 매우 많이 나타나는데, 이들의 주어를 복원하는 것이 매우 까다롭고(선행어를 무엇으로 해야 할지 어려움), 더 근본적으로는 주어 복원의 필요성도 재고해 볼 필요가 있음.</p>
--------------------------------------	--	---

⑤ 상호참조 해결 말뭉치 구축 지침 검토 의견

지침 위치	지침 내용	검토 의견
4쪽	상 호 참 조 해 결 (Coreference resolution)은 임의의 개체(entity)에 대하여 다른 표현으로 사용되는 단어들을 찾아, 서로 같은 개체로 연결해주는 자연어처리 문제이다	- 상호참조대상이 되는 멘션들에 대한 해석의 차이에 대한 명확한 기준이 제시될 필요가 있음
5-6쪽	<p>멘션 탐지 단계는 의존구문 트리에서 등장하는 모든 명사구를 멘션으로 잡는다. 문서 내의 멘션을 탐지하기 위해 다음과 같은 규칙을 적용한다.</p> <ul style="list-style-type: none"> ● 멘션은 기본적으로 형태 단위로 처리한다. ● 수식 정보를 포함한 멘션 생성(즉, 명사구)3 ● 개체명의 원자성4 ● 중심어의 중복 처리5 ● 대명사 분류 	- “형태”의 명확한 기준이 요구됨

5-6쪽	<p>멘션 탐지 단계는 의존 구문 트리에서 등장하는 모든 명사구를 멘션으로 잡는다. 문서 내의 멘션을 탐지하기 위해 다음과 같은 규칙을 적용한다.</p> <ul style="list-style-type: none"> • 멘션은 기본적으로 형태 단위로 처리한다. • 수식 정보를 포함한 멘션 생성(즉, 명사구)³ • 개체명의 원자성⁴ • 중심어의 중복 처리⁵ • 대명사 분류 	<ul style="list-style-type: none"> - 멘션의 기준 명확화 필요 - 멘션 최소 단위: 개체명이 있으면 개체명, 개체명이 없으면 어절, 개체명과 어절에서 조사나 기타 (문장) 부호는 제외 - 개체명의 정보가 작업 환경에 명확히 드러날 필요가 있음
5-6쪽	<p>멘션 탐지 단계는 의존 구문 트리에서 등장하는 모든 명사구를 멘션으로 잡는다. 문서 내의 멘션을 탐지하기 위해 다음과 같은 규칙을 적용한다.</p>	<ul style="list-style-type: none"> - 앞서 국립국어원에서는 멘션의 최장 단위를 정하는 데 있어서 구문 분석 시범 검증용 말뭉치를 기준으로 삼음 - 하지만 멘션의 최소 단위인 개체명과 마찬가지로 최장 멘션도 분명한 기준 없이 본 사업단에서 임의로 분석한 원리를 적용한다면 객관성이 떨어지고 품질이 떨어질 수 있음
12쪽	<p>멘션추출이 가능한 단어(‘이다, 이었다, 로서, 이며’)와 불가능한 단어(‘라면서 등 ‘이다’ 이외의 단어)를 구분하여 가능한 단어는 추출, 불가능한 단어는 제외한다.</p>	<ul style="list-style-type: none"> - 지침을 따르면 아래 예시와 같은 문제가 발생할 수 있음 - 예: 오바마는 미국인이다. - {오바마, 미국인} - 미국인의 인구는 3억 명 가까이 되고 그 중의 한 명이 오바마인데, 지정사구 규칙을 따르게 되면 위와 같이 상호참조할 수 밖에 없음. - 엄밀히 말하면 오바마는 ‘미국인이라는 속성’을 가진 것이고 속성과 특정 개체는 상호참조되면 안됨. - ‘특정한 개체를 지시하는 서로 다른 언어적인 표현을 찾는 것’이 상호참조의

