국립국어원 2021-01-02

발 간 등 록 번 호
11-1371028-000852-01

2020년도 국립국어원 말뭉치 통합 관리 지원

사업 책임자
이 의 종

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2020년도 국립국어원 말뭉치 통합 관리 지원’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2020년 9월 ~ 2021년 1월

2021년 2월 1일

사업 책임자: 이의종(㈜나라지식정보)

사업 수행자 나라지식정보 공동수급체
(㈜나라지식정보, ㈜언어과학)

사업 책임자 이의종

사업 참여자 고동현, 길혜빈, 김선영, 김은수, 김지원, 김태경, 김태우, 김한나, 김희숙, 박승희, 박영훈, 박융배, 박지용, 박진호, 박하선, 박혜승, 배준호, 손지은, 송상헌, 신용남, 신희원, 안대섭, 안의정, 유현조, 유혜선, 윤기현, 윤예진, 이강혁, 이경원, 이규환, 이민우, 이용규, 이재혁, 이주연, 장원철, 장하연, 정규상, 정우헌, 정유남, 정혜주, 조혜미, 최준호, 최 진, 황은하
(총 45명)

〈사업 수행자〉 나라지식정보 공동수급체 (㈜나라지식정보, ㈜언어과학)

사업 책임자	이의종 (㈜나라지식정보)
사업 참여자	고동현 (㈜나라지식정보) 길혜빈 (경희대학교 국어국문학과 박사수료) 김선영 (서울대학교 언어교육원 대우전임강사) 김은수 (한양대학교 한국언어문학과 학사) 김지원 (㈜나라지식정보) 김태경 (한양대학교 ERICA 창의융합교육원 부교수) 김태우 (부산대학교 국어국문학과 조교수) 김한나 (경희대학교 국어국문학과 박사과정) 김희숙 (㈜나라지식정보) 박승희 (㈜나라지식정보 전무이사) 박영훈 (㈜나라지식정보 부장) 박융배 (이화여자대학교 뇌융합과학연구소 빅데이터 경영연구소 연구원) 박지용 (서울대학교 국어국문학과 강사) 박진호 (서울대학교 국어국문학과 교수) 박하선 (㈜나라지식정보) 박혜승 (서울대학교 국어국문학과 강사) 배준호 (㈜나라지식정보) 손지은 (고려대학교 국어국문학과 박사수료) 송상헌 (고려대학교 언어학과 조교수) 신용남 (서울대학교 국어국문학과 박사수료) 신희원 (한양대학교 한국언어문학과 학사) 안대섭 (고려대학교 노어노문학과 강사) 안의정 (연세대학교 학부대학 강사) 유현조 (서울대학교 인문데이터과학연계전공 초빙부교수)

	유혜선 (㈜나라지식정보) 윤기현 (바이칼AI 대표) 윤예진 (서울대학교 국어국문학과 박사수료) 이강혁 (서울대학교 국어국문학과 박사수료) 이경원 (㈜나라지식정보) 이규환 (㈜나라지식정보) 이민우 (사이버한국외국어대학교 한국어학부 조교수) 이용규 (서울대학교 국어국문학과 박사과정) 이재혁 (㈜나라지식정보 수석연구원) 이주연 (㈜나라지식정보) 장원철 (㈜언어과학 상무이사) 장하연 (부산외국어대학교 영어학부 조교수) 정규상 (㈜나라지식정보 부장) 정우헌 (㈜나라지식정보) 정유남 (고려대학교 국어국문학과 강사) 정혜주 (㈜나라지식정보) 조혜미 (서울대학교 영어영문학과 학사과정) 최준호 (서울대학교 국어국문학과 박사수료) 최　진 (㈜나라지식정보) 황은하 (배재대학교 국어국문한국어교육학과 조교수)

<국문 초록>

2020년도 국립국어원 말뭉치 통합 관리 지원

이 사업은 국가 주도로 구축한 다층위 말뭉치에 대한 통합 관리, 운영 방안을 모색하고, 고품질 말뭉치 수요 증대에 따른 지속적인 품질 관리를 지원하며, 말뭉치 사용자 요구에 맞춘 정확하고 빠른 대응 체계를 구축함과 함께 연구, 산업 등 말뭉치 관련 전문 수요자 요구를 관리하는 데에 이바지함에 목적이 있다. 이 목적에 기여하기 위해 '음성 말뭉치의 통합과 정비', '7개 층위 분석 말뭉치의 통합과 정비', '전문가 토론회 개최'라는 세 가지 과업을 수행하였다.

음성 말뭉치의 통합과 정비 과업에서는 2018년도 구축 일상 대화 음성/전사 말뭉치 약 270만 어절에 대하여, 음성이 전사 단위와 일치하도록 221개 음성 파일의 발화 단위별 분절 및 개인정보 비식별화 처리, 즉 음성 정제를 수행하였다. 그리고 약 8,700어절의 대본을 연령대별, 성별로 다양한 낭독자가 낭독, 녹음한 서울말 낭독체 약 88,000개 음성 파일에 대하여 대본과 낭독내용간 오차를 기록하고 전사하였다.

7개 층위 분석 말뭉치의 정비 및 통합 관리 과업에서는 JSON 형식 검증, 층위 통합과 주석 오류 검증, 지침 보완 및 예시 확충, 그리고 유지 보수 사업단 말뭉치와의 비교를 수행하였다. 구체적인 수행 내용은 다음과 같다.

2019년도 구축 7개 층위 분석 말뭉치(형태 분석, 어휘 의미, 개체명, 상호참조 해결, 구문 분석, 의미역, 주격 무형 대용어 복원)를 대상으로, JSON 문서 형식의 일관성과 오류를 검증하였다. 그 뒤에 말뭉치의 층위 형식을 통합하고 말뭉치 내 지정된 범위의 문서에서 주석 오류를 분석하고 수정하였다. 아울러 사업단 내 전문가 의견을 수렴하여 말뭉치 구축 지침을 정비하고 확충 예시 자료를 제안하였다. 이상의 작업과 동시에, '유지 보수 사업단' 말뭉치와의 연계 검증 작업을 진행하였다. 층위 형식이 통합된 말뭉치를 유지 보수 사업단에 전달하고, 각자 수정을 진행한 뒤 사업 종료 시기에 본 사업의 말뭉치와 유지 보수 사업단 말뭉치의 수정 결과를 비교하였다.

전문가 토론회 개최 과업에서는 '인공지능 시대를 향한 우리말 빅데이터의 활용'이라는 제목의 토론회를 개최하였다. 한 명의 사회자와 세 명의 전문가 패널이 참석하는 온/오프라인 병행 토론회를 열고, 강연, 토론 및 청중 질문 시간을 가졌다. 약 150명의 청중이 참여하였다.

국가 주도로 구축한 다층위 말뭉치들의 응용 가능성을 극대화하기 위해 말뭉치의 통합과 정비를 수행하고, 전문가 의견 수렴을 진행하여 앞으로 이루어질 말뭉치의 구축과 활용의 비전을 제시한 데에 본 사업의 의의를 둘 수 있다.

주요어: 말뭉치 통합 관리 지원, 음성 말뭉치 정비, 음성 정제, 층위 통합, 분석 말뭉치 검증

차 례

제1장 사업 개요

제2장 음성 말뭉치의 통합과 정비

제3장 7개 층위 분석 말뭉치의 통합과 정비

제4장 전문가 토론회 개최

제5장 결론과 제언

제 1 장

사업 개요

1. 사업 목적

이 사업은 국가 주도로 구축한 다층위 말뭉치에 대한 통합 관리, 운영 방안을 모색하고, 고품질 말뭉치 수요 증대에 따른 지속적인 품질 관리를 지원하며, 말뭉치 사용자 요구에 맞춘 정확하고 빠른 대응 체계를 구축함과 함께 연구, 산업 등 말뭉치 관련 전문 수요자 요구를 관리하는 데에 이바지함에 목적이 있다.

4차 산업혁명 시대는 언어 데이터를 학습하는 기계학습의 인공지능이 근간이 된다. 이 학습이 목적에 맞게 이루어져 실용적인 결과물을 창출하기 위해서는 정교화된 고품질의 언어 데이터가 필요하다. 그리고 고품질의 언어 데이터가 갖추어야 할 요건을 파악하기 위해서, 그리고 구축된 언어 데이터의 응용 방향과 잠재 가치를 파악하기 위해서 산업계, 학계 및 시민의 의견을 청취할 필요가 있다.

국립국어원에서는 '4차 산업혁명 대비 국어 빅데이터(말뭉치) 구축' 사업을 통해 인공지능에 활용될 수 있는 한국어 기초 자원이 될 수 있는 대규모 말뭉치를 구축해 왔다. 그 결과물은 순차적으로 국립국어원에서 운영하는 「모두의 말뭉치」 웹페이지(https://corpus.korean.go.kr/)를 통해 공개되고 있다. 국가 주도로 이루어진 대규모 문어 원시 말뭉치, 대규모 구어 원시 말뭉치, 그리고 이 원시 말뭉치를 바탕으로 형태 분석, 어휘 의미, 개체명, 상호참조 해결, 구문 분석, 의미역, 주격 무형 대용어 복원 등이 주석된 분석 말뭉치는 학술 연구와 산업 발전에 다방면으로 기여할 수 있는 빛나는 성과이다. 그러나 대규모 말뭉치가 한국어 기초 자원으로서 그 응용 가능성을 더욱 폭넓게 발휘하려면, 지속적인 후속 지원과 자료의 층위 통합, 사용자 의견 수렴이 이루어져야 할 것이다.

이와 같은 후속 관리의 필요에 기여하기 위해 본 사업에서는 구축된 말뭉치의 통합과 말뭉치에 대한 의견 수렴을 수행하고자 한다. 통합을 위해서는 구어 말뭉치의 음성과 전사간의 대응관계 정비, 분석 말뭉치의 층위간 형식 통일, 분석 말뭉치의 주석 정비를 수행한다. 의견 수렴을 위해서는 전문가적 관점에서의 분석 말뭉치 구축 지침 수정 제안과 말뭉치 활용을 위한 전문가 토론회 개최를 수행한다.

2. 사업 수행 범위

본 사업의 수행 범위는 다음과 같다. 첫째, 음성/전사 말뭉치를 정비하고 통합 관리하는 것이다. 둘째, 7개 층위 분석 말뭉치를 정비하고 통합 관리하는 것이다. 셋째, 말뭉치 활용을 위한 전문가 토론회를 개최하는 것이다.

음성/전사 말뭉치의 정비 및 통합 관리의 구체적인 내용은 다음과 같다. 2018년도 구축 일상 대화 음성/전사 말뭉치 약 270만 어절에 대하여, 음성이 전사 단위와 일치하도

록 221개 음성 파일의 발화 단위별 분절 및 개인정보 비식별화 처리를 수행한다. 그리고 약 8,700어절의 대본을 연령대별, 성별로 다양한 낭독자가 낭독, 녹음한 서울말 낭독체 약 88,000개 음성 파일에 대하여 대본과 낭독내용간 오차를 기록하고 전사한다.

7개 층위 분석 말뭉치의 정비 및 통합 관리의 구체적 내용은 다음과 같다. 2019년도 구축 7개 층위 분석 말뭉치(형태 분석, 어휘 의미, 개체명, 상호참조 해결, 구문 분석, 의미역, 주격 무형 대용어 복원)를 대상으로, 층위 형식을 통합하고 말뭉치 내 지정된 범위의 문서에서 주석 오류를 분석하고 수정한다. 지정된 범위의 문서란, 전체 말뭉치의 약 20% 분량에 달하는 문서로 모든 층위에서 동일한 원시 말뭉치에 대응되도록 사전에 지정된 문서들이다. 단, 상호참조 해결 층위는 예외적으로 상호참조 해결 말뭉치와 2019년도 '말뭉치 통합 검증' 사업에서 구축한 검증 말뭉치 사이의 오차분에 대해서만 오류 분석과 수정을 진행한다. 아울러 말뭉치 구축 지침을 정비하고 예시 자료를 확충하며, 말뭉치 사용자 제안 사항 검토 및 수정 후보를 제시한다. 이상의 작업과 동시에, 각 말뭉치의 구축 사업단이 구축 사업의 후속 조치로서 산출한 '유지 보수 말뭉치'와 연계한 검증을 진행한다. 그 절차는 다음과 같다. 층위 통합이 이루어진 말뭉치를 각 구축 사업단에 발송한 뒤, 구축 사업단이 본 과업과는 별도로 외부에서 유지 보수 작업을 수행한 말뭉치를 이후 수령한다. 수령한 유지 보수 말뭉치의 수정 전후를 비교 검증하고, 본 과업의 결과물과 유지 보수 말뭉치의 내용을 비교 검증한다.

말뭉치 활용을 위한 전문가 토론회의 구체적 내용은 다음과 같다. 사업 기간 중 1회의 말뭉치 전문가 토론회를 개최한다. 말뭉치 등 빅데이터 관련 전문가를 초청하여 토론회를 열고, 강연과 토론 및 청중 질의응답을 진행하여 대규모 말뭉치의 활용에 대한 이해를 제고한다.

3. 사업 수행 일정

본 사업은 2020년 9월 4일에 착수하여 2021년 2월 1일까지 약 5개월간 수행되었다. 세부 과업의 구체적인 수행 경과는 다음 표에 나타난 바와 같다. 모든 사업 진행 상황에 대해 매주 주간 보고서를 제출하였으며 특이사항에 대한 보고 및 업무 정보 교환을 수시로 수행하였다. 종료 보고는 2021년 1월 28일에 진행하였다.

<표 1> 사업 수행 경과

과업 구분	9월	10월	11월	12월	1월
2018년도 일상대화 말뭉치 음성 정제					
서울말 낭독체 대본-음성-전사 통합					
7개 층위 분석 말뭉치 검증 및 통합					
7개 층위 분석 말뭉치 전문가 검증 및 유지 보수 말뭉치와의 비교					
전문가 토론회 개최					

제 2 장

음성 말뭉치의 통합과 정비

1. 음성 말뭉치 통합 및 정비의 대상과 범위

본 사업의 음성 말뭉치 통합 및 정비의 대상은 크게 '2018년도 일상 대화 말뭉치'와 '서울말 낭독체 말뭉치'의 두 가지로 대별할 수 있다. 이 두 가지 말뭉치의 통합 및 정비 범위는 다음과 같다.

2018년도 일상 대화 말뭉치의 통합 및 정비는 2018년도에 수집된 일상 대화 음성 파일 221개에 대한 음성 정제를 수행하는 것이다. 221개 파일은 파일당 1시간 15분 내외의 분량으로, 전체 시간은 약 262시간 38분이다. 이 과업의 성격은 2019년도 일상 대화 말뭉치 구축(국립국어원, 2019)의 구조를 참고하여 이해할 수 있다. 2019년도 일상 대화 말뭉치 구축은 '음성 자료 수집', '전사', '음성 정제'의 세 가지 세부 과업을 통해 대규모의 음성 자료와 그에 세밀히 대응되는 전사 자료의 짝을 산출하는 사업이었다. 원칙적으로 본 과업 '2018년도 일상 대화 말뭉치의 통합과 정비'의 목표는 2018년도 일상 대화 말뭉치에 대해 2019년도 일상 대화 말뭉치 구축 사업과 동일한 성격의 결과물을 산출하는 데에 기여하는 것인데, 2018년도 일상 대화 말뭉치는 구축 과정에서 음성 자료 수집과 전사가 이미 이루어졌으므로 본 사업에서는 음성 정제만이 과업 범위가 된다. 즉, 기존 전사 말뭉치에 세밀히 대응되는 음성 데이터를 산출하는 것이 이 과업에서의 '통합'의 의미가 된다. 본 사업에서 수행한 음성 과업의 주요 내용은 '발화 단위에 따른 음성 분절', '분절 단위 시간 정보의 기록', '개인정보 비식별화'이다.

서울말 낭독체 말뭉치의 통합 및 정비는 2005년에 배포된 '서울말 낭독체 발화 말뭉치'에 대한 전사 정비를 수행하는 것이다. 이 말뭉치는 19종의 대본 텍스트와 2대 이상 서울 경기 지역에 거주해 온 서울말 화자 120명이 그 대본을 낭독한 음성 파일 약 88,800개로 이루어진 것으로, 음성 파일의 전체 시간은 약 150시간 44분이다. 그런데 이 말뭉치는 대본을 읽은 낭독 음성이 바탕이 되므로, 이상적으로는 대본과 음성이 일치하여야겠지만, 낭독자의 발화 실수 등으로 대본과 음성이 일치하지 않는 지점들이 있다. 이런 불일치분에 대한 음성 기준의 전사를 수행하여 음성과 텍스트의 대응 정밀도를 높이는 것이 본 과업의 목표이다.

2. 2018년도 일상 대화 말뭉치의 통합과 정비

2.1. 과업 수행 준비

본 과업은 일상 대화 음성 녹음 221개 대화 15,758분 분량의 자료에 대하여, 발화 단위에 따른 음성 분절과 개인정보 비식별화를 수행하는 과업이다. 이를 위하여 국립국어원으로부터 '2018년 국어 말뭉치 연구 및 구축' 사업의 결과물인 221개 음성 녹음 파일과 국립국어원 구어 말뭉치 파일(JSON 형식), 전사 지침 문서와 음성 정제 지침 문서를

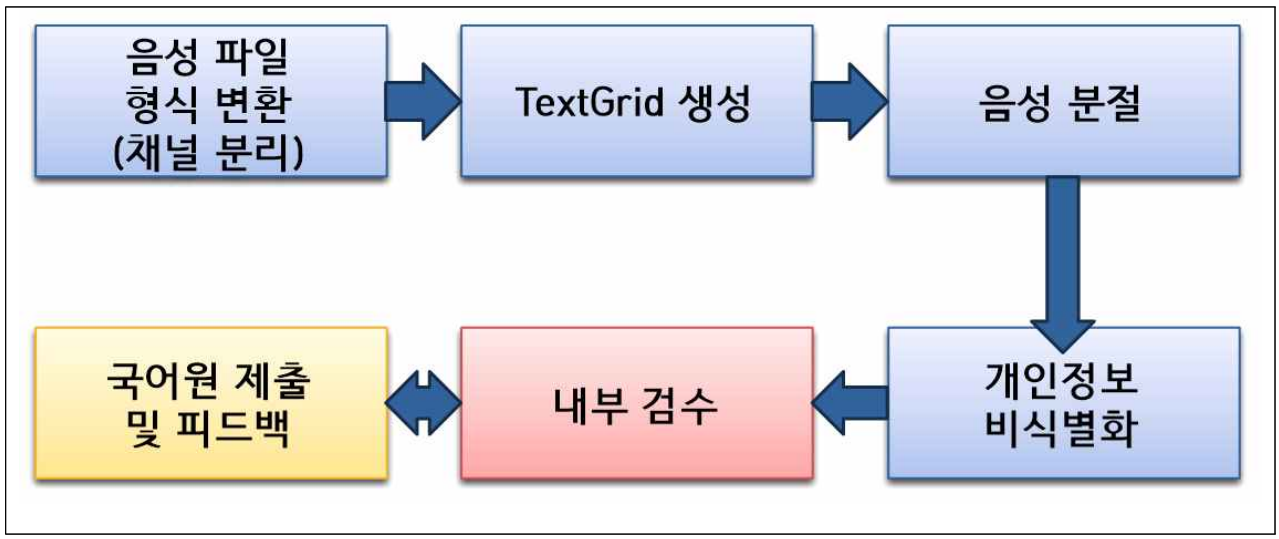
수령하였다.

음성 녹음 파일은 wav 형식 파일로서, 대화에 참여하는 두 사람의 발화가 각기 다른 마이크를 사용해 녹음된 2채널 파일이다. 음성 정제를 위해 우선 발화자별 채널을 분리하였다. 채널 분리에는 FFMPEG 소프트웨어를 사용하였다. 음성 분절에는 praat 소프트웨어를 사용하였으며, 음성 분절 단위 시간 정보의 기록을 위하여 역시 praat 소프트웨어로 TextGrid 파일을 생성하여 제출하기로 하였다.

2019년도 일상 대화 말뭉치 구축 사업에 참여한 경험이 있는 구성원이 주축이 되어 온라인과 오프라인에서 음성 정제 절차 교육 및 실습을 진행하였다.

2.2. 정제

음성 정제 절차 교육 및 실습을 진행한 뒤, 개별 구성원에게 작업 담당 파일을 분배하여 작업을 수행하였다. 이후에 이루어진 정제 및 검수 과정을 도식으로 나타내면 다음과 같다.

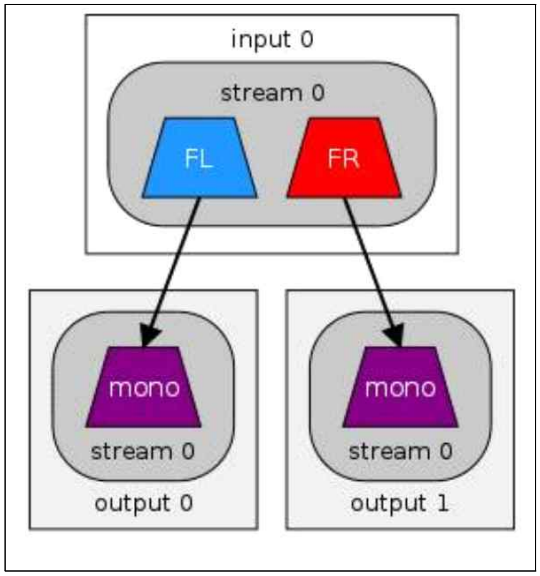


<그림 1> 음성 정제 및 검수 과정

구체적인 음성 정제 작업은 다음과 같은 과정으로 진행되었다.

2.2.1. 음성 파일 형식 변환 (채널 분리)

이 과정은 스테레오 방식으로 저장되어 있는 원본 음성 파일의 채널을 분리하여 2개의 모노 파일로 변환하는 것이다.



<그림 2> 2채널 스테레오 음성 파일의 채널 분리 도식

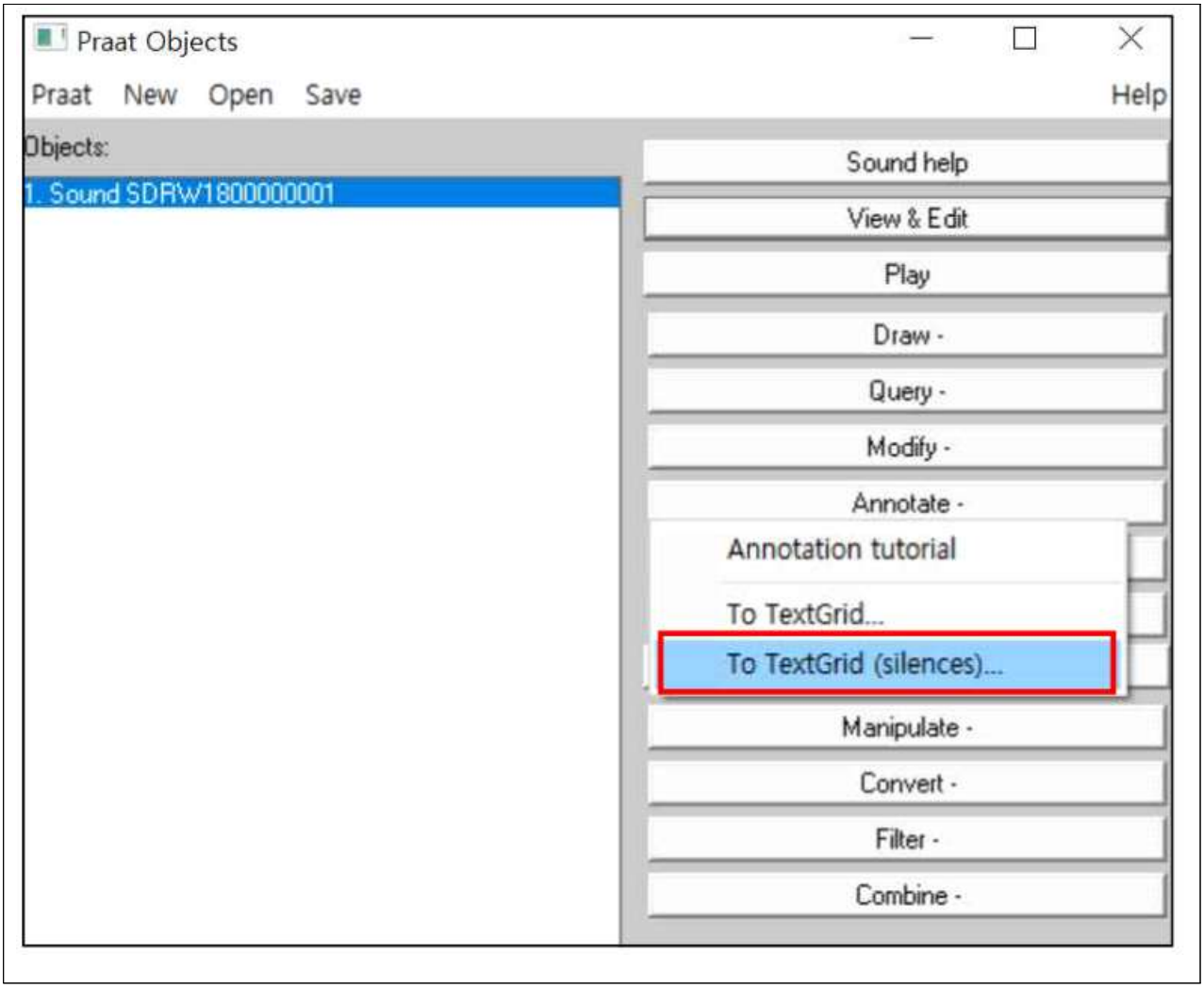
이 작업에는 FFMPEG 소프트웨어의 채널 분절 기능을 사용하였다. 이 과정에서 음성 파일명이 원본 파일에서 바뀌지 않도록 유의하여 진행하였다. 산출된 음성 파일의 형태는 1채널, 샘플링 16kHz, 양자화 16bits wav였다.

2.2.2. 음성 분절

이 과정은 지침에 따라 발화 단위로 음성 파일을 분절하고, TextGrid 파일을 생성하는 것이다. 다음과 같은 절차에 따라 진행하였다.

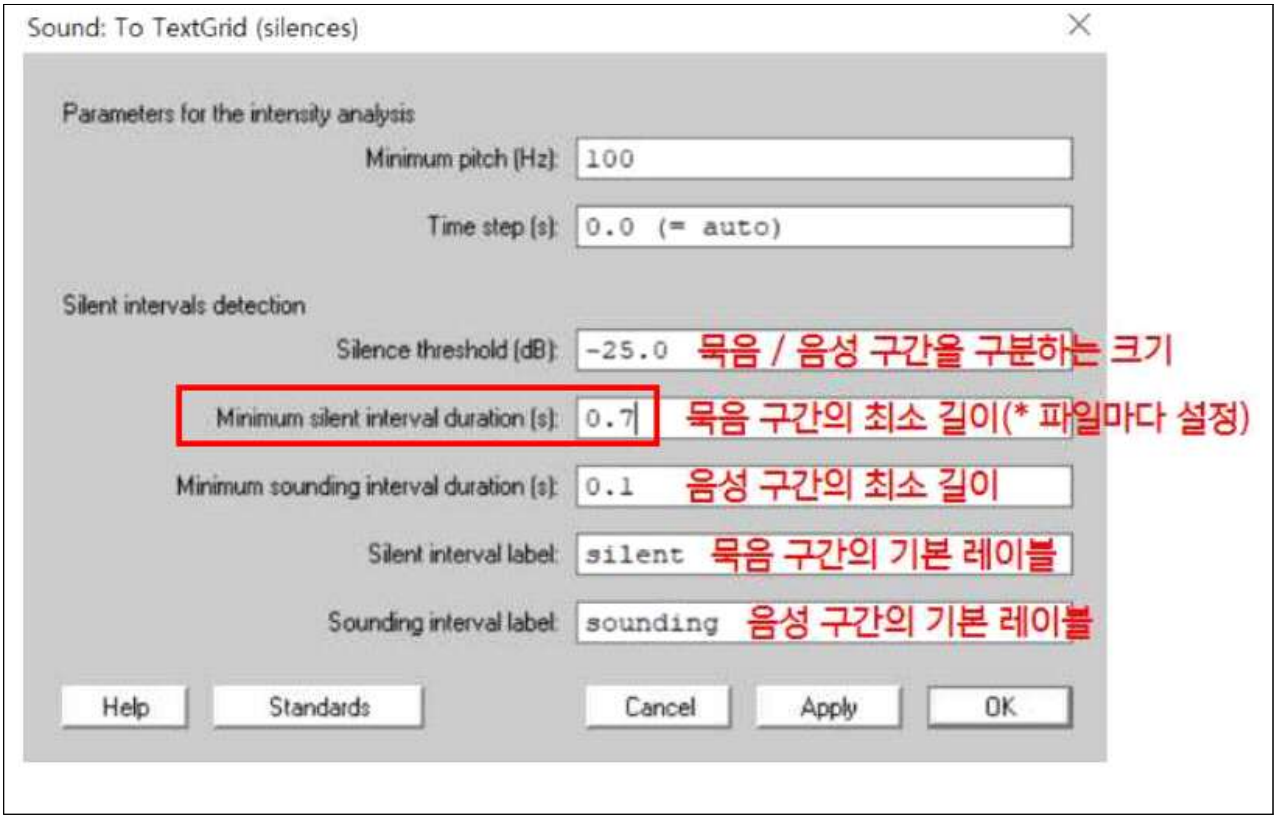
1) TextGrid 생성

praat 소프트웨어를 실행하고, 앞 단계에서 변환한 1채널 음성 파일을 불러와 TextGrid 파일을 생성하였다.