		<p>대전제라고 보면 여기서 속성은 개체를 지시하는 것으로 볼 수 없기 때문</p> <ul style="list-style-type: none"> - 하지만 현재 상호참조되는 대부분은 이 지정사구에 기인한 것이기 때문에 여기에 대한 좀 더 명확한 의미적 구분이 제시될 필요가 있음
17쪽	두 개의 상호참조 해결 사이에 교차되는 멘션이 있는 경우 상호참조에 대한 정의가 필요하다.	<ul style="list-style-type: none"> - ‘교차’의 정의 - ‘엔티티 1과 엔티티 2에서 교집합 멘션이 있으면 하나의 엔티티로 묶어서 상호참조한다’ - 는 적용되는 경우가 매우 한정적이므로 더욱 명확한 정의가 필요
지침에 없는 검토 사항		<ul style="list-style-type: none"> - 특성성과 총칭에 관한 문제
지침에 없는 검토 사항		<ul style="list-style-type: none"> - 의존 명사의 멘션 처리 기준
지침에 없는 검토 사항		<ul style="list-style-type: none"> - 띄어쓰기와 멘션
지침에 없는 검토 사항		<ul style="list-style-type: none"> - 불투명한 문맥(opaque context)의 명사구의 상호참조 여부
지침에 없는 검토 사항		<ul style="list-style-type: none"> - 고정된 숫자 값의 해석 예외

⑥ 구문 분석 말뭉치 구축 지침 검토 의견

지침 위치	지침 내용	검토 의견
없음	띄어쓰기 오류	<ul style="list-style-type: none"> - 실제 작업 시, 띄어쓰기 오류로 인해 형태소 분석이 잘못되었을 때 구문 분석 주석에 대한 TTA 지침이 보완되어야 함 - 구문 분석 전체의 품사와 기능이 두 가지 이상으로 해석될 수 있으므로 주석 일관성에 문제가 생길 수 있음
없음	인용절 내포문	<ul style="list-style-type: none"> - 인용문이 여러 개의 문장으로 되어 있을 때, 각 문장의 root 서술어의 의존 방법이 문제가 될 수 있음 - 지침의 예문은 모두 인용문이 한 개인 예시로 이루어져 있으며, ‘내포문은 내포문 내의 주어와 서술어 간의 의존 관계를 연

		결하여 분석한다(p.12)'라는 규정밖에 없어, 주석자들 사이에 약간의 의견 차이가 발생
없음	몇몇 부사어의 처리	<ul style="list-style-type: none"> - '대해, 통해, 위해, 관해' 등의 부사어는 마치 삽입구처럼 해석되는 경우가 많아 국립국어원에서 추가로 배포한 세부 지침에서 "(1.4) 복문의 해석이 중의적일 때에는 가능한 의미 중에서 가장 가까이에 후행하는 절에 의존한다. 단, 복문의 해석이 단일할 때에는 해석에 따라 분석한다."에 해당하는 것으로 해석하여, 주어를 '대해, 통해, 위해, 관해'에 연결하지 않음 - 그러나 국립국어원에서는 이 해석이 잘못된 것이라고 하였고, 12월 31일 위와 같은 부사절을 모문 주어와 연결하는 것을 채택함
없음	기사 제목	<ul style="list-style-type: none"> - 특수 기호로 묶여 있는 기사 제목의 품사는 무엇인지, 이것이 기사의 첫 번째 문장핵에 의존하여야 하는지, 기사의 끝에 의존하여야 하는지 등의 기준이 필요함

⑦ 의미역 분석 말뭉치 구축 지침 검토 의견

지침 위치	지침 내용	검토 의견
1쪽	필수역은 Korean PropBank와 Etri의 논항 정보(FrameSet)를 기준으로 U-PropBank와 우리말샘을 순차적으로 활용하여 주석한다.	<ul style="list-style-type: none"> - 지침은 참조 격틀에 문제가 있는 경우를 고려하고 있으므로 다음과 같이 참조 격틀에 문제가 있는 경우에 주석 방법을 지침에 추가해야 할 필요가 있음 - 참조 격틀별 문제점 <ol style="list-style-type: none"> 1) Korean Propbank <ul style="list-style-type: none"> - 술어 번호가 없는 경우 (예: 일어나) - 필수 논항의 번호가 빠진 경우 (예: 유행.01) - 격틀 정보와 예문 분석이 일치하지 않는 경우 (예: 떠나.01) 2) U-PropBank

		<ul style="list-style-type: none"> - U-PropBank가 필수 논항에 번호를 부여하는 방식이 Korean PropBank와 ETRI와 달라서 참조 격들 간의 필수 논항 부여방식에 차이가 있음. - 예) Korean PropBank와 ETRI에서 ARG3은 ARG2가 부여된 경우에만 나타나지만 U-Propbank에서는 ARG2가 부여되지 않아도 ARG3을 부여함. - 필수 논항이 잘못 정의됨 (예: 들려오다 000000 대상역이 빠지고 대신 행동주가 정의되어 있음. <p>3) 우리말샘</p> <ul style="list-style-type: none"> - 우리말샘은 격들 정보를 제공하지 않음 - 우리말샘이 제공하는 조사 정보를 치환하는 지침을 추가할 필요가 있음 - 다만, Korean Propbank, ETRI, U-Propbank가 각각 필수 논항에 번호를 부여하는 방식이 다른데, 필수 논항에 대한 번호 부여방식을 통일하는 것은 지침이 다룰 수 있는 범위를 넘어설 수 있음
2쪽	술어의 논항 정보에 이미 행위주가 정의되어 있어서 사동주에 같은 ARG0(행위주)를 주석할 수 없는 경우에는 ARG0을 주석한다.	<ul style="list-style-type: none"> - 해당 지침은 1) 격들에 사동주가 정의되어 있지 않으면 주석자의 판단에 따라 ARG0을 주석해도 된다는 것인지 2) 격들 안에 사동주와 행위주가 모두 정의되어 있는 경우 사동주와 행위주를 구분하여 주석하는 방법인지 명확하지 않아 지침 수정이 필요
2쪽	4. 부가 의미역 태깅 지침	<ul style="list-style-type: none"> - 하나의 주석 단위에 대해서 복수의 부가 의미역 주석이 가능할 때 주석자의 지식과 직관이 아닌 지침에 따라 판단할 수 있도록 부가 의미역에 대한 상세한 추가

		<p>지침이 필요</p> <p>1) ARGM-LOC</p> <ul style="list-style-type: none"> - 추상적인 장소도(예: 편지에서) LOC로 분석될 수 있는지 추가 지침이 필요 <p>2) ARGM-PRD</p> <ul style="list-style-type: none"> - MNR, INS, PRD의 경계를 명확하게 구분할 수 있는 추가 지침이 필요 - 예) 급히 부르다 - 주석자의 직관에 따라 ‘급히’는 MNR이나 PRD가 될 수 있으므로 MNR가 PRD의 경계를 구분하는 구체적인 지침이 필요 <p>3) ARGM-TMP</p> <ul style="list-style-type: none"> - 사건(예: 회의에서)도 TMP로 분석될 수 있는지 추가 지침이 필요 <p>4) ARGM-CAU</p> <ul style="list-style-type: none"> - CAU와 PRP가 모두 가능한 경우에 어느 것을 선택해야 하는지에 대한 추가 지침이 필요 <p>5) ARGM-ADV</p> <ul style="list-style-type: none"> - 부사적 어구가 ADV의 목록에 있어도 해당 어구가 나타난 통사적 구조와 내포하고 있는 의미에 따라 ADV이 아닌 다른 부가 의미역으로 분석되어야 하는 경우가 있으므로 ADV의 정의와 적용 조건을 제약하는 추가 지침이 필요
4쪽	5.4.2. 접속조사	<ul style="list-style-type: none"> - 공동격 조사와 접속조사를 구분하고 공동격 조사에 대한 추가 지침이 필요
4쪽	<p>5.4.2. 접속조사</p> <p>예 4) 너하고 나 (ARG1/ARG2) (ARG1) 달라. (중의적이면 문맥을 고려하여 판단함.)</p>	<ul style="list-style-type: none"> - 의미역 분석은 문장 단위로 이루어지기 때문에 주석 과정에서 문맥을 고려할 수 없으므로 문맥을 고려하지 않고 중의성을 일관성 있게 해결할 수 있는 추가 지침이 필요
5쪽	6.1.1. 기본 원칙	<ul style="list-style-type: none"> - 일부 부가역의 범위와 조사를 활용하는

	8. 일부 부가역은 조사를 포함하여 논항으로 인식한다.	방법에 대한 추가 지침이 필요 - 또한, 신문 기사 제목과 같이 조사가 생략된 경우에 대한 추가 지침도 필요
--	--------------------------------	---