<그림 3> TextGrid 생성

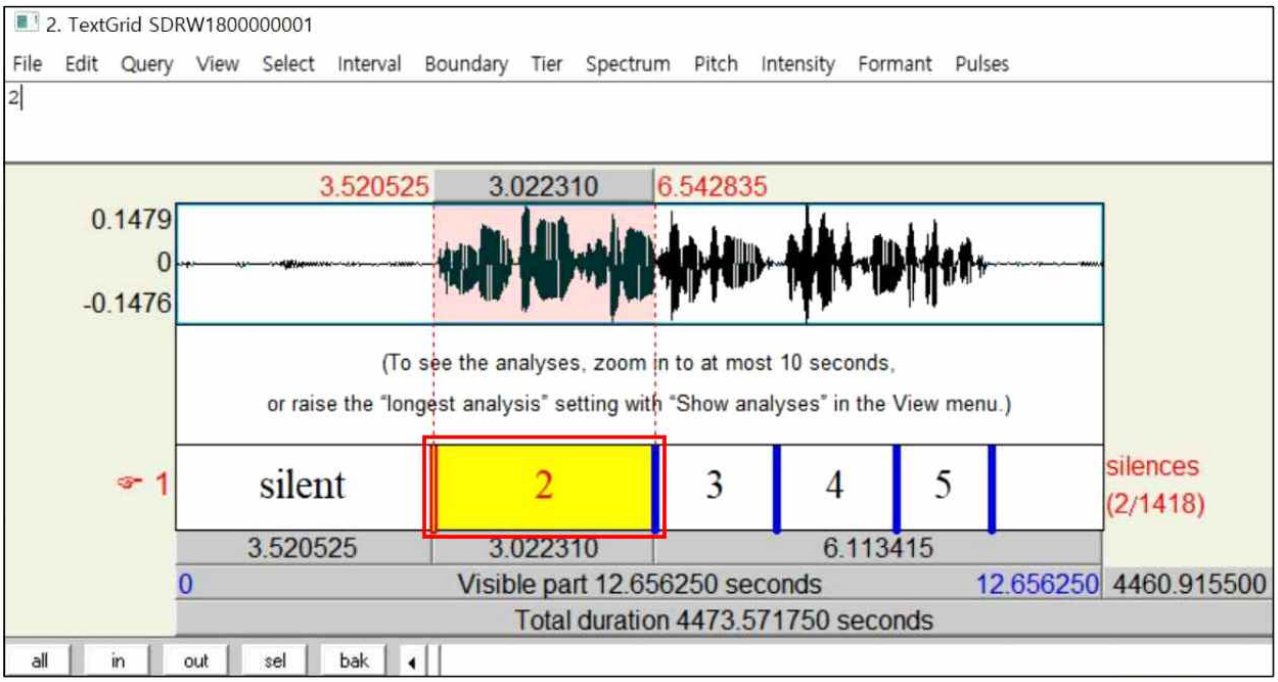
그 다음 minimum silent interval duration(s)를 비롯한 파라미터를 그림과 같이 설정하였다.



<그림 4> 분석 파라미터의 설정

2) 분절 작업

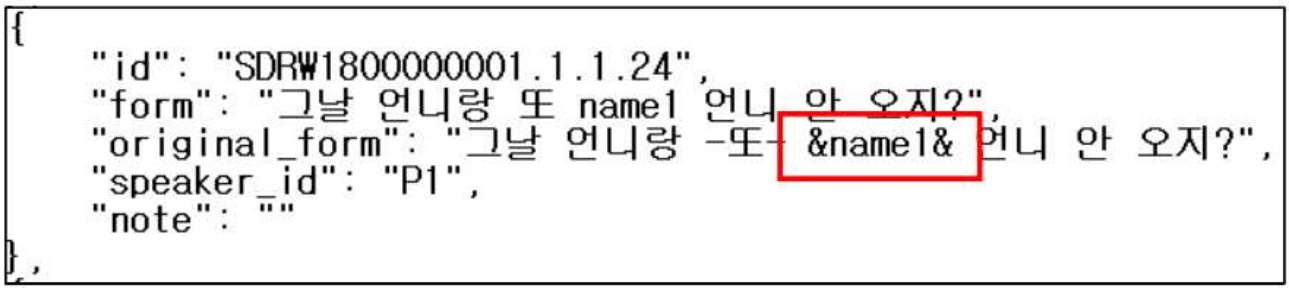
제공된 JSON 전사 말뭉치를 참고하여, 음성을 전사 파일의 전사 단위와 일치하도록 분할하였다. 전사 말뭉치의 발화 번호를 TextGrid에 입력하였다.



<그림 5> praat 환경에서의 음성 분절 작업

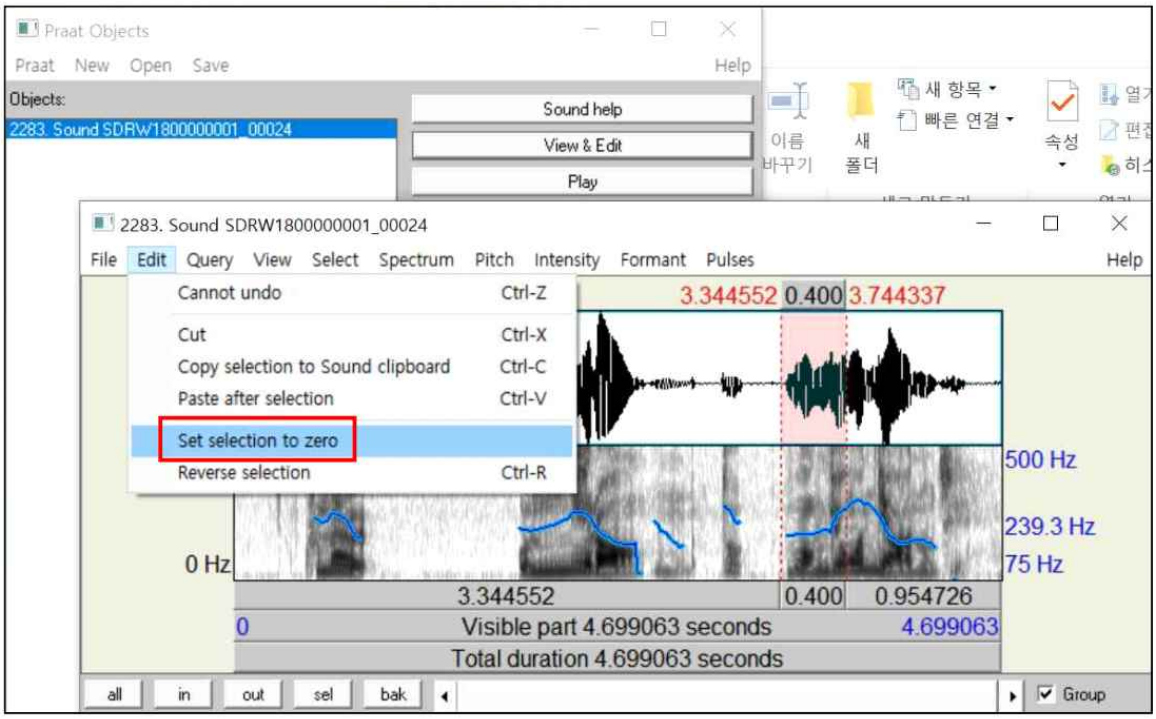
2.2.3. 개인정보 비식별화

이 과정은 발화에서 개인과 관련된 민감한 정보, 즉 이름, 주민등록번호, 신용카드 번호, 전화 번호 등을 음성 정제본에서 묵음으로 처리하는 것이다. 이러한 개인정보는 전사 말뭉치 구축 단계에서 이미 텍스트에 반영되었기 때문에 원칙적으로는 전사 말뭉치의 개인정보 비식별화 기호와 음성 정제본이 대응되도록 묵음 처리를 진행하면 되었다.



<그림 6> 전사 말뭉치의 개인정보 비식별화 처리 방식

묵음 처리는 praat에서 대상 파일을 열고, 묵음 처리 대상 영역을 확인, 선택한 뒤 set selection to zero 명령어로 해당 구간의 음성을 지우는 방식으로 수행하였다.



<그림 7> 정제 음성 파일의 개인정보 묵음 처리

2.2.4. 문제 상황의 처리

전술한 바와 같이 본 과업의 핵심 내용은 이미 음성 자료 수집과 전사가 이루어진 말뭉치에 대하여 음성 정제를 수행하는 것이다. 따라서 원칙적으로는 전사 텍스트를 근거로 하여 전사 단계에서 이루어진 발화 단위 분할과 개인정보 비식별화가 음성 파일에 대응되도록 음성 정제에 반영하면 된다. 그러나 음성 자료 수집 및 전사 단계에서 발생한 오류가 정비되지 않은 채 전사 텍스트에 남아있었기 때문에 음성 정제 과정이 사전에 예상했던 만큼 수월하게 이루어지지는 않았다.

전사 텍스트에 나타난 전사 오류의 유형과 이들을 작업 과정에서 처리한 방법은 다음과 같다. 처리 내역은 별도의 엑셀 파일에 기재하여 제출하였다.

<표 2> 음성 말뭉치 전사 오류의 유형과 처리

구분	내용	처리 방법
발화 단위 분할 구간 오류	구축 지침의 발화 단위 분할 규정을 따르지 않은 경우	과잉분절된 발화 단위를 통합하고 새로이 발화 단위 번호를 부여함
전사 텍스트와 발화 내용이 불일치하는 오류	발화되지 않은 단어가 적히거나 어순이 다르게 전사된 경우	전사에 대응되는 음성이 없거나 식별 불가능한 경우 기호 X로 표시하여 보고함 동시발화가 발생한 경우 '동시발화'라고 표시하여 보고함
발화자 식별 오류	전사 말뭉치에 P1, P2가 잘못 기재된 경우	기호 C로 표시하여 보고함
개인정보 비식별화 오류	비식별화 대상인 단어의 비식별화 처리 누락 또는 비식별화 대상이 아닌 단어의 과잉 비식별화	기호 D로 표시하여 보고함

위 표에 제시된 오류 내용에 대해 자세히 설명하면 다음과 같다.

첫째, 발화 단위 분할 구간 오류이다. 구축 지침에 따르면 하나의 음성 구간 앞, 뒤에는 200msec 이상의 휴지가 포함되어야 하고, 음성 구간 앞, 뒤에 잡음이 포함된 경우에는 가능하면 잡음 외에 200msec 이상의 휴지가 포함되도록 하고 있다. 그런데 전사 텍스트에서는 이 지침에 어긋나게 무리하게 발화 단위를 나누어 과잉 분절한 경우가 있었다.

둘째, 전사 텍스트와 발화 내용이 불일치하는 오류이다. 전사 텍스트에 실제 발화와 다른 단어가 적혀 있거나, 실제 발화와 어순이 다른 경우가 있었다.

셋째, 발화자 식별 오류이다. 2018년도 일상대화 말뭉치는 두 사람(P1, P2)의 대화를 녹음한 자료인데, 발화자 P1과 P2가 전사 텍스트 안에서 뒤바뀌어 기재된 경우가 있었다.

넷째, 개인정보 비식별화 오류이다. 개인정보 비식별화 대상이 되는 정보인데 전사 텍스트에는 비식별화 기호 처리되어 있지 않거나, 거꾸로 개인정보 비식별화 대상이 아닌데 전사 텍스트에는 비식별화 기호 처리된 경우가 있었다.

이와 같이 전사 텍스트의 오류 때문에 발생한 문제 상황에 대해서는 다음과 같이 처리하였다. 발화 단위 분할 구간 오류, 발화자 식별 오류, 개인정보 비식별화 오류는 구어 말뭉치 구축 지침에 맞게 (그러므로, 주어진 전사 텍스트와는 일치되지 않게) 음성 정제 작업을 수행한 뒤, 정제 결과 보고용 엑셀 파일에 발생 위치를 보고하였다.

	A	B	C	D	E	F
1453	SDRW1800000007.1.1.1452	언니 탓 음	P2			1541
1454	SDRW1800000007.1.1.1453	응 그때 고등학교 때 -죽- 교복 없앴다고 난리잖아	P1			
1455	SDRW1800000007.1.1.1454	음	P2			1453
1456	SDRW1800000007.1.1.1455			외부개입으로 대화 중단		X
1457	SDRW1800000007.1.1.1456	네	P2			
1458	SDRW1800000007.1.1.1457	여행 갔다 온 거 많잖아	P1			1456
1459	SDRW1800000007.1.1.1458	저번에 여행 갔을 때 그 골프 투어 갔을 때	P2			
1460	SDRW1800000007.1.1.1459	응	P1			
1461	SDRW1800000007.1.1.1460	한 사람이 이제 회사를 다녀	P2			1459
1462	SDRW1800000007.1.1.1461	어	P1			1459
1463	SDRW1800000007.1.1.1462	신일제약 이사야 근데 그 사람 때문에 좀 날짜가 더 가고 싶은데	P2			
1464	SDRW1800000007.1.1.1463	걔가 회사에서 빡질 -십- 날짜가 얼마 안 되는 거야	P2			
1465	SDRW1800000007.1.1.1464	그게 이제 다른 한 사람이 볼만인 거야 그 사람 위주로 가다 보니까	P2			1463
1466	SDRW1800000007.1.1.1465	비용도 더 많이 들고 그리고 시간은 짧고	P2			1463
1467	SDRW1800000007.1.1.1466	응 그렇지	P1			
1468	SDRW1800000007.1.1.1467	그러니까 내가 갈 때마다 항상 이제 그 언니가 내 뭐라고 해	P2			1466
1469	SDRW1800000007.1.1.1468	응	P1			1466
1470	SDRW1800000007.1.1.1469	근데 얘는 이제 어리니까 말을 못 해	P2			
1471	SDRW1800000007.1.1.1470	어	P1			1469
1472	SDRW1800000007.1.1.1471	그리고 자기 맴에 비싸게 가고 날짜를 줄인다고 하니까	P2			
1473	SDRW1800000007.1.1.1472	이후 미안하다고 하는데도 너무 앞에다 대고 막 그 언니가 또 직선 적이야 또	P2			
1474	SDRW1800000007.1.1.1473	어디 가면 직선적인 사람 있잖아	P2			
1475	SDRW1800000007.1.1.1474	에잇 필요하기는 해	P1			
1476	SDRW1800000007.1.1.1475	어 그래도 이제 우리가 해야 될 말을 그 언니가 다 해	P2			
1477	SDRW1800000007.1.1.1476	그런데 이제 얘가 이제 한번 가끔 가다가 주눅 들어 있어	P2			1475
1478	SDRW1800000007.1.1.1477	이제 뭐라고 이제 자꾸 하니까 너 때문에 비싸게 가고	P2			1475

<그림 8> 정제 결과 보고 엑셀 파일 화면

2.3. 검수 및 피드백

2.3.1. 내부 검수

2019년도 일상 대화 말뭉치 구축 사업에 참여한 경험이 있는 구성원이 주축이 되어, 정제 일정과 동시에 검수를 진행하였다. 모든 정제 음성 파일에 대하여 2회 이상의 검수가 이루어졌다. 검수 과정에서 오류가 발견된 경우, 소규모의 오류는 검수자가 즉시 수정하고, 대규모의 오류는 작업 담당자에게 반환하여 오류를 수정하도록 하였다.

2.3.2. 작업결과물 제출 및 피드백 교환

4회에 걸쳐 음성 정제 결과물 파일을 제출하고, 국립국어원의 검수 의견을 수령하였다. 정제 오류에 대해서는 파일 수정을 진행하고, 질의점에 대해서는 답변을 전달하였다.

3. 서울말 낭독체 말뭉치의 통합과 정비

3.1. 과업 수행 준비

본 과업은 서울말 낭독체 말뭉치 대본 텍스트 19종과 그 낭독 음성 파일 약 150시간

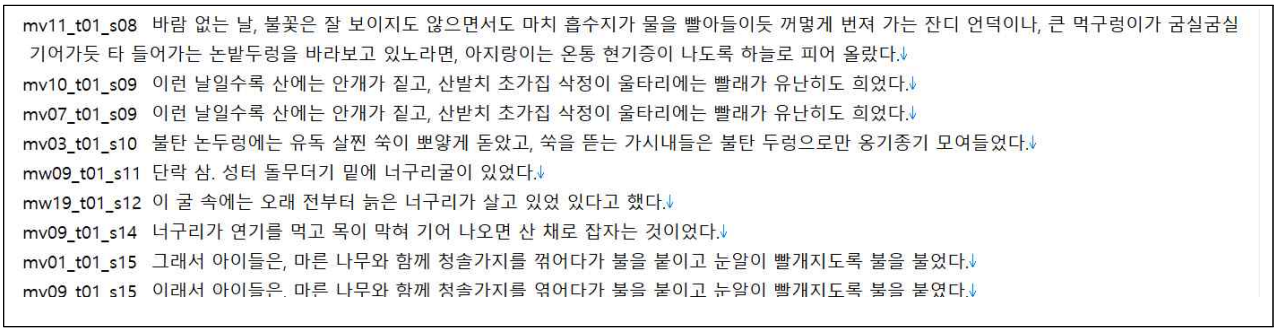
44분 분량에 대하여, 대본과 낭독간의 불일치를 확인하고 불일치를 반영한 텍스트를 작성하는 과업이다. 이를 위하여 국립국어원으로부터 낭독 대본과 낭독 음성 wav 파일을 수령하였다. 낭독 음성 파일은 발화 단위별로 분절되어 있는 것으로서, 이 발화 단위는 대체로 문장 단위와 일치하게끔 대본 단계에서 결정되어 있다. 이 과업은 2018년도 일상 대화 말뭉치 과업과는 달리 녹음 음성을 들으며 대본과의 일치 여부를 확인하는 작업이기 때문에 별도의 음성 처리용 소프트웨어를 사용하지는 않았다.

개별 작업자에게 작업 대상을 분배하기 전, 자료 파일 검토 단계에서 대본의 발화 단위의 수와 음성 파일의 수가 불일치하는 경우가 20여 건 발견되었다. 이는 단락 단위 녹음 파일을 분할하는 과정에서 생긴 오류로 확인되어, 국립국어원에 단락 파일을 요구하여 파일의 재분절을 수행하였다.

3.2. 전사 정비

3.2.1. 대본-음성간 불일치 확인

자료 파일을 검토하고 문제 파일의 재분절을 수행한 뒤, 개별 구성원에게 작업 담당 파일을 분배하여 작업을 수행하였다. 낭독 녹음 내용의 대부분은 대본과 일치하므로, 대본과 다르게 낭독한 발화 단위만 전사하여 기록하는 방식으로 진행하였다.


mv11_t01_s08 바람 없는 날, 불꽃은 잘 보이지도 않으면서도 마치 흡수지가 물을 빨아들이듯 꺼멓게 번져 가는 잔디 언덕이나, 큰 먹구렁이가 굼실굼실 기어가듯 타 들어가는 논밭두렁을 바라보고 있노라면, 아지랑이는 온통 현기증이 나도록 하늘로 피어 올랐다.↓
mv10_t01_s09 이런 날일수록 산에는 안개가 짙고, 산발치 초가집 삭정이 울타리에는 빨래가 유난히도 희었다.↓
mv07_t01_s09 이런 날일수록 산에는 안개가 짙고, 산발치 초가집 삭정이 울타리에는 빨래가 유난히도 희었다.↓
mv03_t01_s10 불탄 논두렁에는 유독 살찐 쑥이 뽀얗게 돋았고, 쑥을 뜯는 가시내들은 불탄 두렁으로만 옹기종기 모여들었다.↓
mw09_t01_s11 단락 삼. 성터 돌무더기 밑에 너구리굴이 있었다.↓
mw19_t01_s12 이 굴 속에는 오래 전부터 늙은 너구리가 살고 있었 있다고 했다.↓
mv09_t01_s14 너구리가 연기를 먹고 목이 막혀 기어 나오면 산 채로 잡자는 것이었다.↓
mv01_t01_s15 그래서 아이들은, 마른 나무와 함께 청솔가지를 꺾어다가 불을 붙이고 눈알이 빨개지도록 불을 불었다.↓
mv09_t01_s15 이래서 아이들은, 마른 나무와 함께 청솔가지를 엮어다가 불을 붙이고 눈알이 빨개지도록 불을 붙였다.↓

<그림 9> 대본-음성간 불일치 기록

정비 과정에서 확인된 대본-음성간 불일치 및 녹음 오류 정보를 간략히 요약하면 다음과 같다.

<표 3> 서울말 낭독체 음성파일 현황

대본 일치 여부	음성 파일 상태	문장수
일치	정상	84,283
불일치	정상	2,737
불일치	나쁨	15
누락	누락	113

전사의 검토에 병행하여 대본의 정비도 수행하였다. 명백한 오자의 수정, 인용 부호

및 각종 구두점 문자의 통일, 오류로 삽입된 특수문자 및 공백의 제거 등이 이에 해당된다.

3.2.2. 문제 상황의 처리

과업 수행 준비 단계에서 음성 파일 분절 오류를 발견하고 수정하기는 하였으나, 그 외에도 많은 음성 파일 오류가 있음을 전사 정비 단계에서 확인하였다. 음성 파일의 분절에 오류가 있는 경우는 ErrorFileSegmentation 태그를, 음질에 문제가 있는 경우는 ErrorSoundQuality 태그를 붙이며 작업을 진행하고, 이후 오류 태그가 붙어 있는 발화 단위들을 모아 일괄 검토하여 문제 상황을 확인하였다.

분절 오류는 대체로 두 가지 유형이 있었다. 하나는 분절되어야 하는 지점에서 분절되지 않은 채, 동일한 덩어리 음성이 두 파일에 중복되어 있는 경우이다.

대본:
t10_s10 "나무꾼 아저씨!"
t10_s11 "살려 주세요."
분절오류 음성 분절:
fx01_t10_s10 나무꾼 아저씨 살려 주세요.
fx01_t10_s11 나무꾼 아저씨 살려 주세요.

<그림 10> 분절 오류의 예 (1)

또 하나는 분절해야 하는 지점을 비껴 정확하지 않은 위치에서 분절한 경우이다.

대본:
t10_s11"살려 주세요."
t10_s12"저는 사냥꾼에게 쫓기고 있어요.“
분절오류 음성 분절:
mw19_t10_s11 살려 주세요. 저는
mw19_t10_s12 사냥꾼에게 쫓기고 있어요.

<그림 11> 분절 오류의 예 (2)

아울러 전사 정비 단계에서 음질 문제가 있는 것으로 파악된 파일 가운데, 단락 파일의 해당 부분을 확인하였을 때는 음질에 문제가 없는 경우가 있어, 음질 문제 중에서도 녹음 단계가 아니라 분절 단계에서 이상이 발생한 경우가 있음을 확인하였다.

전체 음성 파일 중 305개 파일에서 이러한 오류가 있음을 확인하고, 국립국어원으로부터 단락 단위 녹음 파일을 수령하여 파일의 재분절을 수행하였다.

3.3. 검수 및 피드백

3.3.1. 내부 검수

전사 정비 일정과 동시에 검수를 진행하였다. 모든 정비 전사 파일에 대하여 2회 이상의 검수가 이루어졌다. 검수 과정에서 오류가 발견된 경우 작업 담당자에게 반환하여 오류를 수정하도록 하였다.

3.3.2. 작업결과물 제출 및 피드백 교환

4회에 걸쳐 전사 정비 결과물 파일을 제출하였다. 전사 정비 과정에서 국립국어원의 답변 또는 승인이 필요한 사항에 대해 수시로 질의를 전달하고 답변을 확인하였다.

제 3 장

7개 층위 분석 말뭉치의 통합과 정비

1. 7개 층위 분석 말뭉치 통합 및 정비의 대상과 범위

7개 층위 분석 말뭉치 통합 및 정비의 대상이 되는 데이터는 2019년에 구축된 분석 말뭉치로, 구체적으로는 ① 형태 분석, ② 어휘 의미, ③ 개체명, ④ 상호참조 해결, ⑤ 구문 분석, ⑥ 의미역, ⑦ 주격 무형 대용어 복원 말뭉치이다. 본 과업에서는 이들 말뭉치에 대해 층위 통합과 내용 검증을 수행하고, 그 과정에서 확인된 오류 유형을 보고하고 말뭉치 구축 지침의 보완사항을 제안하게 된다. 아울러 각 말뭉치의 구축 사업단이 구축 사업의 후속 조치로서 산출한 '유지 보수 말뭉치'와 연계한 검증 작업도 수행한다.

과업 대상 말뭉치의 규모는 다음과 같다. 형태 분석, 어휘 의미, 개체명, 상호참조 해결, 주격 무형 대용어 복원 말뭉치는 문어 분석 말뭉치, 구어 분석 말뭉치를 모두 대상으로 하나 구문 분석 말뭉치와 의미역 말뭉치는 문어 분석 말뭉치만을 대상으로 하였다.

<표 4> 과업 대상 분석 말뭉치의 규모

분석층위	검증 대상 분량(어절)	비고
형태 분석	300만	
어휘 의미	300만	
개체명	300만	
상호참조 해결	300만	
구문 분석	200만	구어 말뭉치 제외
의미역	200만	구어 말뭉치 제외
주격 무형 대용어 복원	300만	
합계	**1900만**	

7개 층위 분석 말뭉치의 명칭 및 문어 데이터와 구어 데이터의 구별에는 로마자 약호를 사용하기도 한다. 7개 층위 분석 말뭉치를 가리키는 데에 쓰는 약호는 다음과 같다.

<표 5> 7개 층위 분석 말뭉치 명칭 약호

말뭉치	약호
형태 분석 말뭉치	MP
어휘 의미 말뭉치	LS
개체명 말뭉치	NE
상호참조 해결 말뭉치	CR
구문 분석 말뭉치	DP
의미역 말뭉치	SR
주격 무형 대용어 복원 말뭉치	ZA

이에 더하여 문어(신문)을 나타내는 약호로 NX, 구어를 나타내는 약호로 SX가 쓰인다. 예를 들어 NXMP는 문어(신문) 형태 분석 말뭉치, SXMP는 구어 형태 분석 말뭉치를 가리킨다.

향후 여러 층위의 분석 말뭉치가 연계되어 산업과 연구에 응용되기 위해서는, 모든 층위의 분석 말뭉치가 동일한 JSON 형식을 갖추고 있어야 할 것이다. 그러나 기존의 분석 말뭉치는 층위간에 JSON 형식이 일치하지 않고, 한 말뭉치 안에서도 일관성이 없거나, 아예 구축 과정에서 발생한 오류가 남아있는 경우가 있는 것으로 파악된다. 따라서 층위간 JSON 형식의 불일치 및 오류를 바로잡을 필요가 있다. 이 작업이 본 사업에서 말하는 층위 통합이다.

층위 통합이 이루어진 뒤에는 말뭉치의 층위별 주석 내용의 전문가 검증과 수정을 수행한다. 전문가 검증과 수정의 범위는 전체 말뭉치의 약 20% 분량으로, 국립국어원에서 사업 착수 단계에 지정한 문서들을 대상으로 한다. 단, 상호참조 해결 층위는 예외적으로 '2019년 말뭉치 통합 검증' 사업에서 구축한 검증 말뭉치와 분석 말뭉치의 불일치분에 대해 검증 및 수정을 수행한다.[1] 이 과정에서 검증에 참여한 전문가들이 발견한 오류 유형을 기록, 보고하며 말뭉치 구축 지침의 보완사항을 제안한다.