사업 책임자	최기선(한국과학기술원 전산학부 석좌교수)
사업 참여자	강규영 (서울대학교 국어국문학과 박사수료)
	강소연 (고려대학교 언어학 석사)
	강아름 (고려대학교 언어정보연구소 연구교수)
	김건영 (고려대학교 국어국문학과 박사수료)
	김건태 (한국과학기술원 전산학부 석사과정)
	김동환 (한국과학기술원 정보전자연구소 연구원)
	김민지 (서울대학교 국어국문학과 석사과정)
	김선영 (서울대학교 국어학 박사)
	김선주 (미국 하와이주립대학교 언어학 석사)
	김세은 (한국과학기술원 시맨틱웹첨단연구센터 연구원)
	김소희 (고려대학교 언어학 박사)
	김아름 (서울대학교 국어국문학과 박사수료)
	김유겸 (서울대학교 국어국문학과 박사수료)
	김은경 (대전대학교 정경학부 빅데이터 사이언스 전공 교수)
	김주상 (서울대학교 국어학 박사)
	김지성 (한국과학기술원 전산학부 박사과정)
	김진동 (한국과학기술원 전산학부 겸임교수)
	김태우 (인하대학교 한국학연구소 연구교수)
	김태인 (서울대학교 국어학 박사)
	김한결 (서울대학교 국어국문학과 박사과정)
	남상하 (한국과학기술원 전산학부 박사수료)

노소은 (고려대학교 언어학 석사)

노영빈 (한국과학기술원 정보전자연구소 연구원)

목정수 (서울시립대학교 국어국문학과 교수)

박석원 (연세대학교 언어정보학 협동과정 석사과정)

박승희 (나라아이넷(주) 전무이사)

박용배 (서울시립대학교 국어학 박사)

박지용 (서울대학교 국어국문학과 박사수료)

박진호 (서울대학교 국어국문학과 교수)

박형진 (가천대학교 한국어문학과 교수)

박혜린 (고려대학교 언어학과 석사수료)

박혜승 (서울대학교 국어국문학과 박사수료)

백채원 (서울대학교 국어학 박사)

서반석 (서울대학교 국어학 박사)

손지은 (고려대학교 국어국문학과 박사수료)

송상헌 (고려대학교 언어학과 교수)

송영숙 (경희대학교 국어국문학과 박사수료)

신용남 (서울대학교 국어국문학과 박사수료)

심지수 (경희대학교 국어국문학과 박사수료)

안기경 (서울시립대학교 국어국문학과 박사수료)

안상민 (한국과학기술원 인공지능연구소 연구원)

연규동 (연세대학교 인문학연구원 교수)

오규환 (동덕여자대학교 국어국문학과 교수)

오태환 (연세대학교 국어국문학과 박사과정)

유민애 (서울대학교 국어교육연구소 객원연구원)

이경원 (고려대학교 언어학과 석사수료)

이민호 (한국과학기술원 전산학 석사)

이상희 (서울시립대학교 국어국문학과 박사과정)

이신복 (서울시립대학교 국어국문학과 박사수료)

이의중 (서울대학교 국어국문학과 박사수료)

이하은 (한국외국어대학교 정보통신공학과)

이호진 ((주)언어과학 기업부설 연구소 소장)

장고은 (서울대학교 국어국문학과 박사수료)

장원철 ((주)언어과학 상무이사)

장하연 (미국 남가주대학교 언어학과 박사수료)

정용빈 (한국과학기술원 전산학부 석사과정)

정유남 (고려대학교 강사)

정유성 (한국과학기술원 시맨틱웹첨단연구센터 연구원)

정혜린 (서울대학교 국어국문학과 박사수료)

최원석 (나라아이넷(주) 기업부설 연구소 부소장)

최윤지 (인하대학교 한국어문학과 조교수)

최진 (서울대학교 국어국문학과 박사수료)

최현수 (연세대학교 언어정보학 협동과정 석사과정)

함영균 (한국과학기술원 전산학부 박사과정)

허인영 (고려대학교 국어국문학과 박사수료)

허철훈 (한국과학기술원 전산학부 석사과정)

홍은영 (서울대학교 국어국문학과 박사수료)

담당 연구원

이승재(국립국어원 언어정보과장)

이현주(국립국어원 언어정보과 학예연구관)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9757

인쇄일: 2020년 2월 29일

발행일: 2020년 2월 29일

인 쇄: 충대문화사

※ “이 책은 국립국어원의 용역비로 수행한 ‘말뭉치 통합 검증’ 사업의 결과물을 발간한 것입니다.”