층위 통합 및 주석 검증과 동시에, 각 말뭉치의 구축 사업단이 구축 사업의 후속 조치로서 산출한 '유지 보수 말뭉치'와 연계한 검증을 진행한다. 그 절차는 다음과 같다. 층위 통합이 이루어진 말뭉치를 각 구축 사업단에 발송한 뒤, 구축 사업단이 본 과업과는 별도로 외부에서 유지 보수 작업을 수행한 말뭉치를 이후 수령한다. 수령한 유지 보수 말뭉치의 수정 전후를 비교 검증하고, 본 과업의 결과물과 유지 보수 말뭉치의 내용

1) 상호참조 해결 말뭉치가 타 말뭉치와 검증 대상을 달리하게 된 것은 상호참조 해결 말뭉치 고유의 특성 때문이다. 상호참조 해결 말뭉치는 한 문서 내에서 명사구들의 동일 지시 관계를 주석하는 말뭉치이기 때문에, 하나의 주석이 올바른지 판단하기 위해 문서 전체의 맥락을 검토해야 하며, 결과적으로 주석의 적절성을 판단하는 데에 말뭉치를 신규 구축하는 것에 못지않은 노력과 시간이 소모된다. 본 사업에서는 사업 규모와 일정을 고려했을 때 상호참조 해결 층위의 경우 구축 말뭉치와 통합 검증 말뭉치의 주석 차이를 검토하여 말뭉치를 수정하는 것이 현실적인 검증 방법이라고 판단하여 이와 같이 나머지 층위와 다르게 검증을 진행하였다. 이상의 사항은 모두 국립국어원과의 긴밀한 업무 협의에 의해 결정되었다.

을 비교 검증한다.

2. 분석 말뭉치 층위 통합

2.1. 과업 수행 준비

과업의 개시와 함께 말뭉치 수정 이력 관리 및 작업 담당자간 진행상황 공유를 위해 github 페이지를 개설하였다. 이 페이지는 분석 말뭉치 층위 통합뿐 아니라 이후의 분석 말뭉치 과업에도 지속적으로 활용하였다. 주석 전문가 검증의 수행을 대비하여 JSON 형식 파일과 tsv 형식 파일의 상호 변환 작업도 준비하였다.

2.2. 파일 검사 및 JSON schema 검증

말뭉치 파일 형식의 문제와 각 말뭉치가 국립국어원 JSON schema에 부합하는지의 검사를 진행하였다. 검사 결과와 조치 내용은 다음과 같다.

1) 들여쓰기 형식의 말뭉치간 불일치

문어 형태 분석 말뭉치에서는 JSON dump 출력의 들여쓰기를 스페이스 4개로 한 반면, 문어 어휘의미 말뭉치에서는 들여쓰기를 탭 1개로 하여 말뭉치간 형식이 일치하지 않는다는 것이 발견되었다. 이 때문에 문어 형태 분석 말뭉치는 실질적인 말뭉치 정보에 비해 3배 이상의 파일 크기를 갖게 되었다.

이상의 사항을 보고하고, 국립국어원의 업무 지시에 따라 들여쓰기를 스페이스 4개로 통일하였다.

2) 줄바꿈 문자 형식 불일치

문어 형태 분석 말뭉치는 줄바꿈 문자가 LF인 반면, 문어 어휘의미 말뭉치에서는 줄바꿈 문자가 CR-LF로 되어 있는 등 말뭉치간 형식이 일치하지 않는다는 것이 발견되었다.

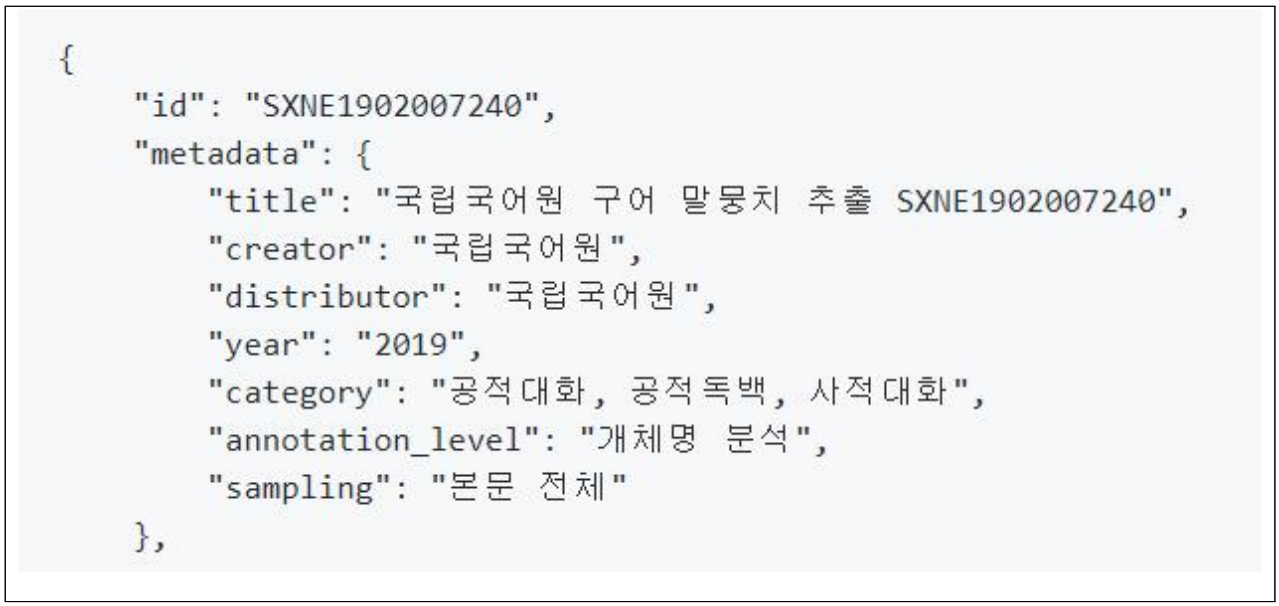
이상의 사항을 보고하고, 국립국어원의 업무 지시에 따라 줄바꿈 문자를 LF로 통일하였다.

3) 데이터 유형 불일치 문제

구어 형태 분석 말뭉치, 구어 어휘 의미 말뭉치, 구어 개체명 말뭉치에서 문장형태(sentence form)가 빈 문자열인 경우에 어절(word), 형태소(morpheme)에 None 값이 부

여되어 있었다. 그런데 국립국어원의 JSON 형식 규정에 의하면 word 및 morpheme은 array값을 가지므로, 이와 같은 처리는 오류이다. 반면 form은 문자열(string) 타입이므로 ''을 가지는 것이 오류가 아니다.

한편, 국립국어원의 JSON 형식 규정에 의하면 말뭉치 내 메타데이터 중 층위명(annotation_level)은 배열(array) 값을 가진다. 그런데 아래와 같이 층위명이 문자열로 입력된 말뭉치가 있었다.



<그림 12> 메타데이터 JSON 형식 오류

이상의 사항을 보고하고, 국립국어원의 업무 지시에 따라 word, morpheme 등이 None으로 된 것을 []으로, annotaton_level을 ["개체명 분석"] 등으로 바로잡았다.

4) 그 외의 문제

그 외의 다양한 유형의 형식적 오류들, 예를 들어 word를 words로 잘못 표기한 오류, NE_id를 ne_id로 잘못 표기한 오류 등을 바로잡았다.

이 단계까지의 수정내역과 수정량을 표로 나타내면 다음과 같다. (단, 말뭉치 파일 전체에 적용되는 수정사항인 들여쓰기 형식의 말뭉치간 불일치 해결, 줄바꿈 문자 형식 불일치 해결은 표에서 제외)

<표 6> 파일 검사 및 JSON schema 검증의 수정내역 및 수정량

말뭉치	변경 전	변경 후	수정량(건)
NXMP	"metadata": {},	"annotation_level": [], "category": "", "creator": "", "distributor": "", "title": "", "year": "" "metadata": { },	1
	"metadata": {},	"metadata": { "author": "", "date": "", "publisher": "" "title": "", },	7,265
SXMP	"metadata": {},	"metadata": { "annotation_level": [], "category": "", "creator": "", "distributor": "", "title": "", "year": "" },	1
	"metadata": {},	"metadata": { "author": "", "date": "", "publisher": "" "title": "", },	423
	"morpheme": null,	"morpheme": [],	2,473
	"word": null,	"word": [],	2,473
NXLS	"metadata": {},	"annotation_level": [], "category": "", "creator": "", "distributor": "", "title": "", "year": "" "metadata": { },	1
	"metadata": {},	"metadata": { "author": "", "date": "", "publisher": ""	7,265

		"title": "", },	
SXLS	"metadata": {},	"metadata": { "annotation_level": [], "category": "", "creator": "", "distributor": "", "title": "", "year": "" },	1
	"metadata": {},	"metadata": { "author": "", "date": "", "publisher": "" "title": "", },	423
	"morpheme": null,	"morpheme": [],	2,473
	"word": null,	"word": [],	2,473
NXNE	"annotation_level": "개체명 분석",	"annotation_level": ["개체명 분석"],	1
SXNE	"annotation_level": "개체명 분석",	"annotation_level": ["개체명 분석"],	1
NXCR	"ne_id": -1	"NE_id": -1	309,355
	"words": [	"word": [	150,082
	"topic": null,	"topic": "",	7,265
	"url": null	"url": ""	7,265
	"ZA": null	"ZA": []	218
	"ZA": null,	"ZA": [],	7,047
	"annotation_level": "상호참조 해결",	"annotation_level": ["상호참조 해결"],	1
SXCR	"ne_id": -1	"NE_id": -1	82,816
	"words": [	"word": [	221,489
	"words": []	"word": []	2,473
	"date": null,	"date": "",	423
	"topic": null,	"topic": "",	423
	"url": null	"url": ""	423
	"ZA": null	"ZA": []	423
	"annotation_level": "상호참조 해결",	"annotation_level": ["상호참조 해결"],	1
NXSR	"metadata": {},	"annotation_level": [], "category": "", "creator": "", "distributor": "", "title": "",	1

		"year": "" "metadata": { },	
	"metadata": {},	"metadata": { "author": "", "date": "", "publisher": "" "title": "", },	7,265
NXZA	"annotation_level": "무형 대용어 복원",	"annotation_level": ["무형 대용어 복원"],	1
SXZA	"annotation_level": "무형 대용어 복원",	"annotation_level": ["무형 대용어 복원"],	1

2.3. 문서, 문장, 어절 정보 검증

말뭉치의 문서, 문장, 어절의 일관성을 검증하였다. 검사 결과와 조치 내용은 다음과 같다.

1) 문서 검사 결과와 조치 내용

각 말뭉치에 포함된 문서들의 일련번호가 모두 일치하는 것을 확인하였다. 그러나 문서의 순서는 일관되지 않았다. 즉, 말뭉치 내 문서들이 문서 번호 순서대로 정렬되어 있지 않았다. 이러한 상태는 데이터 관리 측면에서 좋지 않고, 말뭉치 사용자 입장에서도 문서의 순서가 동일할 것으로 기대할 것이므로, 문서의 순서를 통일할 필요가 있다.

이상의 사항을 보고하고, 국립국어원의 업무 지시에 따라 문서의 순서를 문서 고유번호에 따라 정렬하였다.

2) 문장 검사 결과와 조치 내용

구어 상호참조 해결, 개체명, 주격 무형 대용어 말뭉치에서 문장 형태의 시작에 공백 문자가 삽입된 경우가 발견되었다. 상호참조 해결과 주격 무형 대용어 말뭉치에서는 띄어쓰기 스페이스의 중복 오류도 발견되었다. 그 외에 탭 문자 삽입, newline 삽입 등의 오류 가능성도 검토하였고 오류가 발견되지 않았다.

문장 형태가 빈 문자열인 경우가 확인되었는데, 내용을 검토한 결과 구어를 전사하는 과정에서 특수한 전사 약호로만 이루어진 문장이 기록되었고, 원시 말뭉치에는 내용이 있던 것이 분석 말뭉치로 넘어오면서 빈 문장이 된 것으로 추정되었다. 이는 보기에 따

라서는 오류가 아니라고 판단할 수도 있다. 그러나 빈 문장을 그대로 유지한다고 하더라도 말뭉치 설명에 이에 대한 명시적인 언급이 있어야 할 것이다. 배포된 말뭉치의 사용자에게 예상하지 못한 문제를 일으킬 수 있기 때문이다. 빈 문장 자체를 데이터에서 제외하는 것도 수정 방법의 하나이다. 빈 문장은 이후의 주석이나 분석에 사용될 일이 없기 때문이다.

문서 검사 결과와 문장 검사 결과를 표로 나타내면 다음과 같다. 문어에서는 이와 같은 오류가 발견되지 않았으므로 표 제시를 생략한다.

<표 7> 구어 말뭉치 문서 검사 및 문장 검사 결과

오류 유형	SXMP	SXLS	SXNE	SXCR	SXZA
문장 형태 내 공백문자	0	0	852	2,871	2,871
문장 형태 내 연속된 두 개의 스페이스 문자	0	0	0	769	769
빈 문자열	2,473	2,473	2,473	2,473	2,473

이상의 사항을 보고하고, 국립국어원의 업무 지시에 따라 문장 형태 오류를 바로잡고, 빈 문자열로 이루어진 문장은 해당 문장 데이터를 말뭉치에서 제외하였다.

3) 어절 검사 결과와 조치 내용

어절 형태 내에 공백문자가 있는 오류, 어절의 어절 begin, end 값의 오류, 어절 번호(word id)가 일관성 있게 부여되지 않은 오류, 문장 형태 안의 해당 어절이 어절 형태(word form)와 일치하지 않는 오류가 발견되었다. 검사 결과를 표로 나타내면 다음과 같다.

<표 8> 문어 말뭉치 어절 검사 결과

오류 유형	NXMP	NXLS	NXNE	NXCR	NXZA	NXDP	NXSR
어절 형태 내 공백문자	0	0	0	0	0	0	2
Begin, End 값 오류	0	0	0	0	0	0	128
어절 번호의 비일관성	0	0	1761	0	0	0	0
문장 형태 어절 순서와 어절 형태의 불일치	0	0	0	0	0	26	40

<표 9> 구어 말뭉치 어절 검사 결과

오류 유형	SXMP	SXLS	SXNE	SXCR	SXZA
어절 형태 내 공백문자	0	0	0	0	0
Begin, End 값 오류	0	0	0	0	0
어절 번호의 비일관성	0	0	3842	0	0
문장 형태 어절 순서와 어절 형태의 불일치	0	0	0	0	0

이상의 사항을 보고하고, 국립국어원의 업무 지시에 따라, 어절 번호의 일관성 문제는 어절 번호를 새롭게 부여하는 것으로 해결하고, 어절 형태의 문제는 원시 말뭉치 기준으로 통일하여 바로잡았다. 그에 따르는 begin, end 값 수정도 수행하였다.

4) 문장 번호, 문장 형태 차원에서의 층위 통합

원시 말뭉치와 분석 말뭉치의 문장 번호 짝을 맞추고 말뭉치 간 불일치를 해소하였다. 문어 원시 말뭉치는 문장 번호가 없으며 단락 번호만 존재한다. 원시-분석 말뭉치간 단락 번호 일치를 검증하여 오류가 없음을 확인하였다. 구어 원시 말뭉치는 발화 번호가 있다. 원시 말뭉치의 발화 번호와 분석 말뭉치의 문장 번호 일치를 검증하여 상당수의 오류를 발견하고 불일치를 해소하였다. 전술한 바와 같이 구어 원시 말뭉치에서 문장의 내용이 없는 경우 분석 말뭉치에서는 제거하였다. 이를 통해 문어 분석 말뭉치 7개 층위, 구어 분석 말뭉치 5개 층위에 대해 문장 번호 수준에서 층위간 통합을 수행하였다. 상호참조 해결 말뭉치와 주격 무형 대용어 복원 말뭉치에서는 주석 내용에서 문장 번호가 참조되므로, 문장 번호가 수정되는 경우 해당 문장이 포함된 문서 전체에 대해 주석의 내용 검증과 오류 수정을 수행하였다.

<표 10> 구어 말뭉치 층위별 문장 번호 수정량

말뭉치	수정량
SXMP	103,122
SXLS	103,122
SXNE	103,118
SXCR	140,439
SXZA	240,975

다음으로 원시 말뭉치와 분석 말뭉치 사이의 간 문장 형태(sentence form) 불일치를 검증하였다. 모든 분석 말뭉치의 문장 형태를 원시 말뭉치를 기준으로 수정하였다. 형태 분석, 어휘 의미, 개체명, 구문 분석, 의미역 층위는 문장 단위로 주석이 이루어졌으므로 수정된 문장에 대해 모든 주석을 검토하고 수정하였다. 주격 무형 대용어 복원, 상호참조 층위는 문서 단위로 주석하였으므로 수정된 문장을 포함하는 모든 문서의 주석 가운데 수정된 문장 번호를 참조하는 모든 주석을 검토하고 수정하였다.

<표 11> 말뭉치 층위별 문장 형태 수정량

말뭉치	수정량
NXMP	50
SXMP	255
NXLS	58
SXLS	303
NXNE	14
SXNE	7,989
NXCR	14
SXCR	15,882
NXDP	15
NXSR	24
NXZA	18
SXZA	18,699

이상의 검사가 완료되고 오류가 수정된 상태의 말뭉치를 '층위 통합 말뭉치'라 하고, 구축 사업단에 발송하여 말뭉치 유지 보수 작업이 이루어질 수 있도록 하였다.

3. 분석 말뭉치 주석 검증

3.1. 과업 수행 준비

분석 말뭉치 층위 통합 작업이 이루어짐과 동시에, 말뭉치 구축 지침 숙지 교육과 예비적 주석 검증이 수행되었다. 예비적 주석 검증은 본격적인 주석 검증에 앞서 전문가들이 말뭉치의 오류 유형 파악에 착수할 수 있도록 하는 목적이 있었으며, 최적의 과업 수행 환경을 탐색하는 과정이기도 하였다. 본 과업에서는 일률적인 과업 수행 환경을 제공하기보다는 전문가의 경험과 요구에 맞추어 지속적으로 과업 수행 환경을 변경, 개선해 나가는 방식을 택하였기 때문에, 시행착오와 개선의 기회가 많을수록 좋다고 판단하였다.

3.2. 과업의 진행

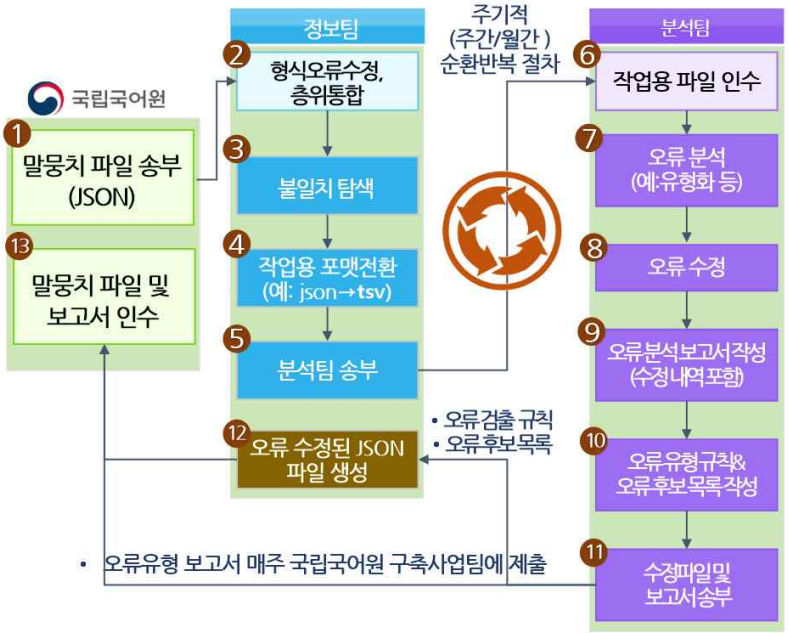
분석 말뭉치 주석 검증은 전체의 20% 분량에 해당되는 지정 문서(상호참조 해결 말뭉치는 예외적으로 검증 말뭉치와의 불일치분)에 대해 주석의 검증과 오류 수정을 수행하는 과업이다. 이 과업에 JSON 형식 파일 그대로를 사용하는 것은 불편과 위험이 큰 일이기 때문에, 작업용 형식으로 변환하여 작업을 수행하는 전문가들에게 제공할 필요가 있었다. 본 사업에서는 JSON 형식 파일을 tsv 형식 파일로 변환하여 전문가들이 각자의 작업 환경에 맞게 텍스트 에디터 등을 이용하여 작업을 수행할 수 있도록 하였다.

본 사업에서는 일률적인 작업 환경을 정하여 사업 내내 사용하는 것이 아니라, 작업 과정에서 지속적으로 전문가들의 피드백을 받아들여 tsv 파일의 양식을 개선하거나, 특수 작업용 파일을 별도로 제공하거나 하였다. 이는 최적의 작업 환경을 과업에 착수하기 전에 미리 알 수도 없었거니와, 효율적인 작업 환경에 대한 전문가들의 아이디어가 과업 중간에도 계속 제시될 수 있다고 내다보았기 때문이다. 결과적으로 층위별로 각기 다른 검증 환경을 제공하게 되었다.

전문가들이 작업 데이터를 분담하여 각자 고립된 환경에서 작업을 수행하는 것이 아니라, 모두가 github 페이지에 작업 현황을 지속적으로 업데이트하여 필요한 경우 누구나 다른 전문가의 진척 상황과 작업 내용을 열람할 수 있도록 하였다. 상시적으로 오류 유형 및 지침 이해에 대한 토의를 상시적으로 협업 지원 도구(JANDI)에 개설된 대화방에서 진행할 수 있었다.

<그림 13> 작업 현황 공유 화면

과업 수행 준비와 진행 과정을 도식으로 나타내면 다음과 같다.



<그림 14> 7개 층위 분석 말뭉치 검증 과업 진행 과정

전문가 주석 검증 과정에서 보편적으로 견지한 원칙은 다음과 같다.

첫째, 지정된 검증 대상에 대한 전수 검증을 수행하였다.

둘째, 빈발하는 오류 유형에 대해서는, 전문가의 제보를 받아들여 동일 유형 오류 후보를 탐색하고, 오류 태그를 부착해 정렬한 뒤 전문가 검증을 위해 제공하는 등 효율적인 과업 수행 방법을 택하였다.

셋째, 한 층위의 주석 정보를 다른 층위의 오류 유형 탐색에 활용하기도 하였다.

넷째, 복수의 팀원이 여러 층위의 지침을 숙지하여 공동 작업하였다.

이상이 검증의 기본 원칙이었으나, 층위별로 말뭉치 특성이 주어진 조건에 따라 검증 환경을 달리하며 과업을 진행하였다. 층위별로 다르게 제공된 작업 환경에 대해서는 후술한다.

분석 말뭉치 주석 검증의 작업 성과량을 수치로 보이면 다음과 같다.

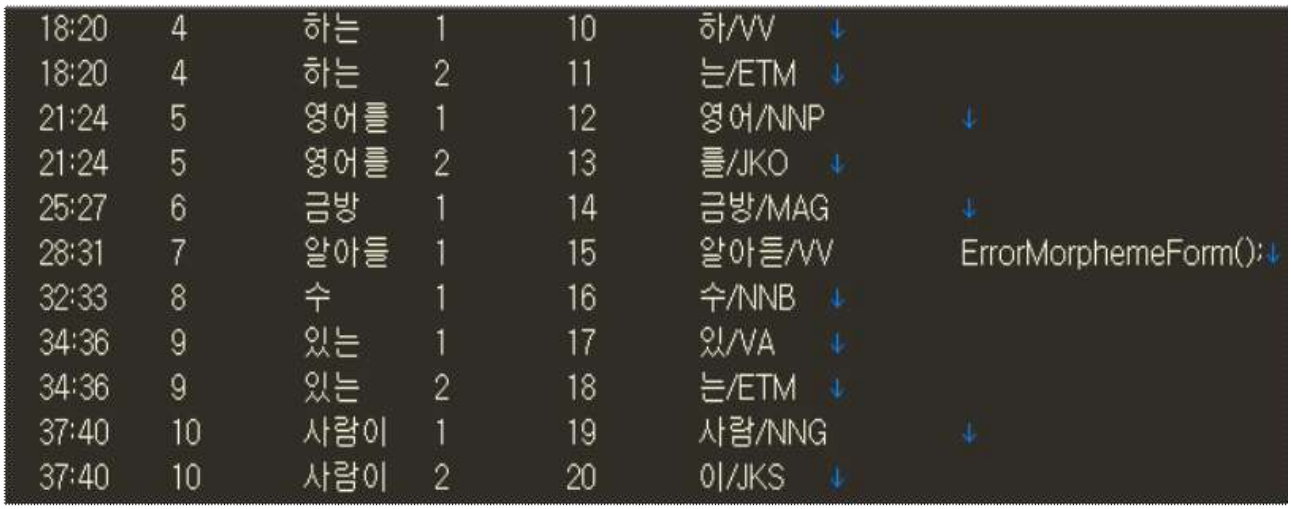
<표 12> 분석 말뭉치 주석 검증의 작업 성과량

층위	작업 분량	수정량
형태 분석	599,731개 어절	1,674개 어절
어휘의미	434,992개 어휘 태그	3,197개 어휘 태그
개체명	130,793개 개체명	개체명 태그 변경 1,489개 개체명 태그 추가 5,013개 개체명 태그 삭제 2,976개
상호참조 해결	29.863개 멘션	멘션 그룹 변경 350개 멘션 삭제 75개 멘션 추가 14,202개
구문 분석	30,066개 문장	기능 태그 변경 5099문장 지배 관계 변경 3069문장
의미역	85,327개 서술어-논항 세트	37,287개 서술어-논항 세트
주격 무형 대용어 복원	79,722개 선행어	21,994개 선행어

아래부터는 개별 분석 말뭉치 층위별로 이루어진 과업 수행 과정을 기술한다.

3.2.1 형태 분석

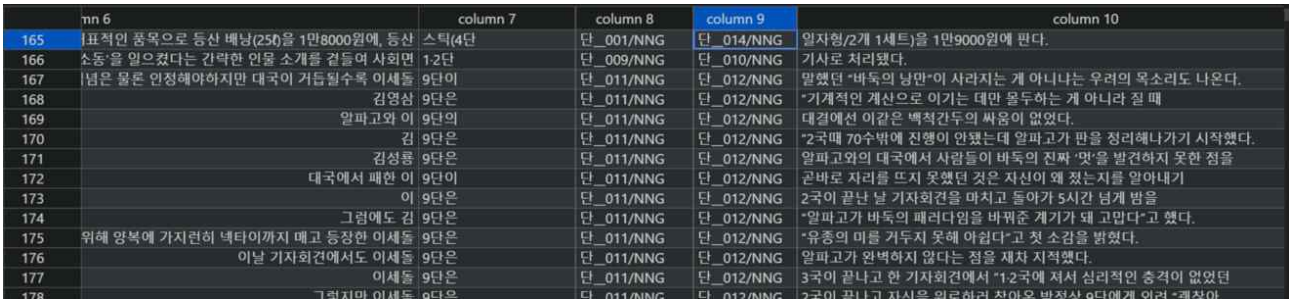
형태 분석 말뭉치 검증을 위하여 우선 예비 검증 작업에서 제보된 오류 유형에 대하여, 오류 후보를 전산 탐색하여 표지를 부착하였다. 어절과 형태소, 품사가 주석되어 있고 오류 후보 표지가 부착되어 있는 tsv 파일을 제공하여 텍스트 에디터 환경에서 검증을 수행할 수 있도록 하였다. 과업 수행 중반에 바이칼AI 자동 형태소 분석의 도움을 받아 바이칼AI의 형태 분석과 구축 말뭉치의 형태 분석 사이에 차이가 있는 지점을 참고 자료로 전문가에게 제공하였다.



<그림 15> 형태 분석 말뭉치 검증 작업 환경

3.2.2. 어휘 의미

어휘 의미 말뭉치는 다의어의 의미 번호를 파악하는 것이 검증 작업의 핵심이므로, 어휘별로 정렬된 데이터가 유용할 것이라고 예비 검증 작업에서 판단되었다. 동일한 어휘, 동일한 의미 번호로 주석된 단어들의 여러 용례 가운데 일부가 이질적인 문맥에서 사용된 것이 눈에 띈다면 곧바로 주석 오류로 의심할 수 있기 때문이다. 이에 작업 대상 어휘를 정렬한 후 앞뒤 문맥을 포함한(KWIC) tsv 파일을 제공하여 검증 작업을 수행하였다.



	mn 6	column 7	column 8	column 9	column 10
165	표적인 품목으로 등산 배낭(25l)을 1만8000원에, 등산	스틱(4단	단__001/NNG	단__014/NNG	일자형/2개 1세트)을 1만9000원에 판다.
166	소동'을 일으켰다는 간략한 인물 소개를 곁들여 사회면	1·2단	단__009/NNG	단__010/NNG	기사로 처리됐다.
167	념은 물론 인정해야하지만 대국이 거듭될수록 이세돌	9단이	단__011/NNG	단__012/NNG	말했던 "바둑의 낭만"이 사라지는 게 아니냐는 우려의 목소리도 나온다.
168	김영삼	9단은	단__011/NNG	단__012/NNG	"기계적인 계산으로 이기는 데만 몰두하는 게 아니라 질 때
169	알파고와 이	9단의	단__011/NNG	단__012/NNG	대결에선 이같은 백척간두의 싸움이 없었다.
170	김	9단은	단__011/NNG	단__012/NNG	"2국째 70수밖에 진행이 안됐는데 알파고가 판을 정리해나가기 시작했다.
171	김성룡	9단은	단__011/NNG	단__012/NNG	알파고와의 대국에서 사람들이 바둑의 진짜 '맛'을 발견하지 못한 점을
172	대국에서 패한 이	9단이	단__011/NNG	단__012/NNG	곧바로 자리를 뜨지 못했던 것은 자신이 왜 졌는지를 알아내기
173	이	9단은	단__011/NNG	단__012/NNG	2국이 끝난 날 기자회견을 마치고 돌아가 5시간 넘게 밤을
174	그럼에도 김	9단은	단__011/NNG	단__012/NNG	"알파고가 바둑의 패러다임을 바꿔준 계기가 돼 고맙다"고 했다.
175	위해 양복에 가지런히 넥타이까지 매고 등장한 이세돌	9단은	단__011/NNG	단__012/NNG	"유종의 미를 거두지 못해 아쉽다"고 첫 소감을 밝혔다.
176	이날 기자회견에서도 이세돌	9단은	단__011/NNG	단__012/NNG	알파고가 완벽하지 않다는 점을 재차 지적했다.
177	이세돌	9단은	단__011/NNG	단__012/NNG	3국이 끝나고 한 기자회견에서 "1·2국에 져서 심리적인 충격이 없었던
178	그럼지만 이세돌	9단은	단__011/NNG	단__012/NNG	2국이 끝나고 자신을 위로하러 찾아온 박정상 9단에게 위로 <패찬아

<그림 16> 어휘 의미 말뭉치 검증 작업 환경

한편 작업 중인 전문가의 요청에 따라 형태 분석과 어휘 의미의 형태소 주석이 불일치하는 경우를 추출하여 제공하기도 하였다. 형태 분석 말뭉치와 어휘 의미 말뭉치의 구축 지침이 같지 않기 때문에, 두 말뭉치의 형태소 주석이 불일치하는 것이 꼭 오류는 아니지만, 그만큼 판단이 까다롭고 쟁점사항이 잠재해 있는 경우라는 예측이 가능하므로 참고 자료로 가치가 있는 것으로 판단하였다.

77	계도 인기를 끌 것으로 본다"고 말했다.	미국/NNP + 산/XSN	미국산__001/NNG	닭고기에 반덤핑관세 중국이 미국산 닭고기에 반덤핑관세를
78	다. 미국산 닭고기에 반덤핑관세 중국이	미국/NNP + 산/XSN	미국산__001/NNG	닭고기에 반덤핑관세를 부과하기 시작했다. 환율과 무역을
79	석도 나오고 있다. 중국 상무부는 26일	"/SS + 미국/NNP + 산/XSN	"/SS + 미국산__001/NNG	수입 닭고기가 중국 국내 산업에 심대한
80	올해 상반기에 지난해 같은 기간에 견줘	미국/NNP + 산/XSN	미국산__001/NNG	닭고기 수입이 6.54% 증가했고, 이에 따른
81	결론 이전의 예비적 조처로 지난달 말	미국/NNP + 산/XSN	미국산__001/NNG	닭고기에 4~30.3%의 상계관세를 매기기 시작했다. 미국산은
82	30.3%의 상계관세를 매기기 시작했다.	미국/NNP + 산/XSN + 은/JX	미국산__001/NNG + 은/JX	중국의 닭고기 수입물량 중 90%가량을 차지한다.
83	격이 묻어난다. 중국 상무부는 "우리는	미국/NNP + 산/XSN	미국산__001/NNG	닭고기 수입을 10년 전부터 허용했지만 미국은
84	했다는 일화를 소개했다. 시 부주석이	서구/NNP + 에/JKB + 는/JX	서구__002/NNP + 에/JKB + 는/JX	호의적이지 않다는 의미인 셈이다. 일본 정부는
85	다고 26일 밝혔다. 그러나 이런 수치는	경제/NNG + 협력/NNG + 개발/NNG + 기구	경제협력개발기구__001/NNP + (/SS + OECD/SL +	회원국 평균인 5.6명과 견줄 때 여전히
86	이하에서 1억 원 이하로 늘어난다. 또	주택/NNG + 금융/NNG + 신용/NNG + 보증	주택금융신용보증기금__001/NNP + 의/JKG	전세자금 대출보증 규모도 지난해 5조8000억 원에서
87	한 뒤 벤처기업을 창업하도록 유도하는	한국/NNP + 형/XSN	한국형__001/NNG	'탈피오트' 프로그램을 만들기로 했다. '탈피오트(Talpiot)'는 히
88	처기업을 창업하도록 유도하는 한국형	'/SS + 탈피오트/NNG + '/SS	'/SS + 탈피오트__777/NNG + '/SS	프로그램을 만들기로 했다. '탈피오트(Talpiot)'는 히브리어로 '
89	'탈피오트' 프로그램을 만들기로 했다.	'/SS + 탈피오트/NNG + (/SS + Talpiot/SL +)	'/SS + 탈피오트__777/NNG + (/SS + Talpiot/SL +),	히브리어로 '최고 중의 최고'를 뜻하는 말로
90	으로 보고 있다. 정부가 이번에 내놓은	한국/NNP + 형/XSN	한국형__001/NNG	탈피오트 프로그램은 지경부의 소프트웨어(SW) 인력양성 프
91	고 있다. 정부가 이번에 내놓은 한국형	탈피오트/NNG	탈피오트__777/NNG	프로그램은 지경부의 소프트웨어(SW) 인력양성 프로그램인 '
92	우려마저 나오고 있다. 18일 금융권 및	나이스신용평가정보/NNP + 에/JKB	나이스신용평가정보__777/NNP + 에/JKB	따르면 지난해 12월 91만9570명이던 금융기관 연체자는
93	였다는 것이다. 그래서 모던한 공간이나	인도/NNP + 풍/XSN + 의/JKG	인도풍__001/NNG + 의/JKG	에스닉한 분위기와도 잘 매치된다. 어느덧 겨울의
94	5년 물타신의 대통령경호대 요별부대	벌/XPN + 아프리카/NNP + 군단/NNG + /SS	벌아프리카__777/NNG + 군단__001/NNG + /SP	청년 미별대 조직 '현명의원회'가 수도 트리폴리를

<그림 17> 형태 분석 말뭉치와 어휘 의미 말뭉치의 연계 검증 작업 환경

3.2.3. 개체명

개체명 말뭉치에는 크게 두 가지 오류 유형이 있었는데, 각각의 오류 유형에 대해 각기 다른 형태의 작업 환경을 제공할 필요가 있었다. 두 가지 오류 유형은 개체명 주석 오류와 개체명 범위/누락 오류였다. 개체명 주석 오류는 개체명 태그를 잘못 붙인 오류

이고, 개체명 범위/누락 오류는 개체명 태그를 붙여야 하는 단어/구의 범위를 잘못 설정하거나 아예 개체명 태그를 붙이지 않은 오류이다. 전자인 개체명 주석 오류는 어휘 의미 말뭉치 검증의 경우와 같이 작업 대상 개체명을 정렬한 후 앞뒤 문맥을 포함한 tsv 파일을 제공하여 검증 작업을 수행하였다. 개체명 범위/누락 오류의 경우 텍스트를 어절별로 세로로 읽어 나갈 수 있는 형태의 tsv 파일을 제공하여 오류를 검증할 수 있도록 하였다. 이 때 작업의 참고를 위해 형태 분석 말뭉치의 주석 정보를 첨부하여 제공하였다.

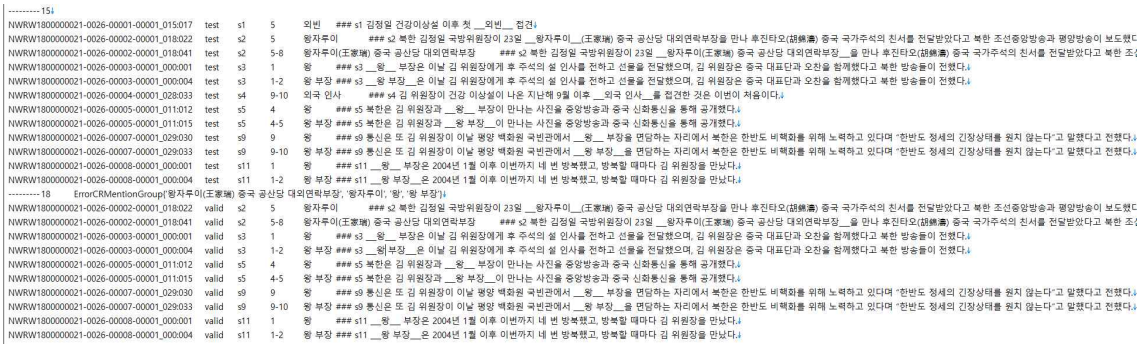
NWRW1800000021-0003-00001-000	"빅3	빅3/QT		"/SS + 빅__888/NNG + 3/SN
NWRW1800000021-0003-00001-000	흔들릴때..."			흔들리/VV + ㄹ/ETM + 때__001/NNG + .../SE + "/
NWRW1800000021-0003-00001-000	유럽차	유럽/LC + 차/AF		유럽__002/NNP + 차__008/NNG
NWRW1800000021-0003-00001-000	美시장	美/LC		美/SH + 시장__005/NNG
NWRW1800000021-0003-00001-000	본격공략			본격__888/NNG + 공략__003/NNG
NWRW1800000021-0003-00002-000	제너럴모터스(GM)	제너럴모터스/OG + GM/		제너럴모터스__777/NNP + (/SS + GM/SL +)/SS
NWRW1800000021-0003-00002-000	포드	포드/OG		포드__888/NNP
NWRW1800000021-0003-00002-000	크라이슬러	크라이슬러/OG		크라이슬러__888/NNP
NWRW1800000021-0003-00002-000	등			등__010/NNB
NWRW1800000021-0003-00002-000	생사의			생사__002/NNG + 의/JKG
NWRW1800000021-0003-00002-000	기로에			기로__002/NNG + 에/JKB
NWRW1800000021-0003-00002-000	선			서/VV + ㄴ/ETM
NWRW1800000021-0003-00002-000	미국의	미국/LC		미국__004/NNP + 의/JKG
NWRW1800000021-0003-00002-000	'빅3'가	빅3/QT		'/SS + 빅__888/NNG + 3/SN + '/SS + 가/JKS
NWRW1800000021-0003-00002-000	유럽	유럽/LC		유럽__002/NNP
NWRW1800000021-0003-00002-000	자동차회사들의	자동차/AF		자동차__001/NNG + 회사__004/NNG + 들/XSN +
NWRW1800000021-0003-00002-000	미국	미국/LC		미국__004/NNP

<그림 18> 개체명 말뭉치 개체명 범위/누락 오류 검증 환경

3.2.4. 상호참조 해결

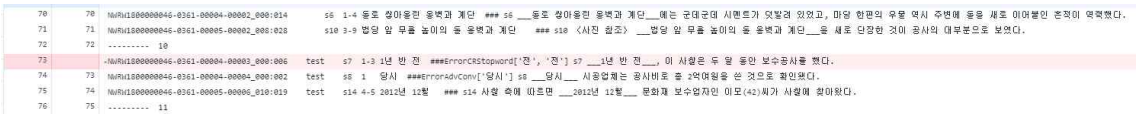
상호참조 해결 말뭉치 검증은 구축 사업 말뭉치와 검증 사업 말뭉치라는 같은 원시 말뭉치에 대한 두 벌의 주석이 있으므로, 이 조건을 최대한 활용할 수 있도록 작업 환경을 구성하였다. 두 벌의 말뭉치를 하나로 합치고, 구축 말뭉치에만 나타나는 멘션, 검증 말뭉치에만 나타나는 멘션 등을 표시하여 확인할 수 있도록 하였다. 또, 두 말뭉치에 모두 나타나되 서로 다른 멘션 그룹에 속하는 멘션을 파악할 수 있도록 하였다.

중첩되는 멘션을 갖는 그룹들의 경우 오류 표지 ErrorCRMentionGroup을 그룹 구분선에 부착하여 그룹 통합 대상을 판단할 수 있도록 하였다. 전문가는 두 그룹의 멘션들이 하나의 그룹으로 통합될 만한 것인지, 혹은 하나의 멘션에 대해 두 사업단에서 각기 다른 주석을 달아 발생하는 것인지 파악한 후 그룹을 통합하거나 새로 생성, 분리하는 등의 작업을 수행할 수 있었다.

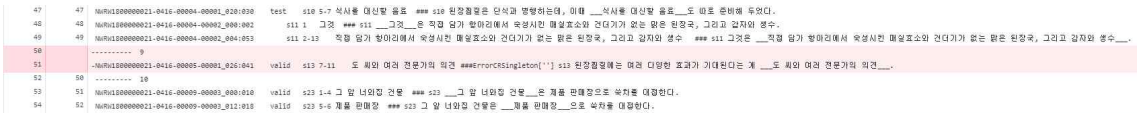


<그림 19> 상호참조 해결 말뭉치 멘션 그룹 오류 검증 작업 환경

그 외에도 불용어 오류 표지를 부착한 작업 환경, 단일 멘션 그룹 표지를 부착한 작업 환경 등을 전문가의 필요에 따라 상시적으로 제공하였다. 단일 멘션 그룹 오류(멘션 그룹 안에 멘션이 하나만 속해 있는 오류)는 작업 절차상 후반부에 이루어져야 하였다. 타 오류의 수정 과정에서 일시적으로 멘션 그룹 안에 멘션이 하나만 속하게 될 수 있기는데, 이를 섣불리 오류로 판정해서는 안 되기 때문이다.



<그림 20> 상호참조 해결 말뭉치 불용어 오류 검증 작업 환경



<그림 21> 상호참조 해결 말뭉치 단일 멘션 그룹 오류 검증 작업 환경

3.2.5. 구문 분석

구문 분석 말뭉치의 검증을 위해서 구문 태그, 기능 태그가 기재된 tsv 파일을 작업 파일로서 제공하여 검증을 수행하였다. 이후 검증이 진행되며 전문가 요청에 따라 오류 유형에 표지를 부착하여 제공하거나 문제 어절을 정렬하여 제공하였다. 다음은 기능 태그가 누락된 NP에 표지(IssueBareNP)를 부착하여 제공하고 수정을 수행한 예이다. (붉은색이 변경 전, 녹색이 변경 후)

```
  NWRW1800000021-0020-00005-00001_019  19  뉴욕      20  NP
  NWRW1800000021-0020-00005-00001_020  20  증시에서  23  NP_AJT
  NWRW1800000021-0020-00005-00001_021  21  주가가    23  NP_SBJ
- NWRW1800000021-0020-00005-00001_022  22  23.2%나  23  NP      IssueBareNP(주가가)(▶23.2%나◀)(폭락했다.);
+ NWRW1800000021-0020-00005-00001_022  22  23.2%나  23  NP_AJT  IssueBareNP(주가가)(▶23.2%나◀)(폭락했다.);
  NWRW1800000021-0020-00005-00001_023  23  폭락했다.  -1  VP
```

<그림 22> 구문 분석 말뭉치 기능 태그 누락 오류 검증 작업 환경

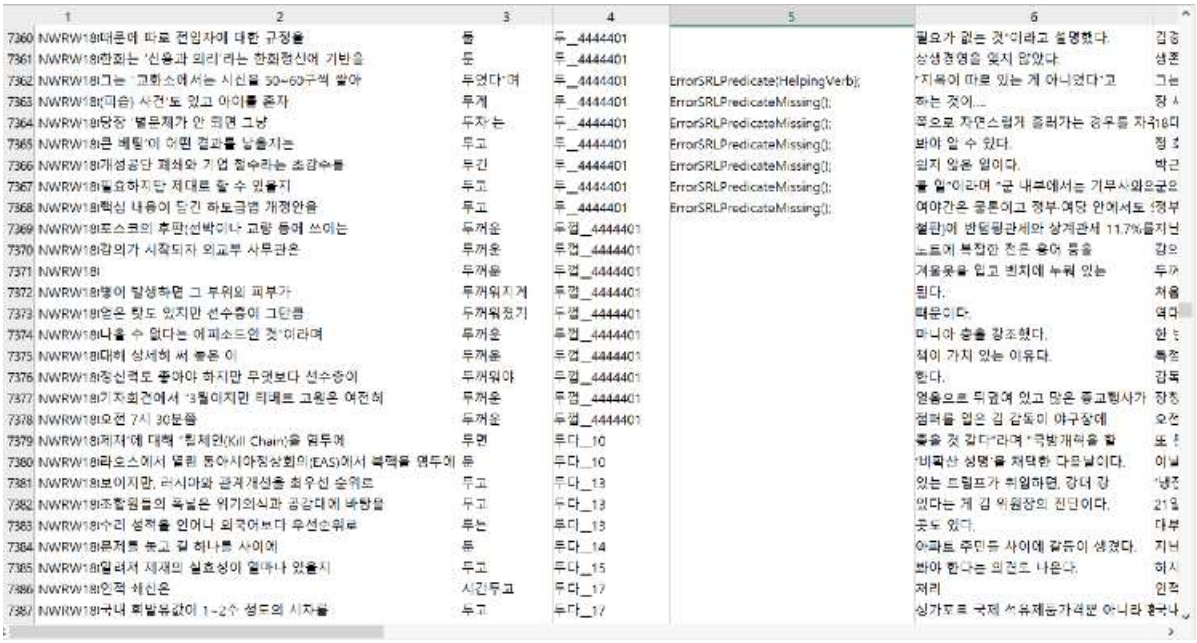
구문 분석 주석 작업에서는 '우리말샘' 구문 틀을 참고하므로, '우리말샘' 용언의 구문틀 정보만 추출하여 목록을 만들어 전문가에게 제공하였다.

	어휘	문형
2	가르치다	【…에게 …을】【…에게 -고】
3	가상하다	【…을】【…을 …으로】【…을 -고】【-고】
4	가정하다	【…을】【…을 …으로】【…을 -고】【-고】
5	가정하다	【…을】【…을 …으로】【…을 -고】【-고】
6	가정하다	【…을】【…을 …으로】【…을 -고】【-고】
7	가짓말하다	【…에/에게 -고】
8	가칭하다	【…을 …으로】【…을 -고】
9	가칭하다	【…을 …으로】【…을 -고】
10	각오하다	【…을】【-기로】【-고】
11	각오하다	【…을】【-기로】【-고】
12	간원하다	【…에게 …을】【…에게 -기를】【…에게 -고】

<그림 23> 구문 분석 말뭉치 검증용 우리말샘 구문틀 목록

3.2.6. 의미역

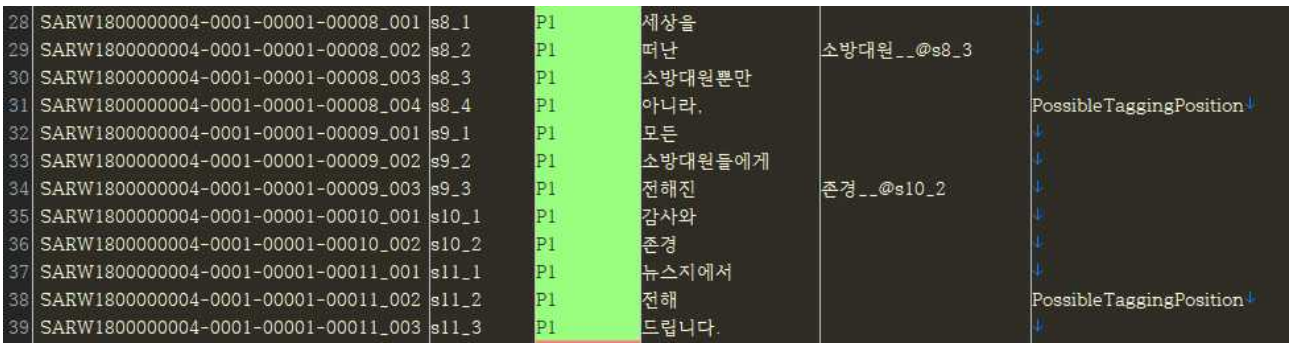
의미역 말뭉치의 검증을 위해 논항 태그, 술어 태그가 기재된 tsv 파일을 제공하였다. 의미역은 일괄적으로 검증하는 것이 효율적인 오류 유형이 많았기 때문에 정렬된 KWIC 환경을 적극 활용하였다. 아래는 서술어 과잉태깅 오류(태깅 배제 대상 서술어에 sense가 태깅된 경우)를 수정하기 위해 제공된 정렬 KWIC 환경의 예이다.



<그림 24> 의미역 말뭉치 서술어 과잉태깅 오류 검증 작업 환경

3.2.7. 주격 무형 대용어 복원

주격 무형 대용어 복원 말뭉치의 검증을 위하여 화자 정보(구어 말뭉치의 경우), 잠재적 복원 대상 어절(즉, 서술어가 포함된 어절)이 표시된 tsv 파일을 작업 환경으로 제공하였다.



<그림 25> 주격 무형 대용어 복원 말뭉치 검증 작업 환경

아울러 복원되어 있는 주어 목록(어절 번호 포함)을 별도 파일로 제공하여 참고할 수 있게 하였다.

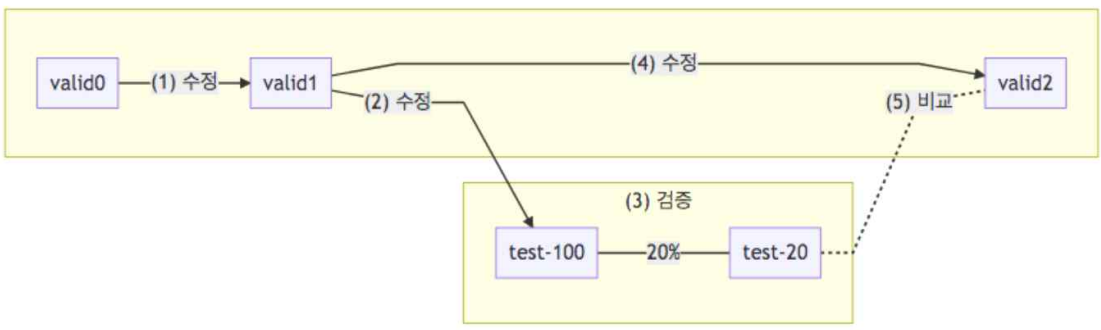
SARW1800000004	P1	존경__@s10_2
SARW1800000004	P1	추모식__@s7_3
SARW1800000004	P1	소방대원__@s6_4
SARW1800000004	P1	소방대원__@s8_3
SARW1800000004	P1	소방대원들__@s3_3
SARW1800000004	P2	딸__@s24_4
SARW1800000004	P2	그__@s27_5
SARW1800000004	P2	규모__@s15_5

<그림 26> 주격 무형 대용어 복원 말뭉치 검증용 복원 주어 목록

4. 유지 보수 말뭉치와의 비교 검증

본 과업은 분석 말뭉치 구축 사업단에서 관리하는 '유지 보수 말뭉치'의 수정 작업과 연계하여 진행되었다. 본 연계 과업에 해당되는 층위는 상호참조 해결, 구문 분석, 의미역, 주격 무형 대용어 복원이었다.

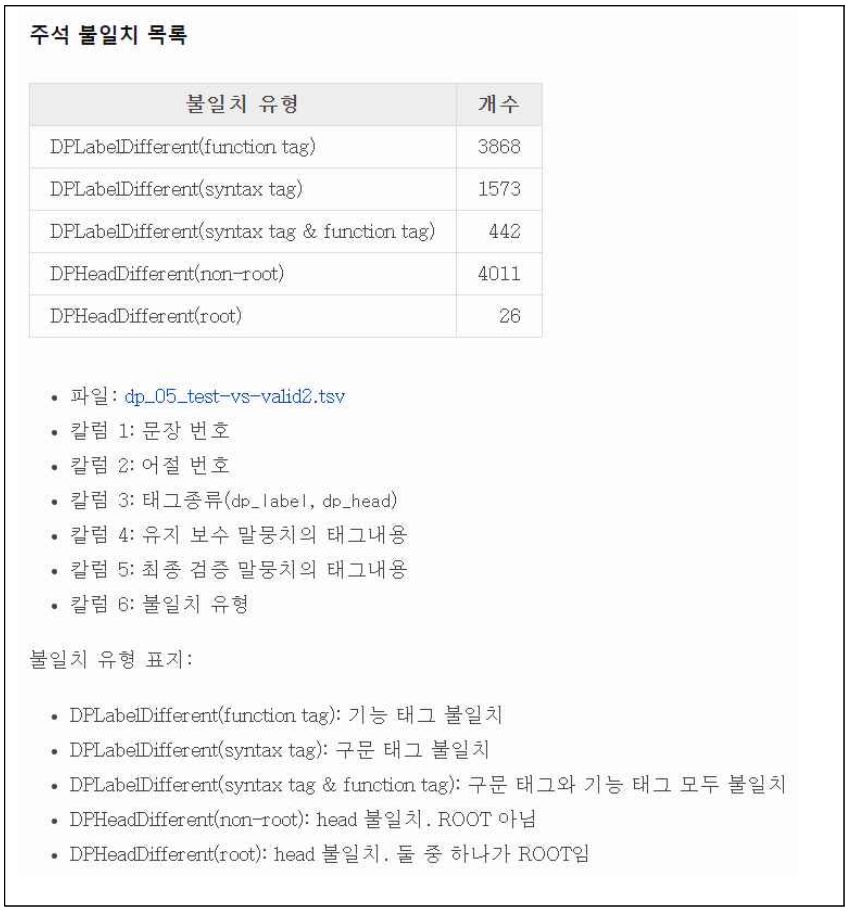
먼저 본 사업의 층위 통합 말뭉치를 구축 사업단에 발송하고, 구축 사업단은 본 사업과는 별도로 유지 보수 업무를 진행하였다. 사업의 마지막 주간에 유지 보수 말뭉치를 본 사업단이 수령하였다. 본 사업단에서는 전문가 주석 검증이 완료된 말뭉치를 먼저 납품하고, 유지 보수 말뭉치를 검토하고 문제 사항을 보고하였다. 그와 함께 유지 보수 말뭉치와 본 과업의 전문가 주석 완료 말뭉치를 비교하고 그 결과를 보고하였다. 구축 사업단의 유지 보수 업무와 연계한 검증 과정을 도식으로 나타내면 다음과 같다.



* vaild0: 최초 말뭉치 (사업 개시 시점)
* valid1: 본 사업단의 중간 층위 통합 말뭉치
* test-100: 구축 사업단의 유지 보수 말뭉치 (전체)
* test-20: 구축 사업단의 유지 보수 말뭉치 (본 사업단과 공통되는 범위의 검증 문서)
* valid2: 본 사업단의 전문가 주석 완료 말뭉치

<그림 27> 유지 보수 연계 검증 과정

이 비교 과정을 거쳐 '본 사업단은 무엇을 수정하였는가', '구축 사업단은 무엇을 수정하였는가', '본 사업단의 수정 내역과 구축 사업단의 수정 내역 사이에 어떤 차이가 있었는가'에 초점을 둔 자료를 추출하고 그 결과를 별도로 제출하였다. 본 사업단의 수정 내역과 구축 사업단의 수정 내역간 차이를 전수 조사하고, 차이의 유형에 따라 통계 및 범례를 제시하였다.



주석 불일치 목록

불일치 유형	개수
DPLabelDifferent(function tag)	3868
DPLabelDifferent(syntax tag)	1573
DPLabelDifferent(syntax tag & function tag)	442
DPHeadDifferent(non-root)	4011
DPHeadDifferent(root)	26

- 파일: dp_05_test-vs-valid2.tsv
- 칼럼 1: 문장 번호
- 칼럼 2: 어절 번호
- 칼럼 3: 태그종류(dp_label, dp_head)
- 칼럼 4: 유지 보수 말뭉치의 태그내용
- 칼럼 5: 최종 검증 말뭉치의 태그내용
- 칼럼 6: 불일치 유형

불일치 유형 표지:

- DPLabelDifferent(function tag): 기능 태그 불일치
- DPLabelDifferent(syntax tag): 구문 태그 불일치
- DPLabelDifferent(syntax tag & function tag): 구문 태그와 기능 태그 모두 불일치
- DPHeadDifferent(non-root): head 불일치. ROOT 아님
- DPHeadDifferent(root): head 불일치. 둘 중 하나가 ROOT임

<그림 28> 유지 보수 말뭉치와의 비교 자료 범례의 예 (구문 분석 층위)

아울러, 본 사업단이 전문가 주석 검증 과정에서 관찰한 사실과 축적한 경험을 활용하여, 유지 보수 말뭉치 수령 후 말뭉치 전체 범위에서 수정해야 할 오류의 후보를 제시하기도 하였다.

<그림 29> 오류 후보의 제시 방법의 예 (의미역 층위)

5. 분석 말뭉치 오류 유형 정리

5.1. 형태 분석

형태 분석 말뭉치에서 발견된 오류의 유형은 다음과 같다.

5.1.1. 과도한 복원/교정

1) 조사 '-이든' 등이 모음으로 끝나는 체언 뒤에 나올 때 지침과 달리 '이'를 입력한 경우

① 특목고든 과학고든: 과학고/NNG+이든/JX (바른 주석: 과학고/NNG+든/JX)

② 사업에 진출할 좋은 기회"라고 말했다.: 기회/NNG+"/SS+이라고/JKQ (바른 주석: 기회/NNG+"/SS+라고/JKQ)

③ 약초라든가 생약이라든가 이런 걸로: 약초/NNG+이라든가/JX (바른 주석: 약초/NNG+라든가/JX)

2) 대상 자료의 형태를 표준 어형 등으로 바꾸어 입력한 경우

① 오래만간에 좋은 강연도 들었고: 오래간만/NNG + 에/JKB (바른 주석: 오래만간/NNG + 에/JKB)

② 그게 산후우울증이였구나 하는게 있었던: 산후/NNG+우울증/NNG+이/VCP+었

/EP+구나/EF (바른 주석: 산후/NNG+우울증/NNG+이/VCP+였/EP+구나/EF)

③ 아침이던지 저녁이던지 스크랩을 한 그런 신문: 아침/NNG+이든지/JX (바 른 주석: 아침/NNG+이던지/JX)

④ 안돼네요.: 안/MAG+되/XSV+네/EF+요/JX+./SF (바른 주석: 안/MAG+돼/XSV+네/EF+요/JX+./SF)

⑤ 뽑아달래는: 뽑/VV+아/EC+달/VX+라는/ETM (바른 주석: 뽑/VV+아/EC+달/VX+래는/ETM)

5.1.2. 어미의 예외적 표기 지침 위반

1) 'ㅅ' 불규칙 용언 뒤에 전성어미가 올 때 지침과 달리 '으'를 표기한 경우

① 황룡사를 지은 진흥왕이: 짓/VV+은/ETM (바른 주석: 짓/VV+ㄴ/ETM)

② 강영주 지음/사계절·3만2000원: 짓/VV+음/ETN (바른 주석: 짓/VV+ㅁ/ETN)

2) '바라다'의 활용형 관련 분석 지침 위반

① 들어서길 바랬던 겁니다.: 바래/VV+었/EP+던/ETM (바른 주석: 바라/VV+았/EP+던/ETM)

5.1.3. 과다/과소 분석

1) 등재어의 분석

① 어 어쨌건 이 지구촌에서는: 어찌/MAG+하/XSV+았/EP+건/EC (바른 주석: 어쨌건/MAG)

② 가족경영은 머잖아 3, 4대로 넘어갈 것이다.: 멀/VA+지/EC+않/VX+아/EC (바른 주석: 머잖/VA+아/EC)

2) - '어요, 지요' 등의 '요' 분석

① 큰 글자를 보지 마요.: 말/VX+아/EF+요/JX+./SF (바른 주석: 말/VX+아요/EF+./SF)

② 제가 북한 문제에 관심 많잖아요.: 많/VA+잖아/EF+요/JX+./SF (바른 주석: 많/VA+잖아요/EF+./SF)

3) '어야지'의 분석

① 버텨야죠: 버티/VV+어야/EC+지/EF+요/JX (바른 주석: 버티/VV+어야지/EF+요

/JX)

4) 비분석 접사의 분석

① 미국 노드롭사는: 노드롭/NNP+사/NNG+는/JX (바른 주석: 노드롭사/NNP+는/JX)

5) 분석 대상 접사의 미분석

① 관례상: 관례상/NNG (바른 주석: 관례/NNG+상/XSN)

② 부부간의: 부부간/NNG+의/JKG (바른 주석: 부부/NNG+간/XSN+의/JKG)

6) 한자 표기에 대한 미분석(고유명사 처리 방식을 우선 적용한 경우)

① 오대양號: 오대양號/NNP (바른 주석: 오대양/NNP+號/SH)

7) '-ㄴ다든지'의 '-든지'에 대한 미분석('-ㄴ다든지'는 '우리말샘'과 지침에서 하나의 어미로 인정되지 않음)

① 십 분 스트레칭을 하고 나간다든지: 나가/VV+ㄴ다든지/EC (바른 주석: 나가/VV+ㄴ다/EF+든지/JX)

② 점심을 설렁탕을 먹는다든지: 먹/VV+는다든지/EC (바른 주석: 먹/VV+는다/EF+든지/JX)

5.1.4. 태깅 방식의 혼동

1) 숫자 사이에 나타난 마침표에 대한 태깅 오류,

① 기본점수(9.50→10.0)는: 9/SN+./SF+50/SN, 10/SN+./SF+0/SN (바른 주석: 9/SN+./SP+50/SN, 10/SN+./SP+0/SN)

2) 숨김표(xx)에 대한 처리 방식(해당 어절이 모두 x로 표기된 경우 SL을 붙여야 함.)

① 지금 xxx 문젠데: xxx/NA (바른 주석: xxx/SL)

5.1.5. 품사 판정 오류

1) 중의성을 지닌 형태 ('일제, 앞서, 다른, 그래도' 등)

① '일제': '우리말샘'에 따르면, '일본 제국'의 뜻과 '일본 제국주의'의 뜻이 있다. 고유

명사 관련 지침에 따르면 '일본 제국'은 고유명사로 처리해야 한다. 그러나 분석 대상 자료에서는 '일제'를 고유명사로 분석한 경우가 없었다.

- 그러나 일제는 조선을 침탈한 뒤: 일제/NNG+는/JX (바른 주석: 일제/NNP+는/JX)

② '앞서'

- 운동하겠다는 마음만 앞서 무리를 하기 때문이다.: 앞서/MAG (바른 주석: 앞서/VV+어/EC)
- 10%포인트 이상 차이를 벌리며 앞서 갔다.: 앞서/MAG (바른 주석: 앞서/VV+어/EC) (이 예는 '앞서가다'를 단일어로 처리해야 하지만 어절이 나뉘어 있어서 그렇게 할 수 없는 경우이다.)
- 우선 한마디로 앞서 여러 편 뭐 굉장히 불상스럽고: 앞서/VV+어/EC (바른 주석: 앞서/MAG)

③ '다른'

- 다른 꿀에다가 설탕을 넣는 경우도 있고: 다르/VA+ㄴ/ETM (바른 주석: 다른/MMD)
- 평소와 다른 모습을 보이는 딸을 다그쳐: 다른/MMD (바른 주석: 다르/VA+ㄴ/ETM)

④ '-ㄴ지': 보절을 이끄는 '-ㄴ지'는 종결어미로 볼 수 있음.

- 이제 누가 어느 당인지조차 헷갈릴 지경이 됐다.: 당/NNG+이/VCP+ㄴ지/EC+조차/JX (바른 주석: 당/NNG+이/VCP+ㄴ지/EF+조차/JX)
- 수비가 좋은 선수가 누구인지 가려낼 수 있다.: 누구/NP+이/VCP+ㄴ지/EC (바른 주석: 누구/NP+이/VCP+ㄴ지/EF)

⑤ '홍콩 달러': 단위사로 쓰인 경우. 고유명사 처리 방식과도 관련됨.

- 31억홍콩달러(4515억원)를: 31/SN+억/NR+홍콩/NNP+달러/NNG+(/SS+4515/SN+억/NR+원/NNB+)/SS+를/JKO (바른 주석: 31/SN+억/NR+홍콩/NNP+달러/NNB+(/SS+4515/SN+억/NR+원/NNB+)/SS+를/JKO)

2) 고유명사 판단: 명백한 오류도 있지만, 달리 처리할 가능성이 높은 경우가 꽤 많다. 지침의 보완이 필요하다고 판단된다. 접사, 한자, 알파벳, 숫자 등이 포함되어 있을 때

고유명사 판단 및 처리 방식과 상충하는 지침 간의 우선 순위 등을 명세할 필요가 있다. 국제기구명, 단체명 등과 관련된 부분도 보완이 필요하다고 여겨진다.

- 사카타공장: 사카타공장/NNP (바른 주석: 사카타/NNP+공장/NNG)
- 장유1동: 장유1동/NNP (바른 주석: 장유/NNP+1/SN+동/NNG)
- 종로1가: 종로1가/NNP (바른 주석: 종로/NNP+1가/NNG)
- 을지로 오가에: 오가/NNP+에/JKB (바른 주석: 오가/NNG+에/JKB)
- 슈퍼맨이 돌아왔다 (TV 프로그램 명칭): 슈퍼맨/NNP (바른 주석: 슈퍼맨/NNG)
- 제3한강교: 제/XPN+3/SN+한강교/NNP (바른 주석: 제3한강교/NNP)
- 제2개성공단을 검토한 바 없다.: 제2개성공단/NNP+을/JKO (바른 주석: 제/XSN+2/SN+개성공단/NNP+을/JKO)
- UN안보리: UN안보리/NNG (바른 주석: UN안보리/NNP)
- 국립과학수사연구원: 국립과학수사연구원/NNP (바른 주석: 국립/NNG+과학/NNG+수사/NNG+연구원/NNG)

5.2. 어휘 의미

어휘 의미 말뭉치에서 발견된 오류의 유형은 다음과 같다.

5.2.1. 동음이의어 수준의 오류

이 오류는 어휘의미 번호를 등재되어 있는 동음이의어(동형어)의 어휘의미 번호로 잘못 부여한 유형의 오류이다. 우리말샘에 별도의 표제어로 각 어휘의미가 등재되어 있는 경우에 해당한다.

① 김 검사장은 서울중앙지검 특수1부장 출신으로 2013년 원전(原電) 비리 합동수사단에 이어 작년 11월부터 방산 비리 합동수사단을 이끌고 있다.: 단__035/NNG (바른 주석: 단__021/NNG)

② 박원순 시장은 2004년 독일에서 세계인권선언 조문이 새겨진 베를린역을 보고 당시 이명박 시장에게 활용안을 권유한 바 있다.: 안__002/NNG (바른 주석: 안__012/NNG)

③ 김대중 전 대통령은 현 정부가 대량살상무기 확산방지구상(PSI) 전면 참여를 발표한 이후 한 언론 인터뷰에서 '서해 충돌 가능성'을 묻는 질문에 "저쪽(북한)에서도 PSI에 가입하면 선전포고로 인정한다고 했으니 가만있지 않을 것"이라고 했다. 선

전__003/NNG (바른 주석: 선전__006/NNG)

5.2.2. 다의어 수준의 오류

이 오류는 어휘의미 번호를 등재되어 있는 다의어의 어휘의미 번호로 잘못 부여한 유형의 오류이다. 우리말샘에 동일한 표제어 아래에 다른 의항으로 각 어휘의미가 등재되어 있는 경우에 해당한다.

① 김희동 진학사 입시분석실장은 “표준점수 만점과 1등급 구분점수 차가 9점으로 높게 나와 수리‘가’형에 강점이 있다면 소신 지원해도 좋을 것 같다”고 말했다.: 만점__002/NNG (바른 주석: 만점__001/NNG)

② 정부는 파르완 주에서 PRT를 운영하는 미국 외에 아프간 중앙정부, 지방정부와의 토지 매입 등에 관한 협의를 거쳐 최종적으로 PRT 설치 장소를 결정할 방침이다.: 지방__005/NNG (바른 주석: 지방__007/NNG)

③ 유모(24·여·서울대 4학년) 씨는 “다른 어학연수 프로그램에 비해 인턴을 병행할 수 있다는 게 장점인데 인턴 취업에 대한 확신이 없어 불안하다”면서 “정부에서는 인턴 임금으로 체류비를 충당할 수 있을 것이란 말만 되풀이하고 있는데 취업이 안 되면 결국 시간낭비만 하는 것 아니냐”고 지적했다.: 시간__004/NNG (바른 주석: 시간__014/NNB)

5.2.3. 우리말샘 등재 어휘의미의 미등재 처리 오류

이 오류는 우리말샘에 해당 어휘가 등재되어 있어서, 적절한 어휘의미 번호를 부여하거나 888을 부여해야 하는 경우이나 그렇지 않고 777을 부여한 오류에 해당한다.

① 1990년 현대중공업 골리앗 크레인에 올랐던 이갑용 전 민주노총 위원장(전 울산 동구청장)과 15만4000볼트의 송전탑에서 해고자 복직을 요구하며 88일(2011년 3월 6일~6월2일) 동안 고공농성을 벌였던 대우조선 비정규직 강병재씨가 편지에 이름을 올렸다. 골리앗__777/NNG (바른 주석: 골리앗__888/NNG)

② 한 쇄신위원은 1일 <한겨레>와 한 전화통화에서 “비정규직법 처리 직후 쇄신안을 공개할 계획이었으나, 비정규직법: 한겨레__777/NNP (바른 주석: 한겨레__888/NNP)

5.2.4. 우리말샘 미등재 어휘의미의 등재 처리 오류

이 오류는 우리말샘에 해당 어휘가 등재되어 있지 않아서, 777을 부여해야 하는 경우

이나 그렇지 않고 다른 어휘의미를 부여한 오류에 해당한다.

① 대법원이 휴대전화 단말기 구입 보조금은 부가가치세 부과 대상이 아니라고 판결했다.: 가치세__001/NNG (바른 주석: 가치세__777/NNG)
② [北 ICBM 2차 도발, 국방위-정보위 보고] 宋국방, 국회 국방위 긴급 현안보고: 정보위__001/NNG (바른 주석: 정보위__777/NNG)
③ 실무차원에서 제시됐던 이 방안은 국토해양부의 재검토 지시를 받았다.: 해양부__001/NNG (바른 주석: 해양부__777/NNG)

5.3. 개체명

개체명 말뭉치에서 발견된 오류의 유형은 다음과 같다.

5.3.1. 개체명 주석 오류

1) LC-OG/OG-LC 오류 : 장소(LC)를 기관(OG)으로, 기관(OG)을 장소(LC)로 잘못 태깅한 경우.

① 강원도는 시간이 지나면 저절로 회복된다며 버티고 있다.: 강원도/LC (바른 주석: 강원도/OG)
② 개성공단으로 갈까: 개성공단/OG (바른 주석: 개성공단/LC)
③ 개성공단이 폐쇄되고: 개성공단/OG (바른 주석: 개성공단/LC)

2) AF-OG 오류 : 건물(AF)과 기관(OG) 주석이 바뀐 오류.

① 올 들어 심각해진 전세난 대책을 찾고 있는 청와대와 정치권은 국토부에 싼 주택 공급을 늘리라고 압박하고 있다.: 청와대/AF (바른 주석: 청와대/OG)
② 여당은 '친이(親李)' 그룹이 해체된 상태이고, 어떻게든 청와대와 거리를 두려는 것이 당의 분위기다.: 청와대/AF (바른 주석: 청와대/OG)
③ 청와대 직원들은 "공무원들이 청와대 지시를 신경 안 쓴 지도 이미 상당히 됐다"고 하고 있다.: 청와대/AF (바른 주석: 청와대/OG)

3) OG-AF 오류 : 기관(OG)과 채널명(AF) 주석이 바뀐 오류.

① MBC의 오락 프로그램인 '세바퀴(세상을 바꾸는 퀴즈)'가 저녁 무렵에는 MBC계열 케이블채널인 'MBC에브리원'에서 방송되더니 이어서 같은 MBC 계열 채널 'MBC 드라마넷'에서 재방송이 나오고 다시 MBC 본 방송에서 '세바퀴'가 방송되더라는

것.: MBC드라마넷/OG (바른 주석: MBC드라마넷/AF)

② 바로 이어서 7시20분부터는 MBC드라마넷이 바통을 이어받아 재방송을 2회 방송했다.: MBC드라마넷/OG (바른 주석: MBC드라마넷/AF)

③ '무한도전'의 경우도 MBC드라마넷과 에브리원, MBC ESPN 3개 채널에서 177회나 방송했다.: MBC드라마넷/OG (바른 주석: MBC드라마넷/AF)

4) OG-CV 오류 : 기관(OG)과 직업(CV) 주석이 바뀐 오류

① "때리고 침뱉어" SNS 확산… 경찰 "뿌리치려다 부딪쳐" 후보측 "고의적… 침뱉진 않아": 경찰/OG (바른 주석: 경찰/CV)

② 이 사진은 '경찰이 김 후보의 얼굴에 침을 뱉고 주먹으로 내리쳤다'는 설명이 붙어 급속히 퍼졌다.: 경찰/OG (바른 주석: 경찰/CV)

③ 김 후보 측은 고의적인 폭행이라고 주장하지만 본보가 사건 당시 경찰 채증반이 촬영한 동영상을 확인한 결과 경찰의 폭력이 아닌 우발적 사건으로 보인다.: 경찰/OG (바른 주석: 경찰/CV)

④ 노 경위를 붙잡던 다른 경찰도 함께 붙들려 들어갔다.: 경찰/OG (바른 주석: 경찰/CV)

5.3.2. 개체명 범위 오류

1) LC + EV > LC 오류 : 장소(LC)와 사건(EV)이 연달아 나올 경우 실제 일어난 사건에 대해서만 태깅하므로 장소(LC)만 태깅해야 한다.

① 새누리당은(새누리당/OG) '**엔엘엘(엔엘엘/LC)** 국정조사 카드'로 맞섰다.

2) LC + FD > FD 오류 : 하나의 분야(FD)로 주석해야 할 것을 국가(LC) + 분야(FD)로 분리 주석한 오류.

① 이들 극장은 최근 한국 영화에 대한 부율을 55 대 45로 투자·배급사 쪽에 유리하게 5% 인상한 뒤, "한국 영화와 외화 간 형평성을 맞추겠다"며 외화 배급사들한테 외화 부율을 기존 '60 대 40'에서 '50 대 50'으로 조정한다고 통보한 바 있다.: 한국/LC + 영화/FD (바른 주석: 한국 영화/FD)

② 문학나눔 사업의 우수문학도서 선정을 출판진흥원의 세종도서 사업에 포함시키는 과정에서 문제가 나타난 것처럼 번역원을 출판진흥원에 흡수 통합할 경우 그동안 쌓아 온 한국문학 세계화의 성과가 무너진다는 것이다.: 한국/LC + 문학/FD (바른 주석: 한국문학/FD)

3) CV 오류: 인간관계 명칭에서는 '-님'을 제외하고 태깅해야 한다.

① 형님: 형님/CV (바른 주석: 형/CV)

② 홀가분…형님께: 형님/CV (바른 주석: 형/CV)

4) TI 오류 : '한 ~ 정도'로 범위를 한정하여 태깅해야 한다.

① 대략 한 대략 한 한 시간 정도 (바른 주석: 대략 한 한 시간 정도/TI)

5.3.3. 개체명 누락 오류

1) CV 누락 오류: 다음 예에서 '에코백'은 가방의 하위 유형이다.

① 에코백(**에코백/CV**) 요새 많이 매잖아요

2) TM 누락 오류: 특정한 패턴이나 모양은 TM으로 주석한다.

① 그거 눌러주고. 하트.(**하트/TM**)

3) DT 누락 오류: 다음 예에서 '당일'은 DT_DAY에 해당한다.

① 당일에는(**당일/DT**) 교통·숙박·식당 대란이 발생했다.

4) EV 누락 오류: 특정 사건은 EV로 태깅한다.

① 대규모 촛불집회가(**촛불집회/EV**) 전국에서 열리는데

5.4. 상호참조 해결

상호참조 해결 말뭉치에서 발견된 오류 유형은 다음과 같다.

5.4.1. 그룹 분리 오류

하나의 그룹으로 처리해야 하는 멘션들을 서로 다른 그룹에 속하도록 태깅한 오류이다.

** 구분선(---------) 기준 위의 멘션들과 아래 멘션들이 서로 다른 그룹에 속해 있음*

①

그래서 [남아공] 최후의 승자는 펠레가 한 번도 언급하지 않은 네덜란드일 것이라는 우스갯소리가 나온다.

이런 추세대로라면 [남아공 대회] 타이틀은 남미국가에 돌아갈 차례이다.

②

중국 양회<[정협]·전인대 > 최대화두 /"사회 불안을 막아라"

[중국의 최고 자문기구인 전국인민정치협상회의](정협)와 대표기구인 전국인민대표대회(전인대)가 각각 3일과 5일 개막한다.

[중국의 최고 자문기구]인 전국인민정치협상회의(정협)와 대표기구인 전국인민대표대회(전인대)가 각각 3일과 5일 개막한다.

5.4.2. 그룹 통합 오류

서로 다른 그룹으로 분리해야 하는 멘션들을 하나의 그룹에 속하도록 태깅한 오류이다.

① '군산시'와 '야미도'의 멘션이 한 그룹에 속한 경우

[새만금 사업지구인 전북 군산시 야미도] 앞 바다에서 고려 청자를 비롯한 각종 옛 도자기들이 무더기로 나왔다.

새만금 사업지구인 [전북 군산시] 야미도 앞 바다에서 고려 청자를 비롯한 각종 옛 도자기들이 무더기로 나왔다.

국립해양문화재연구소는 지난해 9월 이후 [야미도] 해역에 대한 수중발굴 조사 결과, 옛 도자기 2293점을 찾아내 건져올렸다고 4일 밝혔다.

② '2002 월드컵', '2006 월드컵' '2010 월드컵'의 멘션이 한 그룹에 속한 경우

드디어 [월드컵]이다.

이번 [월드컵]에서 사람들은 어떤 응원가를 목이 터져라 부르게 될까?

[2002년 한·일 월드컵] 서울시청 앞에서 가장 많이 울려 퍼졌던 노래는 역시 '오 필승 코리아'다.

이후 윤도현 밴드는 국민 가수 반열에 올랐고, 이 노래는 지금도 [월드컵] 응원가 하면 가장 먼저 떠오를 만큼 확고한 대표성을 지니게 됐다.

[2006년 독일 월드컵] 4년 전 최고의 응원가였던 '오 필승 코리아'가 사라지는 기현상이 벌어졌다.

이동통신사들의 [월드컵] 마케팅 싸움 때문이었다.

[2010년 남아공 월드컵] 4년 전에도 그랬지만, 이번에는 당시와 비교가 안 될 정도로

많은 응원가들이 경쟁적으로 쏟아지고 있다.

노브레인의 '대한의 전사들이여', 레이지본의 '우린 모두 챔피언', 카라의 '위 아 위드 유', 티아라의 '위 아 더 원', 슈퍼주니어의 '빅토리 코리아', 노라조의 '자블라니 잡아라', 투에이엠(2AM)의 '넘버원', 박현빈의 '앗 뜨거 [월드컵]', 컬투와 캔의 '나는 대한민국이다' 등 날마다 쏟아지는 응원가들은 이루 다 셀 수 없을 정도다.

'오 필승 코리아'처럼 자생적으로 태어난 응원가와 달리 홍보·마케팅 효과를 노리는 기업과 [월드컵] 인기에 편승하려는 가수가 인위적으로 만든 응원가가 바람직한지에 대한 의문도 나온다.

5.4.3. 어절 내 멘션 추출 오류

한 어절 내에서 멘션을 분리하여 추출한 오류이다.

①

[송파구청]장 등은 지난 2월 박원순 서울시장, 3월 김상범 행정1부시장, 지난달 중순 문승국 행정2부시장을 만나 청사 이전 계획을 설명해왔다.

청사 이전 계획은 지난해 9월께 롯데물산이 지상 123층짜리 롯데슈퍼타워 건립을 위해 맞붙은 [송파구청]사 부지 매입 의사를 밝히면서 구체화해왔다.

②

예. [한국]사가 이제 에~

영어로 읽어보며는 아 이 [한국]문화가 그게 아니다.

5.4.4. [전], [후] 중심어 추출 오류

[전]과 [후]가 중심어로 있는 멘션을 추출한 오류이다.

① [11년 전] 비슷한 사건이 있었기 때문이다.

② 2006년 독일 월드컵 [4년 전] 최고의 응원가였던 '오 필승 코리아'가 사라지는 기현상이 벌어졌다.

5.4.5. 부사어+명사구 추출 오류

명사구의 범위를 넘어 부사어와 명사구로 이뤄진 멘션을 추출한 오류이다.

① [대부분 노인]으로 양파·마늘·벼 농사를 짓고 산다.

② [11년 전 비슷한 사건]이 있었기 때문이다.

5.4.6. 구체적 정보 없는 시간 명사 추출 오류

구체적인 정보가 없는 시간 명사끼리 상호참조하도록 태깅한 오류이다.

①

[오늘]은 뭐가 문제인지 일단 짚어봤습니다.

[오늘] 말씀 여기까지 듣고 다시 한 번 또 연결하겠습니다.

[오늘] 말씀 감사합니다.

네. [오늘] 제목을 딱 봤는데 오늘 재미 있는 주제네요.

네. 오늘 제목을 딱 봤는데 [오늘] 재미 있는 주제네요.

과학으로 입증되어진. 참 네 [오늘].

아 재밌었습니다 [오늘].

저는 요만하고 물러가야겠습니다 [오늘]은.

5.4.7. 단일 멘션 그룹 형성 오류

단일 멘션으로 상호참조 그룹이 형성되어 있는 오류이다.

① '[나]'와 '타자(他者)'가 주체성을 지키며 맺는 관계가 '사랑'이며 타자의 사랑에 의해 내 존재는 정당화된다.

② 특히 <제빵왕 김탁구>는 [서인숙](전인화)이 아들 마준(주원)을 위해 김탁구(윤시윤)를 원양어선에 팔아넘기는 등 내 자식을 위해서라면 다른 사람의 자식은 죽어도 된다는 설정이 습관처럼 등장한다고 분석했다.

③ 특히 "박 대통령은 아베 총리를 만나 (위안부 문제 등에 대한) [그의 생각]을 하나하나 확인하고 그걸 국민에게 설명해 줄 필요가 있다.

④ 당의 한 3선 의원은 "내년 총선을 앞두고 '의원 유권자들'이 철저히 자기이해에 따른 전략적 투표를 고민하고 있다"며 "눈 따로, 입 따로, 손 따로 움직이는 게 원내대표 선거라지만 이번처럼 결과를 예측하기 힘든 건 [처음]"이라고 말했다.

5.5. 구문 분석

구문 분석 말뭉치에서 발견된 오류의 유형은 다음과 같다.

5.5.1. 인용표지('-고') 결합 어절의 CMP 분석 오류

인용표지 ‘고’ 결합 어절이 CMP로 분석되려면 <표준국어대사전>이나 <우리말샘>의 구문틀에 ‘-고’를 가진 서술어에 의존하는 경우라야 한다. 그러나 구문틀에 ‘-고’를 갖지 않은 서술어에 의존하는 경우에도 CMP로 분석되고(①), 구문틀에 ‘-고’를 가진 경우라도 CMP가 부착되지 않는 오류가 있었다(②).

①
낸다고(VP_CMP) 걱정하면
아니라고(VP_CMP) 걸어봅니다.
사례”라고(NP_CMP) 격려한

②
한다”고(VP) 강조했다고
달라고(VP) 건의했다고
것”이라고(VP) 말했다.

5.5.2. 기호를 포함한 어절의 구문태그 분석 오류

한 어절이 중간에 기호를 가진 경우 기호의 앞부분과 뒷부분에 따라 각각 구문태그, 기능태그를 부여해야 하는데, 기호가 없는 경우와 마찬가지로 분석한 오류이다.

①
것”이라고(VNP_CMP) 말했다.
의문”이라고(VNP_CMP) 감싼
’편견’이었다고(VNP_CMP) 느끼고
것”이라며(VNP) 말했다.

5.5.3 명사구의 기능태그 누락

문장에서 명사구가 서술어에 의존하는 경우 서술어의 구문틀에 의해 주어, 목적어, 부사어, 보어로 분석되어야 하는데 기능태그 분석이 누락되어 있는 오류이다. 특히 조사가 없는 경우에 명사구의 기능태그 분석 누락이 빈번하다.

** 아래 붉은색 줄이 수정 전, 초록색 줄이 수정 후의 모습임.*

```
  NWRW1800000021-0020-00005-00001_019  19  뉴욕      20  NP
  NWRW1800000021-0020-00005-00001_020  20  증시에서  23  NP_AJT
  NWRW1800000021-0020-00005-00001_021  21  주가가    23  NP_SBJ
- NWRW1800000021-0020-00005-00001_022  22  23.2%나   23  NP      IssueBareNP(주가가)(▶23.2%나◀)(폭락했다.);
+ NWRW1800000021-0020-00005-00001_022  22  23.2%나   23  NP_AJT  IssueBareNP(주가가)(▶23.2%나◀)(폭락했다.);
  NWRW1800000021-0020-00005-00001_023  23  폭락했다.  -1  VP
  NWRW1800000021-0020-00008-00001_001  1   14일      3   NP_AJT
  NWRW1800000021-0020-00008-00001_002  2   월스트리트저널  3  NP
```

```
  NWRW1800000021-0022-00009-00001_001  1  그는      7  NP_SBJ
  NWRW1800000021-0022-00009-00001_002  2  또래들과는  3  NP_AJT
  NWRW1800000021-0022-00009-00001_003  3  달리      7  AP
- NWRW1800000021-0022-00009-00001_004  4  TV        5  NP      IssueBareNP(▶TV◀)(볼);
+ NWRW1800000021-0022-00009-00001_004  4  TV        5  NP_OBJ  IssueBareNP(▶TV◀)(볼);
  NWRW1800000021-0022-00009-00001_005  5  볼        6  VP_MOD
  NWRW1800000021-0022-00009-00001_006  6  시간도    7  NP_SBJ  IssueDoubleSubject(그는)(▶시간도◀)(없다.);
  NWRW1800000021-0022-00009-00001_007  7  없다.     -1 VP
```

```
  NWRW1800000021-0023-00003-00001_002  2  영화배우  4  NP
  NWRW1800000021-0023-00003-00001_003  3  톰        4  NP
  NWRW1800000021-0023-00003-00001_004  4  크루즈(47)는  19  NP_SBJ
- NWRW1800000021-0023-00003-00001_005  5  18일      9  NP      IssueBareNP(▶18일◀)(열린);
+ NWRW1800000021-0023-00003-00001_005  5  18일      9  NP_AJT  IssueBareNP(▶18일◀)(열린);
  NWRW1800000021-0023-00003-00001_006  6  서울      7  NP
  NWRW1800000021-0023-00003-00001_007  7  용산구    8  NP
  NWRW1800000021-0023-00003-00001_008  8  그랜드하얏트호텔에서  9  NP_AJT
  NWRW1800000021-0023-00003-00001_009  9  열린      10  VP_MOD
  NWRW1800000021-0023-00003-00001_010  10  기자간담회에서  19  NP_AJT
```

5.5.4. 이중주어문, 이중목적어문 분석에서 보조사 결합 어절의 주어, 목적어 분석 오류

보조사('는', '도', '만' 등) 결합 어절의 기능태그 오분석 비율이 상대적으로 높다. 대체로 주어나 목적어로 분석되는데 서술어의 구문틀에 비추어 보면 부사어로 분석해야 하는 경우가 상당히 많다.

** 아래 붉은색 줄이 수정 전, 초록색 줄이 수정 후의 모습임.*

```
  NWRW1800000021-0220-00009-00001_006  6   사거리    8   NP
  NWRW1800000021-0220-00009-00001_007  7   500km     8   NP
  NWRW1800000021-0220-00009-00001_008  8   이상의    9   NP_MOD
- NWRW1800000021-0220-00009-00001_009  9   탄도미사일은  12  NP_SBJ  IssueDoubleSubject(한국군은)(▶탄도미사일은◀)(개발할);
+ NWRW1800000021-0220-00009-00001_009  9   탄도미사일은  12  NP_OBJ  IssueDoubleSubject(한국군은)(▶탄도미사일은◀)(개발할);
  NWRW1800000021-0220-00009-00001_010  10  6개월     11  NP
  NWRW1800000021-0220-00009-00001_011  11  내에      12  NP_AJT
  NWRW1800000021-0220-00009-00001_012  12  개발할    13  VP_MOD
  NWRW1800000021-0220-00009-00001_013  13  수        14  NP_SBJ
  NWRW1800000021-0220-00009-00001_014  14  있고      22  VP
```

```
- NWRW1800000021-0302-00004-00003_001  1   요새는   12  NP_SBJ
+ NWRW1800000021-0302-00004-00003_001  1   요새는   12  NP_AJT
  NWRW1800000021-0302-00004-00003_002  2   비행기에 3   NP_AJT
  NWRW1800000021-0302-00004-00003_003  3   가지고   4   VP
  NWRW1800000021-0302-00004-00003_004  4   들어가기 5   VP_SBJ
  NWRW1800000021-0302-00004-00003_005  5   편하도록 8   VP_AJT
  NWRW1800000021-0302-00004-00003_006  6   홍삼액   7   NP
  NWRW1800000021-0302-00004-00003_007  7   분말을   8   NP_OBJ
  NWRW1800000021-0302-00004-00003_008  8   뭉쳐     10  VP
  NWRW1800000021-0302-00004-00003_009  9   은단처럼 10  NP_AJT
  NWRW1800000021-0302-00004-00003_010  10  만든     11  VP_MOD
  NWRW1800000021-0302-00004-00003_011  11  제품도   12  NP_SBJ  IssueDoubleSubject(요새는)(▶제품도◀)(나와);
  NWRW1800000021-0302-00004-00003_012  12  나와     13  VP
  NWRW1800000021-0302-00004-00003_013  13  있다.    -1  VP

  NWRW1800000021-0331-00004-00001_022  22  법인의   23  NP_MOD
  NWRW1800000021-0331-00004-00001_023  23  자산을   24  NP_OBJ
  NWRW1800000021-0331-00004-00001_024  24  동결하는 25  VP_MOD
- NWRW1800000021-0331-00004-00001_025  25  제도도   26  NP_SBJ  IssueDoubleSubject(금융위는)(▶제도도◀)(검토하고);
+ NWRW1800000021-0331-00004-00001_025  25  제도도   26  NP_OBJ  IssueDoubleSubject(금융위는)(▶제도도◀)(검토하고);
  NWRW1800000021-0331-00004-00001_026  26  검토하고 27  VP
  NWRW1800000021-0331-00004-00001_027  27  있다.    -1  VP
```

5.5.5. 최상위 지배소 분석 오류

문장의 가장 마지막 어절이 언제나 최상위 지배소이며 지배소 값으로 '-1'을 가지는데 이러한 최상위 지배소는 한 문장에 단 하나만 나타나야 한다. 그러나 일부 경우에 '-1'을 두 개 이상 가진 경우가 있다.

** 초록색 줄이 문장 중간에 나타난 지배소 값 '-1'을 가진 어절*

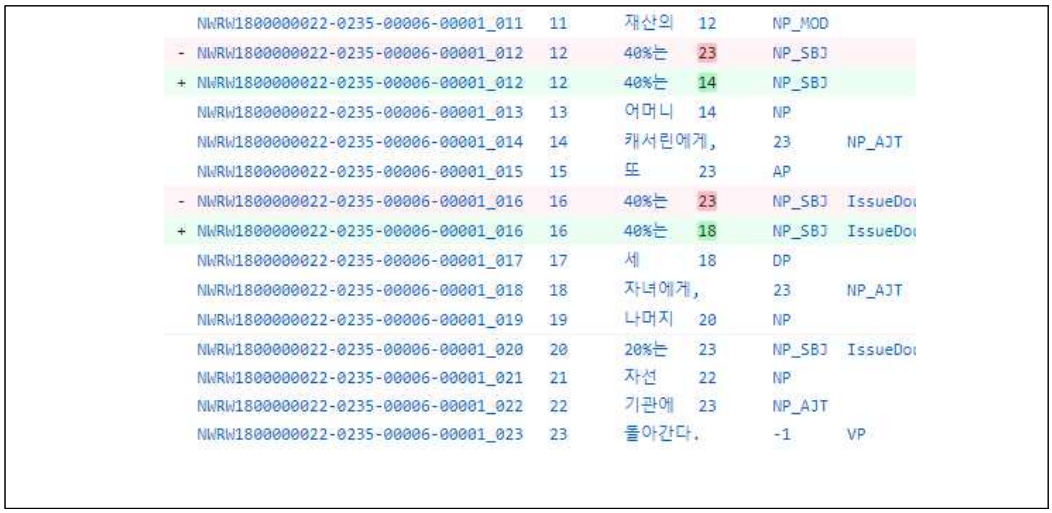
```
NWRW1800000021-0219-00008-00004_0011  민주노총은      NP_SBJ   5   ↓
NWRW1800000021-0219-00008-00004_0022  지난달        NP       3   ↓
NWRW1800000021-0219-00008-00004_0033  말           NP       4   ↓
NWRW1800000021-0219-00008-00004_0044  보도자료를     NP_OBJ   5   ↓
NWRW1800000021-0219-00008-00004_0055  통해         VP       -1  ErrorDPHead();↓
NWRW1800000021-0219-00008-00004_0066  "한국노동사회연구소의 NP_MOD  10  ↓
NWRW1800000021-0219-00008-00004_0077  '비정규직법      NP       8   ↓
NWRW1800000021-0219-00008-00004_0088  시행         NP       9   ↓
NWRW1800000021-0219-00008-00004_0099  2년          NP       10  ↓
NWRW1800000021-0219-00008-00004_01010 분석보고서'에    NP_AJT   11  ↓
NWRW1800000021-0219-00008-00004_01111 따르면        VP       22  ↓
NWRW1800000021-0219-00008-00004_01212 법           NP       13  ↓
NWRW1800000021-0219-00008-00004_01313 시행         NP       14  ↓
NWRW1800000021-0219-00008-00004_01414 기간인        VNP_MOD  17  ↓
NWRW1800000021-0219-00008-00004_01515 2007년       NP       16  ↓
NWRW1800000021-0219-00008-00004_01616 7월~2008년   NP       17  ↓
NWRW1800000021-0219-00008-00004_01717 8월          NP_AJT   22  ↓
NWRW1800000021-0219-00008-00004_01818 기간제        NP       19  ↓
NWRW1800000021-0219-00008-00004_01919 노동자가       NP_SBJ   22  ↓
NWRW1800000021-0219-00008-00004_02020 25만         NP       21  ↓
NWRW1800000021-0219-00008-00004_02121 명           NP       22  ↓
NWRW1800000021-0219-00008-00004_02222 감소했다.      VP       -1  ErrorDPHead();↓
```

NWRW1800000030-0334-00001-00001_0011	[서울시장	NP	2	↓
NWRW1800000030-0334-00001-00001_0022	후보	NP_OBJ	3	↓
NWRW1800000030-0334-00001-00001_0033	찾기	VP	4	↓
NWRW1800000030-0334-00001-00001_0044	…	X	7	↓
NWRW1800000030-0334-00001-00001_0055	한나라당	NP	6	↓
NWRW1800000030-0334-00001-00001_0066	내부	NP	7	↓
NWRW1800000030-0334-00001-00001_0077	격론]	NP	-1	ErrorDPHead();↓
NWRW1800000030-0334-00001-00001_0088	親朴	NP	-1	ErrorDPHead();↓
NWRW1800000030-0334-00001-00001_0099	"이기태(애니콜	NP	17	↓
NWRW1800000030-0334-00001-00001_01010	신화	NP	13	↓
NWRW1800000030-0334-00001-00001_01111	前	DP	12	↓
NWRW1800000030-0334-00001-00001_01212	삼성전자	NP	13	↓
NWRW1800000030-0334-00001-00001_01313	부회장)·황창규(반도체	NP	14	↓
NWRW1800000030-0334-00001-00001_01414	신화	NP	17	↓
NWRW1800000030-0334-00001-00001_01515	前	DP	16	↓
NWRW1800000030-0334-00001-00001_01616	삼성전자	NP	17	↓
NWRW1800000030-0334-00001-00001_01717	사장)	NP_AJT	18	↓
NWRW1800000030-0334-00001-00001_01818	같은	VP_MOD	19	↓
NWRW1800000030-0334-00001-00001_01919	기업인	NP_OBJ	20	↓
NWRW1800000030-0334-00001-00001_02020	찾아보자"	VP	-1	ErrorDPHead();↓

NWRW1800000049-0217-00006-00003_0011	인터파크는	NP_SBJ	-1	ErrorDPHead();↓
NWRW1800000049-0217-00006-00003_0022	“개인정보	NP	3	↓
NWRW1800000049-0217-00006-00003_0033	보호	NP	4	↓
NWRW1800000049-0217-00006-00003_0044	조치	NP	5	↓
NWRW1800000049-0217-00006-00003_0055	의무를	NP_OBJ	7	↓
NWRW1800000049-0217-00006-00003_0066	일부	NP	7	↓
NWRW1800000049-0217-00006-00003_0077	위반했기	VP	8	↓
NWRW1800000049-0217-00006-00003_0088	때문에	NP_AJT	11	↓
NWRW1800000049-0217-00006-00003_0099	유출	NP	10	↓
NWRW1800000049-0217-00006-00003_01010	사건이	NP_SBJ	11	↓
NWRW1800000049-0217-00006-00003_01111	일어났다고	VP	12	↓
NWRW1800000049-0217-00006-00003_01212	단정하기	VP_SBJ	13	↓
NWRW1800000049-0217-00006-00003_01313	어렵고.	VP	19	↓
NWRW1800000049-0217-00006-00003_01414	주민등록번호·금융정보	NP	15	↓
NWRW1800000049-0217-00006-00003_01515	등	NP	18	↓
NWRW1800000049-0217-00006-00003_01616	가장	AP	17	↓
NWRW1800000049-0217-00006-00003_01717	민감한	VP_MOD	18	↓
NWRW1800000049-0217-00006-00003_01818	정보는	NP_SBJ	19	↓
NWRW1800000049-0217-00006-00003_01919	유출되지	VP	20	↓
NWRW1800000049-0217-00006-00003_02020	않았다.	VP	-1	ErrorDPHead();↓

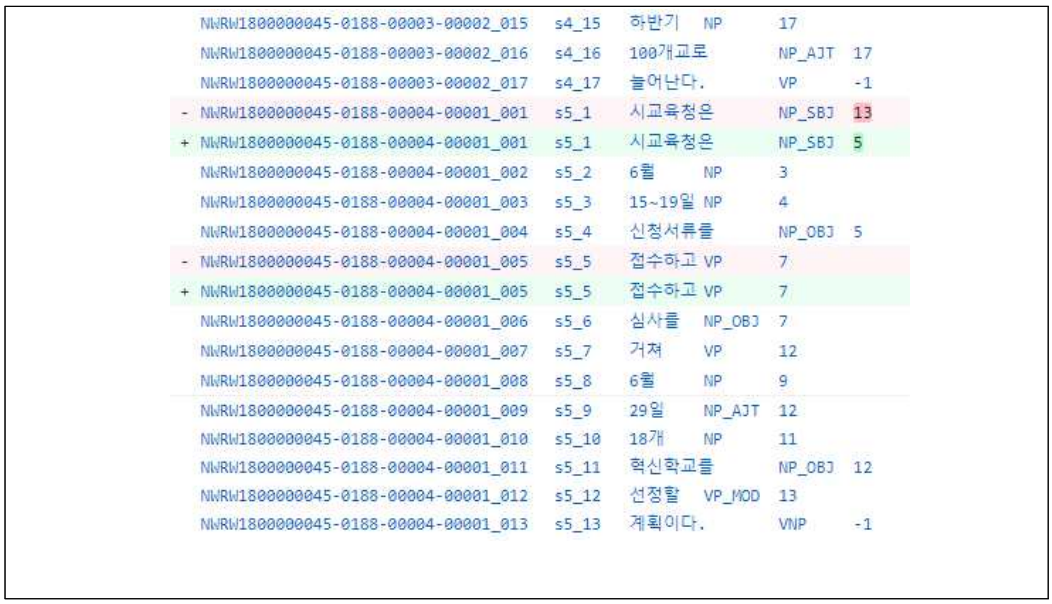
5.5.6. 지배 관계 분석 오류

피지배소-지배소 관계 분석 오류가 다수 존재한다.



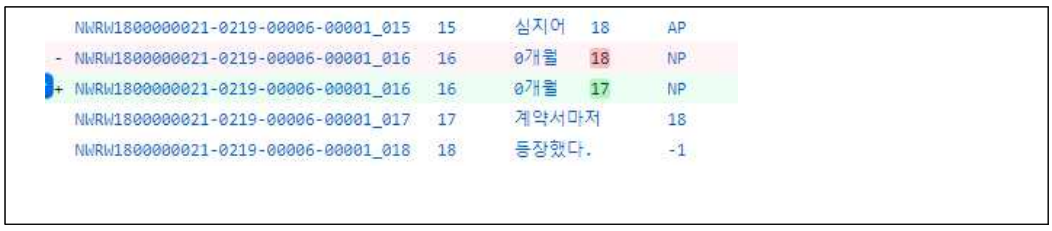
```
  NWRW1800000022-0235-00006-00001_011  11  재산의  12  NP_MOD
- NWRW1800000022-0235-00006-00001_012  12  40%는  23  NP_SBJ
+ NWRW1800000022-0235-00006-00001_012  12  40%는  14  NP_SBJ
  NWRW1800000022-0235-00006-00001_013  13  어머니  14  NP
  NWRW1800000022-0235-00006-00001_014  14  캐서린에게,  23  NP_AJT
  NWRW1800000022-0235-00006-00001_015  15  또  23  AP
- NWRW1800000022-0235-00006-00001_016  16  40%는  23  NP_SBJ  IssueDo
+ NWRW1800000022-0235-00006-00001_016  16  40%는  18  NP_SBJ  IssueDo
  NWRW1800000022-0235-00006-00001_017  17  세  18  DP
  NWRW1800000022-0235-00006-00001_018  18  자녀에게,  23  NP_AJT
  NWRW1800000022-0235-00006-00001_019  19  나머지  20  NP
  NWRW1800000022-0235-00006-00001_020  20  20%는  23  NP_SBJ  IssueDo
  NWRW1800000022-0235-00006-00001_021  21  자선  22  NP
  NWRW1800000022-0235-00006-00001_022  22  기관에  23  NP_AJT
  NWRW1800000022-0235-00006-00001_023  23  돌아간다.  -1  VP
```

<그림 30> 서술어 생략문에서의 분석 오류



```
  NWRW1800000045-0188-00003-00002_015  s4_15  하반기  NP  17
  NWRW1800000045-0188-00003-00002_016  s4_16  100개교로  NP_AJT  17
  NWRW1800000045-0188-00003-00002_017  s4_17  늘어난다.  VP  -1
- NWRW1800000045-0188-00004-00001_001  s5_1  시교육청은  NP_SBJ  13
+ NWRW1800000045-0188-00004-00001_001  s5_1  시교육청은  NP_SBJ  5
  NWRW1800000045-0188-00004-00001_002  s5_2  6월  NP  3
  NWRW1800000045-0188-00004-00001_003  s5_3  15~19일  NP  4
  NWRW1800000045-0188-00004-00001_004  s5_4  신청서류를  NP_OBJ  5
- NWRW1800000045-0188-00004-00001_005  s5_5  접수하고  VP  7
+ NWRW1800000045-0188-00004-00001_005  s5_5  접수하고  VP  7
  NWRW1800000045-0188-00004-00001_006  s5_6  심사를  NP_OBJ  7
  NWRW1800000045-0188-00004-00001_007  s5_7  거쳐  VP  12
  NWRW1800000045-0188-00004-00001_008  s5_8  6월  NP  9
  NWRW1800000045-0188-00004-00001_009  s5_9  29일  NP_AJT  12
  NWRW1800000045-0188-00004-00001_010  s5_10  18개  NP  11
  NWRW1800000045-0188-00004-00001_011  s5_11  혁신학교를  NP_OBJ  12
  NWRW1800000045-0188-00004-00001_012  s5_12  선정할  VP_MOD  13
  NWRW1800000045-0188-00004-00001_013  s5_13  계획이다.  VNP  -1
```

<그림 31> 부사절에서의 분석 오류



```
  NWRW1800000021-0219-00006-00001_015  15  심지어  18  AP
- NWRW1800000021-0219-00006-00001_016  16  0개월  18  NP
+ NWRW1800000021-0219-00006-00001_016  16  0개월  17  NP
  NWRW1800000021-0219-00006-00001_017  17  계약서마저  18
  NWRW1800000021-0219-00006-00001_018  18  등장했다.  -1
```

<그림 32> 나열 명사구에서의 분석 오류

<그림 33> 명사구 접속에서의 분석 오류

5.6. 의미역

의미역 말뭉치에서 발견된 오류의 유형은 다음과 같다.

5.6.1. 주석 누락 오류

필수 논항 또는 부가역을 주석하지 않고 누락한 오류이다.

① 유럽 업체들은(ARG0) 과거 미국 시장을 공략했다가
② 미국의 대규모 공적자금 투입과 각국의 유동성 공급으로(ARGM-CAU) 한동안 안정을 되찾았던

5.6.2. 논항 주석 오류

필수 논항의 논항 번호 또는 부가역의 부가어 주석을 잘못한 오류이다.

① 그는(ARG1 → ARG0) 종합 144.62점을 얻어 2위 이준형(능내초·117.56점)을 큰 점수 차로 제쳤다.
② “다른 어학연수 프로그램에 비해(ARGM-EXT → ARGM-CND) 인턴을 병행할 수 있다는

5.6.3. 부가어 중복 주석 오류

동일한 부가어가 서로 다른 서술어에 연결되어 주석된 오류이다.

① 6월에는(ARGM-TMP) 136조4430억 원(14.6%)에 이르러 해당 기업에 대한 발언권은 그만큼 커졌다. (‘커졌다’와의 연결을 제거해야 함)

5.6.4. 논항 과잉 주석 오류

보조 용언, 의사 보조 용언, 서술어 배제 리스트에 속한 용언에 논항이 주석된 오류이다.

① 승객과 승무원 등 18명은 3개의 구명보트를 타고 표류하다 부근을 지나던 어선에(ARG2) 의해 구조됐으나 나머지 탑승객의 생사는 확인되지 않고 있다.
② 결과적으로 예전보다 더 행복하게 사는 것(ARG1) 같다"고 말했다.

5.6.5. 서술어 과잉 주석 오류

보조 용언, 의사 보조 용언, 서술어 배제 리스트에 속한 용언이 서술어로 주석된 오류이다.

① 출시한 새로운 상품에 대해 안내드리려고 하는데요….”(하__4444401)
② 교원단체의 반발을 사고 있는 시간선택제 교사 제도는 새로 선발하는 것이 아니라 재직 중인 교사가 전환하는 방식을 도입하기로 했다.(하__4444402)
③ 서울시가 펴낸 <서울교통사>에 따르면(따르__4444401) “일본에서 강판을 제작해 현장에선 볼팅·가설만 하던” 시기였다.

5.6.7. 그 외

그 외에 서술어 sense 번호 오류, 서술어 sense 번호 누락 오류, 비서술어가 서술어로 주석된 오류, 명사구 범위 지정 오류, 논항 형태(argument form) 오류, 서술어 형태(predicate form) 오류 등이 발견되었다. 아울러 의미역 말뭉치 구축 지침에는 서술어의 sense 번호와 격틀 구조에 대하여 Korean Propbank, ETRI, Upropbank, 우리말샘을 참고하게 되어 있었으나, 본 과업의 업무 지시에서는 Upropbank의 사용을 배제하고 Korean Propbank, ETRI, 우리말샘만을 사용하도록 하였으므로 Upropbank 기준으로 주석된 서술어와 격틀은 모두 오류로 간주되어 우리말샘 기준의 서술어 번호와 격틀로 수정되었다.

5.7. 주격 무형 대용어 복원

주격 무형 대용어 복원 말뭉치에서 발견된 오류의 유형은 다음과 같다.

5.7.1. 인용문 주어 불일치 오류

인용문 내부에 있는 서술어의 주어를 인용문 외부의 주어로 복원한 오류이다. 이 경우 인용문 내부의 주어로 복원대상을 수정한다.

** 변경사항을 [-삭제된내용-] {+추가된내용+}으로 표시함*

①

s7_8 딸을
s7_9 그리워했는지
s7_10 알고 [-할머니__@s6_10-]{+나__@s7_3+}
s7_11 싶다"며
s7_12 "죽기 [-할머니__@s6_10-]{+나__@s7_3+}
s7_13 전에
s7_14 북한을
s7_15 방문해 [-할머니__@s6_10-]{+나__@s7_3+}

5.7.2. '누군가', '무언가' 미복원 오류

문서 내에서 복원 가능한 선행어가 없을 때 '누군가' 또는 '무언가'로 복원해야 하지만 복원되지 않은 오류이다.

** 변경사항을 [-삭제된내용-] {+추가된내용+}으로 표시함*

①

s8_3 "대통령으로서
s8_4 무슨
s8_5 정책을
s8_6 내놔도 {+누군가__@-1+}
s8_7 계속
s8_8 반대만
s8_9 하는 사람__@s8_10
s8_10 사람을

s8_11 보면서 누군가__@-1
s8_12 참으로
s8_13 답답함을 {+무언가__@-1+}
s8_14 느낄 누군가__@-1
s8_15 때가
s8_16 있다"면서

5.7.3. 관형사형 주어 미복원 오류

지침에 따라 관형사형의 주어를 복원해야 하는데 복원되지 않은 오류이다.

①
s9_9 재판부가
s9_10 평결과
s9_11 다른 {+판결__@s9_12+}
s9_12 판결을
s9_13 할
s9_14 수도
s9_15 있으나,

5.7.4. 특정 구문 복원 오류

특정 구문 표현의 주어 복원이 잘못된 오류이다.

1) '-기로 하다'에서 '하다'의 주어가 복원되지 않은 오류
①
s7_1 BMW도
s7_2 새로운 1시리즈__@s7_3
s7_3 '1시리즈'를
s7_4 미국
s7_5 시장에
s7_6 최초로
s7_7 선보이고
s7_8 깜찍하고 차__@s7_10

s7_9 귀여운 차__@s7_10
s7_10 차로
s7_11 인기를
s7_12 끌고 미니__@s7_14
s7_13 있는
s7_14 '미니'의
s7_15 대리점을
s7_16 10여
s7_17 곳
s7_18 늘리기로 BMW__@s7_1
s7_19 했다. {+BMW__@s7_1+}

2) '-ㄴ것으로 보이다/알려지다/나타나다'에서 내포문 '-ㄴ'의 주어가 복원되지 않은 경우. 이때 내포문의 주어를 복원하고, 복원된 서술어의 주어는 삭제해야 한다.

①
s3_20 가능성이
s3_21 높은 {+가능성__@s3_20+}
s3_22 것으로
s3_23 보인다.
s4_1 국토부의

②
s3_3 상당수가
s3_4 완치
s3_5 이후에도
s3_6 정신적
s3_7 고통을
s3_8 호소하는 {+상당수__@s3_3+}
s3_9 것으로
s3_10 나타났다.

5.7.5. 구어 선행어 복원 순서 오류

문서 내 가장 가까운 선행어(전방+후방 조응)로 복원한다는 지침에 어긋나게 복원된 오류이다.

①

s13_1P2	저희	
s13_2P2	많이	
s13_3P2	드시는	
s13_4P2	소고기	
s13_5P2	돼지고기	
s13_6P2	닭고기	
s13_7P2	이용해서	[-저__@s115_2-]{+저__@s34_1+}
s13_8P2	고기	
s13_9P2	요리하고요.	[-저__@s115_2-] {+저__@s34_1+}

5.7.6. 구어 담화 인칭 불일치 오류

구어에서 화자간의 대화 중 인칭이 불일치한 오류이다.

①

s284_1	P2	이케	
s284_2	P2	불을	
s284_3	P2	끄고	[-저희__@s193_1-] {+누군가__@-1+}
s284_4	P2	해주시면	[-저희__@s193_1-] {+누군가__@-1+}
s284_5	P2	훨씬	
s284_6	P2	좋고요. 멸치__@s283_6	

6. 분석 말뭉치 구축 지침 보완 제언

6.1. 형태 분석

6.1.1. 고유명사 처리 지침

2019년 형태 분석 말뭉치를 검토한 결과 고유명사의 처리에 있어서 분석가 간의 차이가 나타나거나 유사한 언어 표현이 서로 다른 대우를 받는 경우가 상대적으로 많았다. 고유명사의 외연을 정확히 한정하는 것은 지난한 일이지만, 작업의 일관성과 효율성 확

보를 위해서는 구축 지침을 보완할 필요가 있다고 여겨진다.

1) 인명 관련

고유명사가 일반명사로 쓰이게 되는 경우, 형태 분석의 차원에서는 이를 어떻게 처리할지에 대한 지침이 보완될 필요가 있다고 여겨진다. 가령 '슈퍼맨'의 경우 고유명사로도 쓰이지만 능력이 뛰어난 사람을 나타내는 말로도 쓰이는데, '슈퍼맨이 돌아왔다'의 분석에서 '슈퍼맨'을 고유명사로 처리한 것이 있었다. 이와 같은 처리도 가능함을 인정할 때, 고유명사와 일반명사 양자의 의미를 지닌 표현에 대한 처리 방식을 명시할 필요가 있다고 생각된다. 가령, "'우리말샘'에서 일반명사로 등재된 항목이 해당 의미로 쓰인다면 일반명사로 처리한다. 인명 이외의 고유명사도 마찬가지 방식으로 처리한다."와 같은 조항을 넣을 수 있다고 여겨진다.

분석 결과 중 "'빅보이' 이대호"에서 '빅보이'를 고유명사로 처리한 경우가 있었다. 그런데 '빅보이'는 그 자체로는 고유명사라 보기 어려운 표현이다. 지침 중 혼성어와 관련하여 별명의 처리 방식이 나와 있기는 하나, 단순히 지시적 속성만을 가지고 고유명사로 판정할 수 없음을 밝힐 필요가 있다.

2) 지명 관련

구축 지침에서는 고유의 지명 표현 뒤에 지리적 특성을 나타내는 '내륙, 대륙, 바다, 거리' 등과 행정구역으로서의 '도', '시', '동' 혹은 지역적 특성을 나타내는 '공업지대'는 물론 '지역, 지방, 지구' 등 구역의 특성을 나타내는 말이 한 어절에 나타났을 때 전체를 고유명사로 처리한다고 밝히고 있다. 그런데 '을지로3가, 율량2지구, 석관1동' 등의 표현의 경우에도 해당 원칙이 적용되는지는 불명확하다. 분석 결과에서는 '을지로3가' 류는 대개 하나의 단어로 처리하고 '율량2지구, 석관1동'은 분절하거나 하나로 처리하는 등 차이가 나타났다. 이는 숫자의 처리 방식과 연관되는 문제이므로 처리 지침을 명시할 필요가 있다. 또 '은평3구역'이나 '은평갑'과 같은 표현들의 처리 방식도 명시한다면 작업 결과의 일관성을 높일 수 있으리라 여겨진다.

3) 단체명

고유명사 처리에 있어서 가장 많은 혼란이 나타난 부분이다. 국제기구, 시민단체, 이익단체 등의 명칭은 지침상 어느 부분에 근거하여 처리하면 되는지 뚜렷하지 않고, 고유명사 판단 기준이 일반적인 인식과 차이가 있는 듯하다. 일례로 '안전보장이사회'는 모두 일반명사로, '경제협력개발기구'는 모두 고유명사로 처리되었는데, 지침만으로는 그

근거를 알기 어렵다. 또 '참여연대'는 모두 고유명사로 처리되었고, '중소기업중앙회'는 일반명사로 처리되는 등, 지침만으로는 분석 결과를 예측하기가 어려운 경우가 많다. 그리고 기관명과 단체명, 회사명 등에 대한 고유명사 판정 기준이 다르다는 점은 고유명사 판정을 위해 해당 단체의 속성도 파악해야 한다는 부담이 따른다. 이 밖에도 고유명사의 판정에 혼동을 일으키기 쉬운 요소가 많다. 한편으로는 판단 기준을 단순화하고 다른 한편으로는 해당 단어가 우리말샘에 등재되어 있는지 여부를 판단 기준에 추가함으로써 판정 기준을 보다 단순하게 바꾸는 것을 고려할 수 있다.

4) 기타

말뭉치 중 '제3한강교'와 '제2개성공단'이 있었는데, 모두 '제3', '제2'를 분석한 것이었다. 그러나 전자가 실체를 가리키는 말이라면 후자는 비유적으로 쓰인 말로서 경우에 따라 '의'의 삽입이 가능하다. '영동6교' 등도 같은 성격이 문제라 할 수 있다. 이러한 점을 고려하여 지침을 명세한다면 오류의 가능성이 낮아질 것으로 보인다.

다음으로, 한 어절에 대해 복수의 IC분석이 가능한 경우의 분석 방식에 대해 지침에서 다루고 있는데 대등한 자격을 지닌 요소가 셋 이상 나열될 경우에도 그 지침이 적용되는지 명확하지 않다. 가령 '중·일·대만'의 분석이 '중·일/NNP + ·/SP + 대만/NNP'과 같이 이루어진 예가 있는데 이처럼 대등한 자격의 요소가 셋 이상 나열될 때에도 다른 곳에서 제시한 바와 같이 형태 분석을 하는 것이 나을지 혹은 각각을 분석하는 것이 나을지 재고의 여지가 있다. '남-북-미-중'과 같은 표현이 있을 때 기본적인 형태 분석 방식에 따르면 '남북/NNP+미중/NNP'과 같이 분석하게 되는데 국가명 약어나 기타 다른 경우에도 대등한 요소가 셋 이상 연쇄될 때에는 각각을 분석하는 방식을 취하는 것이 보다 직관에 부합하는 방식이라 여겨진다.

6.1.2. '우리말샘'에서 미명세된 용법

1) '-ㄴ지'

'-ㄴ지'의 경우 '우리말샘'에서 연결어미와 종결어미로서의 용법이 모두 인정되지만 연결어미의 용법은 "얼마나 부지런한지 세 사람 몫의 일을 해낸다."와 같이 후행절 명제 내용에 대한 근거 추정의 용법만 서술되어 있다. '-ㄴ지'는 보어절을 이끌며 쓰이는 경우가 많은데 그와 관련된 내용이 '우리말샘'에 명시적으로 서술되지 않은 것이다. 그러다 보니 보어절을 이끄는 '-ㄴ지'를 연결어미로 분석한 오류가 많이 나타난 것으로 판단된다. 이와 관련하여 지침에서 조사와의 결합 가능성이나 통사적 특성 등 판단 기준을 제시한 것은 필요한 일이다. 다만, '-ㄴ지'가 '모르다'의 목적어로 쓰이는 경우만이 아니

라 주절 서술어의 대상이 되는 경우, 모두 종결어미로 판정할 수 있음을 명시할 필요가 있다고 판단된다.

2) ‘그래도’

‘그래도’의 형태는 접속부사 혹은 ‘그러-’나 ‘그렇-’의 활용형 중 하나인데, “저희 부서가 작년에 그래도 성과를 많이 냈어요.”에서와 같이 ‘그래도’가 ‘만족스럽지는 않지만, 적어도’ 정도의 의미를 나타내며 양태 부사처럼 쓰이는 경우가 많다. 그리고 이때의 ‘그래도’는 정확히 무엇을 지시하거나 대용한다고 하기 어려운 경우가 많다. 그러다 보니 형태 분석을 어떤 식으로 할지, 그리고 동사와 형용사 중 어느 것으로 분석을 할지 고민이 되기 마련이다. 이와 관련하여 지침에서 세부 조항을 통해 분석가의 판단 기준을 제시해 주면 작업의 효율성이 향상될 것이라 여겨진다. 일반적인 접속부사와 달리 다양한 위치에 나타날 수 있다는 점을 중시하면 활용형으로 판단할 수 있을 듯하고, ‘그렇다’의 의미 중 ‘만족스럽지 않다’가 있으므로 이를 지침에서 밝혀주는 것이 좋을 것이라 생각된다.

3) ‘그래서인지’

‘그래서인지’의 경우 ‘그래서’를 접속부사로도 볼 수 있고, 활용형으로 볼 수도 있다. ‘그래서인지’는 절 경계에서도 쓰일 수 있지만 문두에서 쓰이는 경우가 많은데, 문두에 쓰인 경우 ‘그래서’의 분석에 도움이 되는 판단 기준을 찾는 것이 쉽지 않은 듯하다. 이런 경우에는 분석 작업의 효율성을 도모하여 하나의 처리 방식을 제시할 수 있을 듯하다. 지침에서 ‘그런데도’에 대한 분석 방식을 제시한 것처럼 ‘그래서인지’의 분석 방식을 제시하면 분석가의 부담을 줄여줄 수 있을 것이다.

6.2. 어휘 의미

6.2.1. 어휘 의미 번호 ‘888’의 부여 기준

어떤 사전이든 해당 언어의 모든 어휘에 대해 기술하는 것을 목표로 하지만, 실제로 그러한 목표를 이룰 수는 없다. 실생활에서 사용되는 표현들은 여러 기제로 인해 사전의 뜻과 다른 뜻으로 쓰이게 되는데, 이러한 경우 888을 주어야 할지, 아니면 기존의 의미를 주어야 할지 현재의 어휘의미 분석 지침으로는 판단하기가 어렵다. 예컨대 ‘손’이 ‘결국 손을 다 들었다’처럼 ‘포기하다’의 의미를 이루는 ‘손을 들다’의 구성 요소로 포함된다면, 이는 ‘신체 기관’인 ‘손’을 들어 포기를 표하는 행위를 비유하는 것으로 이해하여 ‘손__001/NNG (사람의 팔목 끝에 달린 부분. 손등, 손바닥, 손목으로 나뉘며 그 끝

에 다섯 개의 손가락이 있어, 무엇을 만지거나 잡거나 한다)'으로 분석해야 할지, 아니면 우리말샘에 등재되지 않은 새로운 의미를 얻었다고 판단하여 '손__888/NNG'으로 분석해야 할지 판단하기가 어렵다. 어느 정도의 비유적인 표현까지는 등재된 의미번호를 부여한다는 기준을 세우거나, 혹은 정확히 등재 의미가 아닌 경우에는 모두 사전 기술의 미비로 판단하고 888로 부여한다고 기준을 세울 필요가 있어 보인다.

예컨대 구축 지침의 세부 사항으로 "<우리말샘>에 등재된 어휘가 사전에 기술된 어휘의미로는 정확히 설명할 수 없는 비유적 표현이나 관용구의 일부로 쓰일 경우, 해당 어휘의미에는 888을 부여한다"와 같은 내용을 추가할 수 있다. 그 예문으로 앞서 든 '손을 들다'와 같은 예를 제시하여, 이와 같은 경우에 '손__888/NNG'을 부여한다고 설정할 수 있다.

6.2.2. 다의어의 판단 기준

현재 구축 지침에서는 "<우리말샘>에 제시된 갈래 뜻 사이에 구분이 불분명할 경우에는 <우리말샘>의 예문을 최대한 검토하여 확정한다. 이때 한쪽이 예문이 없는 것이라면 예문이 있는 쪽의 갈래 뜻을 선택한다. 단, 양쪽이 모두 예문이 없다면 포괄적인 갈래 뜻을 선택한다."라고 되어 있고, 후속 장에서 일반명사나 고유명사 등의 일부 판단 예를 제시하고 있다. 다만 실제 데이터를 분석해보면 판단 대상이 되는 어휘의미 양쪽 모두 예문이 있고 예문으로는 데이터의 실제 용례를 판단하기 어려운 경우가 있는데, 이러한 경우에는 거의 전적으로 분석가의 직관에 의존하고 있다. 예컨대 '가톨릭교회__001'은 '가톨릭교를 믿는 교회'이고, 예문으로 "가톨릭교회가 신도의 영성 생활에 주력해야 할 것은 물론이다" 하나를 가지고 있다. '가톨릭교회__002'는 '로마 교황을 최고 통치자로 한 교회들을 통틀어 이르는 말'이고, 예문으로 "그런 높은 첨탑건립은 특히 거의 모든 로마 가톨릭교회에서 두드러진다" 하나를 가지고 있다. 이런 경우 실제 말뭉치에서 두 의미를 명확히 구분하기는 어렵다.

이런 경우까지 포괄할 수 있는 지침을 마련할 필요가 있어 보인다. 예컨대 <2.나항>의 "… <우리말샘>의 예문을 최대한 검토하여 확정한다. …"라는 부분을 보다 정확히 윤문하여, 예컨대 "예문의 다른 명사나 서술어와의 공기 관계를 최대한 고려하여 확정한다"라고 고치거나, "예문에서 드러나는 구체성/추상성, 개별성/조직성, 일반성/특수성, 포괄성/국한성 등의 구분에 유의하여 확정한다"와 같이 고쳐서, 판단 근거를 보다 명확히 제시할 수 있을 것이다. 또한 "의미번호 기술과 예문만으로는 도저히 판단할 수 없는 경우, 번호가 빠른 의항을 선택한다"와 같이 판단이 도저히 되지 않는 경우에 취할 수 있는 기준을 마련해줄 필요도 있어 보인다.

6.3. 개체명

6.3.1. 개체명 주석 대상의 예시 추가

개체명 주석 지침은 구축 작업자로 하여금 해당 태그가 어느 정도 범위의 개체명까지 커버하는지 가늠할 수 있는 정보를 주어야 한다. 이를 위하여 각 태그 세목별로 설명과 예시를 확충할 필요가 있다. 설명과 예시 확충이 필요한 세목과 추가될 내용의 제안을 아래 보인다.

1) 동물 몸의 한 부분(신체부위) 명칭 태그에 대하여, 동물뿐만 아니라 사람의 신체부위에 대해서도 주석하도록 되어 있으므로 '인간'을 명세할 필요가 있다. '동물, 인간 몸의 한 부분(신체부위) 명칭'이라고 표현을 정정하는 것이 좋을 것이다.

2) 학문 분야 및 학파(FD)의 예시를 추가할 필요가 있다. 경찰학의 학위 분야로 '과학수사'가 학문 분야로 나타나는데 FD 태깅이 가능하다. 현재 예시는 '~학'으로 끝나는 경우이므로 좀 더 다양한 학문 분야(언어정보, 응용언어, 정보기술 등)를 예시로 추가할 필요가 있다.

3) 사회과학 이론/법칙/방법/원리/사상, 정치사상(TR)의 범위를 좀 더 명시적으로 제시할 필요가 있다. 19년 말뭉치에서 '친노', '비박', '진보', '보수' 등에 TR 태깅이 되어 있지만 이는 정치사상이라 보기 어렵다. 우리말샘 미등재어이면서 일부 정치 세력이나 무리를 일컫는 용어는 태깅하지 않도록 다음처럼 예를 제시하는 것이 좋을 것이다.

→ 진보/TR 개혁 세력에 (X) : 일부 정치적 서향을 같는 무리는 태깅하지 않음.
진보신당/OG (O) : 특정 정당을 언급함.
→ 친박/TR들이 (X), 친박신당/OG(O)

4) 도로/철로 명칭, 운하, 거리(AF) 관련 지침에는 도로명 '~로'에 대한 예시만 나와 있다. 실제로 '광화문 네거리'와 같이 '~거리'도 AF 태깅 대상이 되므로 예를 추가할 필요가 있다.

5) 무기 명칭(AF) 관련 지침에 따르면 문맥에 따라 무기로 사용되었다고 판단할 경우 AF로 태깅할 수 있는데, 이를 더 정밀하게 "문맥에 따라 무기로 사용되었다고 판단할

경우 AF로 태깅할 수 있음. 단, 무기 명칭이 아닌 문맥에서는 AF로 태깅하지 아니함.” 이라고 진술하는 것이 좋을 것이다. 아울러 다음과 같은 예를 추가할 수 있다.

예) 시퍼렇게 날이 선 칼/AF을 빼어 들고 우리를 쫓아왔다. (O)
수정과 만들 때에는 저처럼 칼/AF을 넣어서 요렇게 만듭니다. (X)

6) 조약, 협정, 협약은 OGG 계열의 관련 개체명으로 태깅한다는(대부분은 OGG_POLITICS) 지침에, 예시를 추가할 필요가 있다. 다음과 같은 예를 추가할 수 있다.

예) 자유무역협정/OG, 정책협약/OG

7) 미디어 기관/단체(OGG_MEDIA)의 매체 유형을 좀 더 다양하게 제시할 필요가 있다. 2019년도 구축 말뭉치에서는 나타나지 않았지만 유튜브(Youtube), 카카오TV 등을 OGG_MEDIA로 볼지, ‘카카오톡’, ‘트위터’와 마찬가지로 TMI_SEVICE로 볼지 검토가 필요하다.

8) “‘시, 군, 구’도 기관으로 나올 시, OG로 태깅한다. (예 : 양구군은 ... (중략) 군은→ 양구군/OG, 군/OG)”는 지침을 다음과 같이 명확화하는 것이 좋을 것이다. “‘시, 군, 구’가 독립적인 기관으로 나올 시 각각을 OG로 태깅한다. 단, ‘시, 도’ 전체를 일컬을 경우에는 하나의 개체로 태깅한다.” 다음과 같은 예를 제시할 수 있다.

예) 서울시/OG에서는 ... 시/OG에서 25개의 구/OG와 협력하여
전국 16개의 시·도 교육청/OG (O)

9) “‘명사(국가/도시/행정구역명/)+명사’ 구성에서 앞부분에 해당하는 ‘명사(국가’는 LC로 일괄 태깅한다.”는 지침에, ‘행정구역명 + 직위’에 해당하는 예시를 추가할 필요가 있다. 다음과 같은 예를 추가할 수 있을 것이다.

예) 마포구청장, 중구청창, 동대문구청장 : 마포구/LC + 청장/CV, 중구/LC + 청장, 동대문구/LC + 청장/CV
→ ‘이름(성) + 직위명’이 나타날 경우에는 CV로 태깅한다.

예) 김/PS 구청장/CV, 이/PS 시장/CV
→ 전체 기관장을 나타낼 때에는 통합하여 CV로 태깅한다.
예) 중소기업은행장/CV, 한국은행장/CV

10) 언어 명칭(CV_LANGUAGE)에는 '한자'와 같은 문자 명칭은 태깅하지 않는다는 예를 추가하는 것이 좋을 것이다. 다음과 같은 예를 추가할 수 있다.

예) 한자/CV로(X)

11) 음식/곡물 명칭, 음식재료, 음식의 유형(CV_FOOD)에, '-하다'가 결합하여 나타나는 개체명에는 태깅하지 않는다는 설명과 예를 추가할 것을 제안한다. 다음과 같은 예를 추가할 수 있다.

예) 양념/CV하다(X), 국수/CV하다(X)

그리고 음식재료, 반찬, 양념의 하나로 쓰이는 경우 하나의 개체명으로 태깅하도록 한다. 우리말샘 '다진 양념', '볶은고추장' 등재어를 참조할 수 있다. 다음과 같은 예를 추가할 수 있을 것이다.

예) 마늘/PT(X), 파/PT(X)
볶은고추장/CV(O), 볶은 마늘/CV(O), 다진 마늘/CV, 다진 파/CV(O), 다진 양념/CV(O)

12) 의복/섬유 명칭, 의복유형(CV_CLOTHING)에 모자, 가방, 양말, 신발의 하위 유형은 포함되나, 액세서리 류는 포함되지 않는다는 것을 나타내 주어야 한다. 다음과 같은 예를 추가할 수 있다.

예) 넥타이/CV(O), 구두/CV(O), 샌들/CV(O), 데님/CV(O), 에코퍼/CV(O), 에코백/CV(O)
목걸이/CV(X), 반지/CV(X)

13) CV_POSITION의 예를 추가하고, '직급 + 직위'의 경우 각각 태깅한다는 것을 잘

보여주는 예를 더할 수 있다.

예) 난민, 시민, 아동, 유학생, 에듀맘, 의료진, 후원자
예) 9급 공무원 : 9급/CV + 공무원/CV, 부장 교사 : 부장/CV + 교사/CV, 기간제/CV + 교사/CV

14) DT_DURATION에 대하여 설명과 예시를 다음과 같이 추가할 수 있다. "'월 3만 9천원'에서 '월'은 태깅하지 않는다. '총 6개월'에서 '총'을 태깅하지 않는다."

예) 총 6개월/DT

15) QT_AGE에 대하여 설명과 예시를 다음과 같이 추가할 수 있다. "문맥에서 특정일을 나타내는 것이 아니라 '나이'를 의미할 때에는 QT로 태깅한다."

예) 82년생/QT 김지영/PS
1958년 6월 21일생/QT인 이 후보자는 60세 정년인 경찰공무원법에 따라 2년 임기가 끝나기 전인 2018년 6월에 퇴직해야 한다.

16) AM_PART 태깅 대상에 대한 예시를 다음과 같이 추가할 수 있다.

→ 동물, 인간 몸의 한 부분(신체부위) 명칭
예) 치아

17) 구축 지침에 따르면 "세균", "바이러스", "박테리아" 등은 비세포성 생물이므로 AM_OTHERS 에 해당되고, 질병의 원인인 '바이러스'는 TMM_DISEASE 로 태깅한다. 지침이 모호해지는 지점을 명시적으로 설명할 필요가 있다. 현재 지침에 따르면, 총칭적인 의미로 '바이러스'가 나오면 AM으로 특정명이 같이 나오는 '코로나 바이러스/TM' 처럼 태깅해야 한다. '바이러스에 걸리다', '바이러스 확산', '세균 감염' 등과 같은 경우에는 어떻게 태깅해야 할지 검토가 필요하다.

18) TM_direction에 대해 설명과 예시를 다음과 같이 추가할 수 있다. "방향을 나타내는 경우에만 태깅함. 안, 밖, 내, 외, 곁, 옆, 위, 아래 등은 지시적인 표현으로 명시적인

방향을 나타내지 않으므로 태깅하지 아니함. 지리적으로 특정 방향을 나타내는 개체명에 한해 태깅함.”

예) 옆/TM으로 밀면(X)
앞뒤/TM로 먼저 익혀 주면(X)
오른쪽/TM, 왼쪽/TM(O)
동서남북/TM(O)

19) TM_SHAPE에 대해 설명과 예시를 다음과 같이 추가할 수 있다. “특정한 모양, 특정한 패턴, 특정한 무늬, 특정한 형태 명칭에 대해 태깅함.”

예) 하트 모양/TM, 별 모양/TM으로 만들고 (O)
체크 무늬/TM(O), 호피 무늬/TM(O)

20) 증상/증세/질병(TM_DISEASE)의 예시가 확충될 필요가 있다. 특정 증세에 한해만 태깅할지, ‘우울증’, ‘조울증’, ‘치매’, ‘비만’, ‘스트레스’ 등 일반적인 질병에 대해서 태깅할지 검토가 필요하다. 19년도 구축 말뭉치에서는 발견되지 않았다.

21) 하드웨어 용어, 전자기기에 해당하는 제품 TM_HW에 대한 설명 추가
→ ‘전화’가 하드웨어, 전자기기일 경우에 태깅하는 것인지 일상적인 전화를 이용한 통화의 행위에서까지 모두 포함하는 것인지 불분명합니다. ‘전화를 걸다’, ‘전화하다’는 미태깅으로 처리했습니다.
→ 전자기기를 나타낼 경우에 한해 태깅함.
예) 전화/TM를 걸다(X), 전화/TM하다(X)
전화/TM를 가지고 있지 않아서 연락이 안 닿았다.(O)

22) TM_SITE에 웹 사이트의 주소도 해당하는지 예시가 추가되어야 한다. 개인 블로그, 트위터 계정도 태깅 대상인지, 개인 정보와 관련하여 태깅 여부 검토되어야 한다.

예) @82kimjiyoug, youtube.com/~

22) 구어 말뭉치의 줄임 표현의 경우, “원문서에서 줄임 표현의 경우, 형태 분석 결과

를 반영하여 태깅한다.”고 지침을 수정 및 명세화하는 것이 좋다고 생각된다. 다음과 같은 예를 추가할 수 있다.

예) 오승재 기잡니다 : 오승재/PS 기자/CV
러시안데요 : 러시아/LC

6.3.2. 줄임말, 축약형 태깅에 대하여

예) ‘김변님’ : 김/NNP + 변님_777/NNG

위와 같은 대상은 개체명 층위에서는 김/PS로 처리하였다. ‘변호사’에 해당하는 ‘변’에 대해서 처리할 수 있는 지침이 마련되어 있지 않다. 언어 현실을 고려할 때, ‘김 박(사)’, ‘김 변(호사)’와 같은 형태가 빈번히 나타나므로 처리 기준이 필요하다.

6.3.3. 신조어 태깅에 대하여

‘댕댕이’, ‘유튜버’ 등에 대해서 개체명 태깅을 할지, 우리말샘 등재 여부를 고려하여 지침에 반영할 필요가 있다.

6.3.4. 경제, 행정, 자격증 관련 전문 용어의 태깅 분류명

자금조달비용지수(코픽스), 코스피, 금융영어자격증(ICFE), 비즈니스영어자격증(BEF), 등 정보 추출에서 중요한 정보일 수 있으나 현재 15개 대분류, 146개 세분류에서는 해당 항목을 분류할 수 없으므로 관련 항목에 대한 검토가 필요하다.

6.4. 상호참조 해결

6.4.1. 구어 축약형

문어 말뭉치와 달리 구어에서는 다양한 축약형이 사용되곤 한다. 가령 ‘서울대학교의 연구 결괍니다만은’과 같은 자료가 존재하여 이때의 [서울대학교의 연구 결과]를 선행하는 [연구 결과]와 상호참조 그룹으로 묶어야 하는 경우가 있을 수 있다. 실제로 ‘이겁니다’의 경우에 [이겁]을 멘션으로 추출한 경우가 있는데, 이러한 경우에 대해 지침에서 명확히 설명해준다면 작업자들의 말뭉치 구축 과정에서 오류가 줄어들 수 있을 듯하다.

지침의 예)
1.4. 구어 자료의 처리

1.4.3. 축약형의 처리

구어 자료에는 구어 표현의 특성상 축약형이 다수 존재할 수 있다. 멘션 태깅은 형태 단위로 이뤄지므로 축약형의 경우 멘션의 원래 형태를 살려 태깅한다.

예1)

<예문> 서울대학교 연구 결괍니다만은(SBRW1800000012-0001)

<보기> [서울대학교](O), [서울대학교 연구 결과](O)

<오류> [서울대학교 연구결괍](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 전 만족해요. 정말로(SBRW1800000209-0001)

<보기> [저](O)

<오류> [전](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

6.4.2. 불용사전 목록 예문 보충

구축 지침에서는 멘션의 중심어가 불용사전에 포함된 경우 추출하지 않아야 함을 설명하며 불용사전의 목록을 제시하고 있는데, 이들에 예문이 보충될 필요가 있어 보인다. 가령 불용사전에는 '도시', '~사이', '~기', '신'과 같은 것들이 예문 없이 제시되어 있다. 이러한 형태를 중심어로 가진 모든 명사구가 추출 대상이 될 수 없는 것은 아니기에, 추출하지 않아야 할 대상들에 대한 예문이 보충되어야 작업자들의 오류를 줄일 수 있을 것이다.

예1)

<예문> 도시 모르겠어.(자체 생성 예시문)

<보기>

<오류> [도시](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

예2)

<예문> 그가 집에 간 사이, 난리가 났다.(자체 생성 예시문)

<보기> [그](O), [집](O), [난리](O)

<오류> [그가 집에 간 사이](X), [사이](X)

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

6.4.3. 지정사 관련 규정

현 지침에는 지정사가 포함된 경우도 상호참조의 대상이 된다고 명시하고 있으나, 'NP1이 NP2이다'에서 NP1과 NP2의 의미적 관계에 대해 자세히 기술하고 있지는 않다. 엄밀히 말하면 특정한 개체와 그 개체의 속성은 서로 상호참조를 할 수 없는 것이 맞지만, 본 사업에서는 NP1과 NP2의 관계가 '속성'일 경우에는 서로 상호참조하는 것으로 분석하였다. '속성'이 아닌 '서술'이나 '내용 기술'의 경우에도 상호참조 하는 것으로 태깅된 경우가 많아, 이와 같은 내용을 지침에 자세히 반영할 필요가 있다.

지침의 예)

2.2.3. 지정사가 포함된 경우

(추가) ▶ 'NP1은(는) NP2이다'와 같은 지정사 구문에서 NP1과 NP2의 관계가 '속성'일 경우에는 상호참조 태깅을 하며, NP1과 NP2의 관계가 '서술'이나 '내용 기술'의 경우에는 상호참조 태깅을 하지 않는다.

예1)

<예문> 박근혜 문재인 후보 모두 진정성과 원칙을 중시하는 정치인들이다.

(NWRW1800000036-0424)

<보기> [박근혜](O), [문재인](O), [박근혜 문재인 후보](O), [진정성과 원칙을 중시하는 정치인들](O)

<오류>

<상호참조 태깅> {[박근혜 문재인 후보], [진정성과 원칙을 중시하는 정치인들]}(O)

예2)

<예문> 가계부채는 한국 경제의 가장 큰 위험 요인이다.

(NWRW1800000044-0026)

<보기> [가계부채](O), [한국](O), [한국 경제](O), [한국 경제의 가장 큰 위험 요인](O)

<오류>

<상호참조 태깅> {∅}(예문 내 멘션들 중 상호참조 관계 없음)

6.5. 구문 분석

6.5.1. 부사격 조사가 생략되거나 보조사로 대체된 명사구의 AJT 분석 예시

AJT 분석 관련한 예문 추가가 필요하다. 특히 이중주어문, 이중목적어문 분석에서 보조사 결합 어절의 주어, 목적어 분석 오류의 경우, 문장의 주어, 목적어는 이미 있고, 문제가 되는 어절을 부사격 조사 결합형으로 대치할 수 있는 경우에는 AJT로 분석하도록 지침이 정해지는 것이 좋으리라 생각된다. 이렇게 하는 경우 대체로 수량사구가 AJT로 판정되는 예가 많을 것이다. 다음은 제안하는 지침 내용의 예이다.

부사격 조사 결합형으로 상정되지 않는 경우, 즉 주격·목적격 조사의 생략형일 수밖에 없는 경우에 한하여서만 SBJ, OBJ로 분석한다. 부사격 조사의 목록은 형태 층위 TTA 지침을 참고한다. (TTAK.KO-11.0010/R1, pp.24-25)

1	업종별	NP	3	
2	설비	NP	3	
3	가동률	NP	4	
4	상승폭은	NP_SBJ	8	‘올라’의 주어로만 생각되므로 기존의 분석 유지.
5	섬유가	NP_SBJ	8	
6	1월보다	NP_AJT	8	
7	4.5%포인트	NP_SBJ	8	‘4.5%포인트만큼’의 생략형으로 보아 AJT로 수정.
8	올라	VP	10	
5	“거래량과	NP_CNJ	6	
6	유동성이	NP_SBJ	7	
7	적은	VP_MOD	9	
8	모든	DP	9	
9	금융상품은	NP_SBJ	13	‘금융상품에서’처럼 부사어로 해석되므로 AJT로 수정.
10	변동성이	NP_SBJ	11	
11	클	VP_MOD	12	
12	가능성이	NP_SBJ	13	
13	높다”며	VP	31	

단, 서술어가 '걸리-'일 경우, '걸리-'의 구문틀에 따르면 수사구가 주어이므로 SBJ로 분석해야 한다.

걸리다
「동사」 【…에/에게】
「023」시간이 들다.
차가 밀려 다음 정거장까지 20분이 걸렸다.
이 일을 하는 데 세 시간이 걸린다.
초보자에게는 세 시간이 걸리는 일이 숙련공에겐 한 시간 정도가 걸린다.

호남고속철이 1시간33분 걸리다.
호남고속철이 NP_SBJ → 걸리다.
1시간33분 NP_SBJ → 걸리다.
'호남고속철이'는 표면에 주격 조사 '이/가'가 나타난 것을 존중하여 SBJ로 분석함. 만약 '호남고속철은'이었다면 '호남고속철로'라는 부사어로 고쳐 이해할 수 있으므로 AJT로 분석해야 할 것임. '1시간33분'은 주격 조사가 나타나지 않지만 '걸리-'의 주어는 시간을 나타내는 수사구이므로 SBJ로 분석함. '1시간33분이나'처럼 보조사 결합형으로 나타나더라도 SBJ로 분석함.

6.5.2. 후행절 서술어가 (의사) 보조 용언 구성일 때 선행절 서술어의 지배소

선행절 서술어가 후행절의 본용언, 보조용언 중 어느 쪽에 의존해야 하는지 지침이 불분명하다고 생각된다. ① 내포문의 서술어를 모문의 본용언, 보조용언 어느 쪽에 연결해도 무방한 경우가 있는 반면(예. 멜라닌은 자외선을 차단해서 자외선으로부터 피부를 보호해 준다.: 차단해서 → 보호해/준다), ② 보조용언에 연결해야 자연스러운 경우도 있다(예1. 식음료 업체들이 설 연휴가 끝나자 제품 가격을 일제히 올리기 시작했고, 일부 공공요금은 이미 올랐거나 인상을 앞두고 있다.: 올랐거나 → 있다. 예2. 거래 규모가 워낙 방대해 현실적으로 관리감독이 쉽지 않다.: 방대해 → 않다.).

'선행절 서술어를 후행절 본용언에 의존한다'와 같은 기계적인 지침보다, 의미를 고려하여 지침을 마련하고 적절한 예가 제시되어야 한다. 예를 들어 위의 ①의 경우에는 가장 가까운 서술어인 본용언 '보호해'에 의존하도록 하고, ②의 경우에는 보조용언에 의존하도록 한다고 규정할 수 있다. ②의 예1은 '~ 올랐거나'와 '~앞두고 있다.'가 나타내는 시제가 각각 과거, 현재로 다르기 때문에 보조용언에 의존하는 것이 자연스럽다. ②의 예2는 "관리감독이 쉽지 않은 이유는 거래 규모가 방대하기 때문"이라는 뜻인데, 만약 '방대해'를 본용언 '쉽지'에 연결하면 "거래 규모가 방대해서 관리감독이 쉬움"에 대한 부정이 되므로 이와 같은 구문 분석에 따른 해석이 실제 해석과 차이가 나므로 실제 해석이 유도되는 구조적 분석이 이루어지도록 '방대해'를 '않다.'에 연결해야 한다.

6.5.3 술어 생략 세부 유형

아래 지침 내용에서 '김동률로,' '장기하로,' '소식을,' 등, 생략된 서술어를 대신하여 문장 성분을 지배하는 명사구가 어디에 의존하는지 예시되어 있지 않다.

> 가. 소설가 **김영하에서** 그를 찜한 가수 김동률로, 가수 **이적에서** 인디뮤지션 장기하로, 드림위즈 최고 경영자(CEO) **이찬진에서** 신세계그룹 정용진 부회장으로, 네오위즈 CEO 허진호에서 마이크로소프트의 창업자 빌게이츠로, 마우스를 한 번 클릭해 찜하는 것만으로 이들이 트위터에 올리는 메시지가 실시간으로들어온다.
>
> 김영하에서 → NP_AJT 김동률로,
>
> 이적에서 → NP_AJT 장기하로,
>
> 나. '미국 드라마(미드)' **마니아라면** 한국 방영을 앞둔 최신 드라마 소식을, 정보기술(IT) 분야 종사자라면 해외 유저의 아이패드 사용기를 소개하는 식으로 SNS 세계의 '인맥 허브'에 등극할 수 있다.
>
> 마니아라면 → VNP 소식을,

위에서 (가) '김동률로,' '장기하로,' '부회장으로,' '빌게이츠로,'가 '실시간으로들어온다.'에 의존하고, (나) '소식을,'이 '소개하는'에 의존한다는 내용을 추가할 것을 제안한다.

6.6. 의미역

6.6.1. '없다'의 ARG-NEG 주석 대상에서의 배제

의미역 작업 지침이 몇 차례 개정을 거치면서, 어떤 규정이 더 이상 적용될 수 없는 단어가 나타나는 경우가 있다. '없다' 용언의 경우 현재의 구축 지침에는 ARGM-NEG를 '없다'에 태그하는 경우가 사라졌다고 생각된다. '-ㄹ 수 없다'는 의사 보조 용언이므로 주석 대상이 아니고, '수밖에 없다'는 '강조'로 부정에서 제외된다. 본용언 '없다'는 본용언이므로 부가역으로 태깅하지 않는다. 따라서 ARGM-NEG의 주석 대상에서 '없다'는 제외된다는 것을 명시하는 것이 좋을 것이다. 아울러 부사어 '안', '못'을 NEG에 추가할 것을 제안한다.

6.6.2. 부가의미역의 추가

현행 지침에는 13종의 부가의미역이 존재하나 부가어 모두를 주석하기에는 부족하다.

예)

14일 월스트리트저널 인터넷판에 따르면(ARGM-CND) 이미 250억 달러의 자금을
그는 "질리지 않고 재미있게(ARGM-MNR) 계속할 수 있고

위 예에서 '따르면'과 '재미있게'가 각각 'CND', 'MNR' 주석되어 있는데 'CND', 'MNR'의 정의에 부합되지 않는다. 현행 지침에는 이를 대체할 만한 부가의미역이 존재하지 않는다. 문장에 실현되는 부가어의 다양성을 포착하기 위해 부가의미역이 확대되는 것이 좋다고 생각된다. 위의 예의 경우 '에 따르면'은 'SRC(source)', '재미있게'는 'EMT(emotion)' 정도의 부가역을 신설하여 주석하면 될 것이다.

6.6.3. 'A/V-게'의 PRD 주석 가능 여부

'깔때기 모양으로 파인'에서 '깔때기 모양으로'의 자리에 '둥글게' 등의 부사어가 올 수 있다. '둥글게' 등을 PRD로 주석 가능한지가 명확하지 않다. 구축 말뭉치에서 'A/V-게'에 PRD를 주석한 경우는 발견되지 않았다. '둥글게'와 같이 결과 상태를 나타내는 경우 PRD로 주석하도록 할 것을 제안한다.

6.6.4. 명사가 격틀에 영향을 미치는 경우

서술어 앞에 결합하는 명사에 따라 새로운 논항이 추가되는 경우가 있다.

예)
대통령은 **야당 대표와** 회동을 가졌다.
남자는 **그 노트를** 필요로 했다.
나는 그 **남자와** 친분이 있다.

위 예에서 '야당 대표, 그 노트, 그 남자'는 '가졌다, 했다, 있다'의 논항으로는 볼 수 없다. '가졌다, 했다, 있다'가 앞의 논항과 결합하여 생긴 논항에 해당한다. 현재 구축 지침상에는 이와 관련된 지침 내용이 없다. VP[NP-V]가 새로운 논항을 취하는 사례를 조사하고 이에 대한 지침 내용이 추가되어야 한다.

6.6.5. predicate lemma의 결정 규칙

현재의 predicate lemma 결정 규칙엔 혼동의 여지가 있다고 판단된다. 다음과 같이 지침 텍스트를 변경할 것을 제안한다.

- 술어 정의(ID)는 '능동형의 어근' 또는 '기본형'에 아래 숫자를 더하여 주석한다.
- Korean PropBank와 Etri는 논항 정보(FrameSet)에 제시된 lemma 형식을 그대로 따른다.

술어 정보 출처	논항 구조 ID	주석 ID		
		lemma		sense_id
K-propbank	어근.**	동사 '**하다'	어근	44444**
		형용사, 그 밖의 동사,	어간	
Etri	어근.**	'**하다' 동사	어근	55555**
		형용사, 그 밖의 동사	어간	
U-propbank	기본형 ******	기본형		6******
우리말샘	기본형 ***	기본형		***

6.7. 주격 무형 대용어 복원

6.7.1. 서술어 형태(Predicate form)

주격 무형 대용어 복원 말뭉치 구축 지침에 있는 '산출물 후처리 기준'이란 'Predicate form'의 기호 및 조사 삭제 기준에 관한 것이다. 그런데 세 가지 기준 중에서 문법 형식에 관한 기준은 '조사 이외 어미 유지'라고만 제시되어 있고 이에 따른 상세 설명이 없다.

지침내용:

2-2) 조사 이외 어미 유지: 물타기"라는, 물타기라는, 먹자"며, 먹자며 -> 물타기"라는, 물타기라는, 먹자"며, 먹자며(해당 술어의 주어는 생략된 술어 '말하다'의 주어로 처리 혹은 복원. 따라서 어미 혹은 축약형 어미를 삭제하면 분석이 달라지므로 삭제 X)

수정제안:

2-2) 조사와 어미 처리: 조사는 생략 가능하며, 이때 생략 가능한 조사 목록은 다음과 같다(생략 가능한 조사 목록 제시). 한편 술어에서 어미 혹은 축약형 어미를 삭제할 경우 문장 분석이 달라질 수 있다. 예를 들어 '라는'과 '"며'는 각각 '라고 하는',

'라고 하며'의 축약형이며, 이때 생략된 주어 '말하다'의 주어로 처리 또는 복원한다.

위 문법 형식과 더불어 구어 형식 오류 목록 중에서 predicate form에 '게(것이)'가 빠져있는데, 이때 word form과 predicate form이 일치한 형태가 맞는지(어려운게 - 어려운게), 빠진 형태(어려운게 - 어려운)가 맞는지에 대한 지침이 없어 이에 대한 설명이 필요하다.

6.7.2. 서술어에 해당하지 않는 VP(_MOD) 목록

구축 지침에는 서술어에 해당하지 않는 VP(_MOD)에 관한 설명이 다음과 같이 제시되어 있다.

지침 내용:

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 {잇는다|이었다}.]

(생략어 지침 7쪽) VP로 분석된 일반 용언은 담화나 문서 내에서 개체와 개체 또는 개체와 값 간의 사건, 행동및 상태와 같은 주요한 정보를 서술하는 데 쓰인다. 이러한 서술어에 해당하는 VP를 일차적인 지배소 후보로 생각할 수 있다. 서술어에 해당하지 않는 VP는 생략어의 지배소로 여기지 않는다. (예: 관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고, 그러면서 등) 그리고, 하나의 다어절 용언은 서술의 의미가 들어있는 용언에만 태깅한다

그런데 위 두 지침만으로는 서술어에 해당하지 않는 VP(_MOD)의 전부를 알 수 없고, 이에 따른 부가 설명이 반드시 필요하다.

수정제안:

(1) 서술어가 아닌 VP(_MOD) 목록 확대하고 위에 명시된 예들뿐만 아니라 관련된 형태를 모두 제시(예: 관해, 대해, 의해, 향해, 인해, 통해, 따라, 아니라, 불구하고, 그러면서, 이어, 관한, 대한, 의한, 향한, 인한, 통한, 따른, 아닌, 같은, ~에 따르면, ~를

넘어, ~시점을 맞아, ~에 비해, 위해, 위한, 앞서 등)

(2) 위 생략어 지침 내용에서 “서술어에 해당하지 않는 VP는 생략어의 지배소로 여기지 않는다. 그러나 문면의 주어가 나타날 경우 이들 VP(_MOD)에 후행하는 서술어의 주어를 복원한다.”는 문구와 관련 예문 추가.

예1) 철수는 과학에 대해 공부했다. --> 문면의 ‘철수는’은 ‘대해’의 주어임. 따라서 ‘공부했다’의 주어를 복원해야 함.(‘철수’로)

예2) 과학에 대하여 철수는 공부했다. --> ‘대하여’의 주어가 생략됨. 그러나 복원하지 않음. 문면의 ‘철수는’은 ‘공부했다’에 의존.

예3) 철수는 책을 읽고 과학에 대하여 생각했다. --> 문면의 ‘철수는’은 ‘읽고’에 의존. ‘대하여’의 주어가 생략됨. 그러나 복원하지 않음. ‘생각했다’의 주어는 복원함.(‘철수’로)

예4) 철수는 책을 읽은 후 과학에 대하여 생각했다. --> 문면의 ‘철수는’은 ‘대하여’에 의존. ‘읽은’, ‘생각했다’의 주어를 복원해야 함.(‘철수’로)

(3) 주어 복원 대상/대상 아닌 유사 구문 비교 제시

예:

-어야/아야 하다의 ‘하다’ 복원 대상 아님.

-어야 되다의 ‘되다’ 복원 대상.

-게 하다의 ‘하다’ 복원 대상 아님.

-게 되다의 ‘되다’ 복원 대상.

‘~기도 하다’의 ‘하다’ 복원 대상 아님.

‘~기로 하다’의 ‘하다’, 복원 대상.

‘~면 (안)되다’의 ‘되다’ 복원 대상.

6.7.3. ‘그러하다/그렇다’

문어에서는 ‘그럼에도 불구하고’와 같이 접속부사는 아니지만 ‘그러하다/그렇다’가 쓰여 접속부사적 기능을 하는 관용적 표현들이 많이 나타나고, 구어에서는 ‘그렇습니다/그렇죠’ 등이 많이 나타나는데, 이러한 경우 주어 복원 여부가 매우 까다롭기 때문에 보다 명확한 지침이 필요하다.

위 내용은 현재 구어 지침에는 없는 내용이기 때문에 구어에서 상대방 발화에 맞장구치는 '그러하다/그렇다', '맞다' 등의 분석에 대하여 세부 지침을 정해야 한다(또는 이러한 표현은 주어가 불분명하므로 복원하지 않는 것으로 재고할 수도 있다).

주어 복원이 필요하다고 판단될 경우 다음과 같은 지침을 추가할 것을 제안한다.

구어에서 '그렇습니다', '그렇죠(그쵸)', '그렇군요' 등, 그리고 '맞아요' 등의 맞장구 표현이 나타날 경우 선행어의 기준을 다음과 같이 설정한다. 1) 선행 문장에 주어가 있을 경우 해당 주어로 복원. SBRW1800000096.1.1.213 2 za_ante 인터뷰__@s208_2 뭐 그렇지만은 2) 선행 문장에 주어가 없을 경우 '무언가'로 복원. SBRW1800000101.1.1.353 1 za_ante 무언가__@-1 그렇다고 복지가 올라간 것도 아니잖아요.

제 4 장

전문가 토론회 개최

1. 전문가 토론회 개요

전문가 토론회는 국가 주도의 대규모 말뭉치 구축 이후에 현장에서 어떻게 말뭉치 데이터를 활용할 수 있을 것인지를 모색하는 행사였다. 행사의 개략적인 정보는 다음과 같다.

□ 개요

○ (행사명) 2020년 국립국어원 인공지능 시대를 향한 우리말 빅데이터의 활용

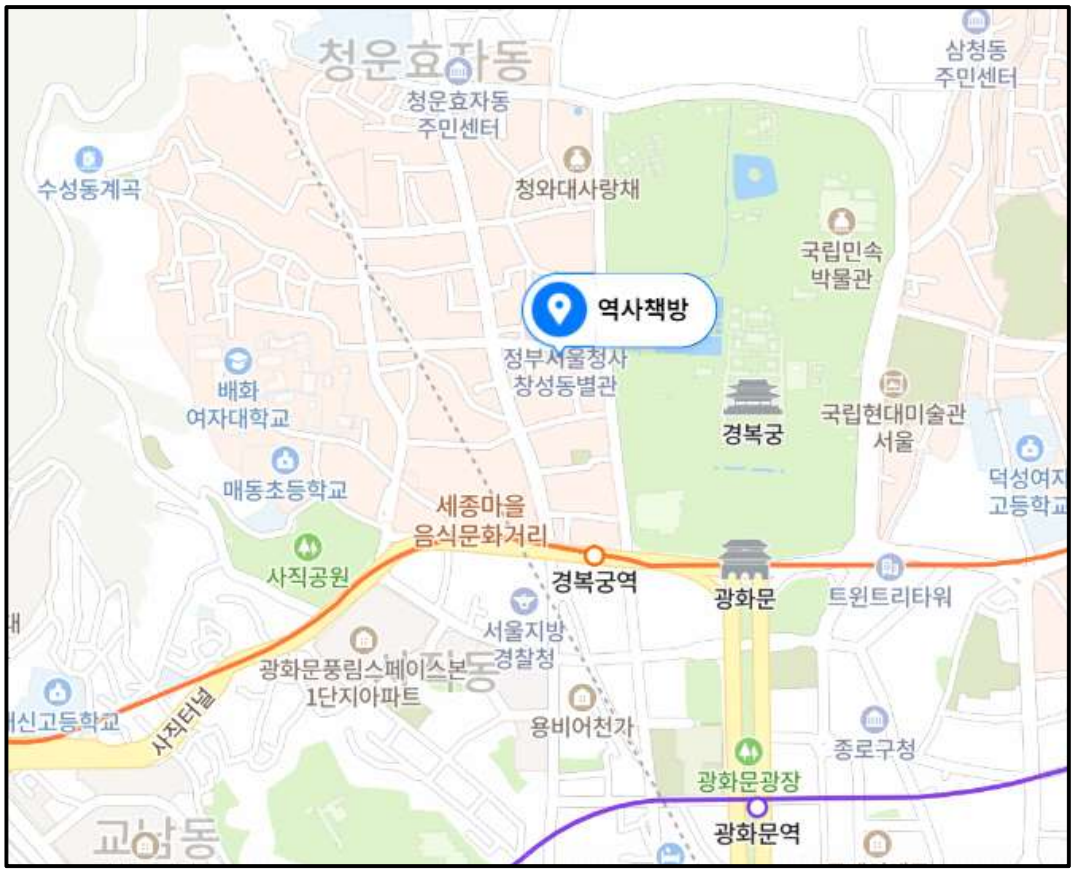
○ (행사일정) ‘20. 10. 14. (수) 14:00 ~ 16:30

○ (참여인원) 사회자(고려대 송상헌 교수), 발표자 3인(네이버 강인호 책임리더, NC 소프트 이연수 실장, 솔트룩스 이경일 대표) 외 온라인 참여자(150명)

○ (행사장소) 역사책방, 서울시 종로구 자하문로 10길 24

- (주행사장) 역사책방 1층

<그림 34> 역사책방 위치



○ (행사내용) 인공지능 시대를 향한 우리말 빅데이터의 활용

2. 행사 일정

일시		행사			비고
		구분	주요내용	연설자	
2020.10.14.(수)					
11:00~13:55	175‘	행사장 조성 및 리허설	행사장 조성 및 리허설		
14:00~14:10	10‘	사회자 인사 및 개회사	행사 개회사	송상헌 교수 소강춘 원장	라이브 송출 사전녹화본 송출
14:10~14:40	30‘	강연 1	우리말 빅데이터를 이용한 AI 서비스 트렌드	강인호 리더	사전녹화본 송출
14:40~15:15	30‘	강연 2	우리말 빅데이터 활용 사례 및 서비스 응용	이연수 실장	사전녹화본 송출
15:15~15:20	05‘	휴식	휴식		
15:20~15:50	30‘	강연 3	우리말과 인공지능	이경일 대표	사전녹화본 송출
15:50~16:30	30‘	자유토론	자유토론 및 Q&A	송상헌 교수 강인호 리더 이연수 실장 이경일 대표	라이브 송출

3. 행사장 구성

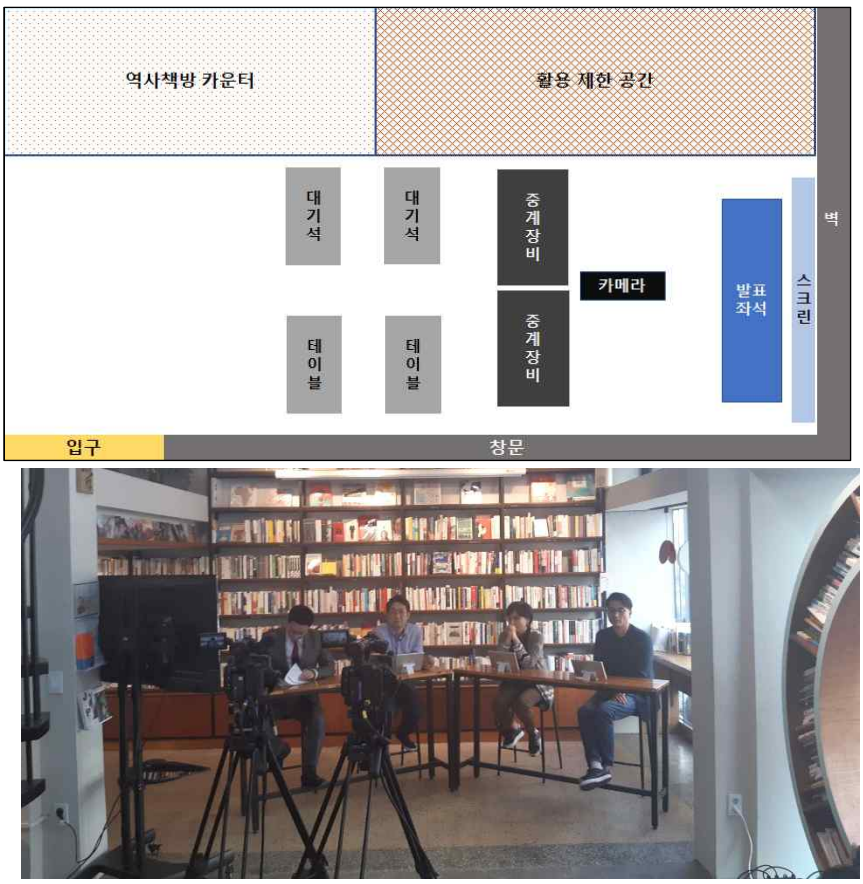
□ 행사장 개요

ㅇ (장소) 역사책방

ㅇ (행사대상) 국립국어원 빅데이터 관련 전문가 및 언론

ㅇ (행사장 인원 구성) 국립국어원 관계자, 연설자, 행사진행인력

구분	구성
국립국어원	관계자 및 담당자 3명
사회자	고려대학교 송상헌 교수
강연자	네이버 강인호 자연어처리 책임리더
	엔씨소프트 이연수 실장
	솔트룩스 이경일 대표
행사진행인력	중계엔지니어 2인 / 헤어,메이크업 담당 1인
	행사담당자 5인(나라지식정보, 애스톤)



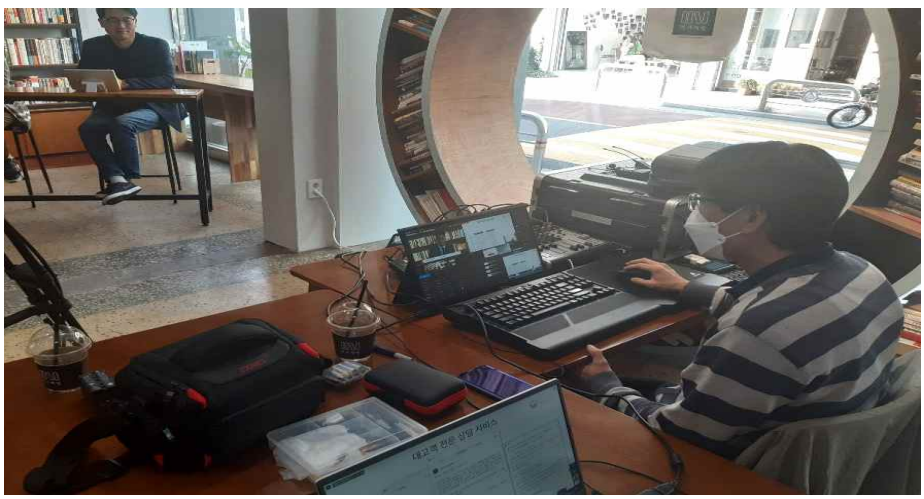
<그림 35> 행사장 구성도

4. 행사 운영

4.1. 온라인 웨비나 플랫폼

여러 사람이 밀집하는 환경 조성을 피하기 위하여, 온라인 화상회의/웨비나 플랫폼인 ZOOM을 활용하여 행사를 운영하였다. 행사는 순수 웨비나 방식과 사전녹화본 송출을 결합한 방식으로, 다음과 같이 혼합적으로 구성되었다.

- 행사 소개 및 사회자 멘트 (현장 라이브 송출)
- 강연 송출 (강연자 3인 동일 행사장 사전 녹화분 송출)
- 자유토론 및 질의응답 (플랫폼을 통한 실시간 질의 및 답변 라이브 송출)



<그림 36> 전문가 토론회 행사 운영 플랫폼

4.2. 사전 홍보 및 신청

국립국어원 내부 전문가 및 언론 홍보대상으로 사전 참가 신청 링크(구글폼 주소)를 포함한 포스터를 배포하였다.

국립국어원 관련 사업 진행의 경험이 있거나, 현재 진행 중인 15개 사업단에 참가 요청을 하였고, AI/NLP 관련 그룹, 개인 등 50여 곳에 행사를 안내하였다.

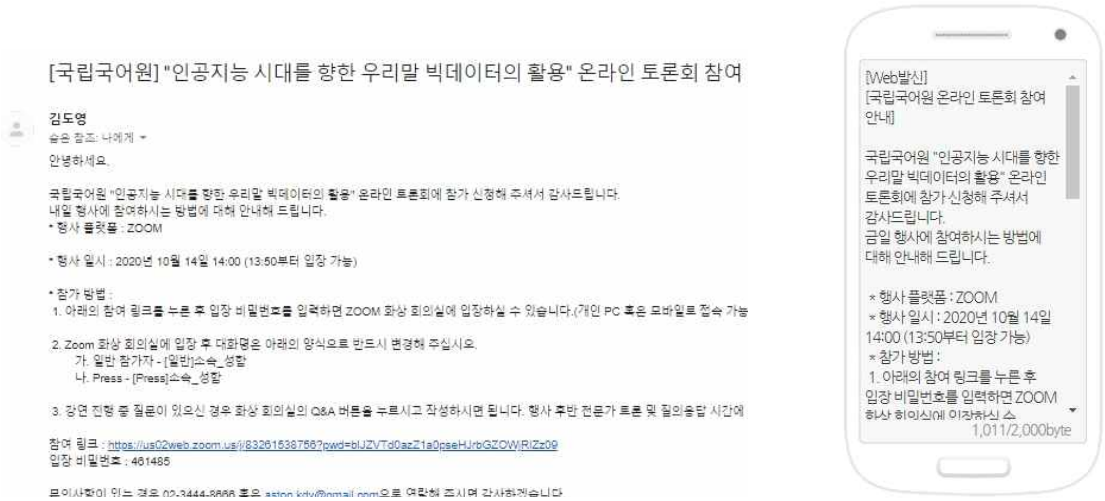
21개 언론 매체에 별도로 접촉한 결과 10개 매체가 보도하였다.

<표 13> 전문가 토론회 언론 매체 보도

신문사	기사 링크	기사 제목
연합뉴스	https://www.yna.co.kr/view/AKR20201008148700005?input=1195m	[문화소식] 국립국어원, 인공지능과 우리말 자료 활용 토론회
뉴시스	https://newsis.com/view/?id=NISX20201008_0001191723&cID=10701&pID=10700	인공지능 시대, 우리말 빅데이터 어떻게 활용할까
ZD-Net	https://zdnet.co.kr/view/?no=20201010135131	AI 시대, 우리말 빅데이터 구축은?... 온라인 토론회 14일 열려
한국강사신문	http://www.lecturernews.com/news/articleView.html?idxno=53451	국립국어원, '인공지능 시대를 향한 우리말 빅데이터의 활용' 온라인 토론회 개최
매일경제	https://www.mk.co.kr/news/culture/view/2020/10/1039670/	우리말 빅데이터 활용 토론회 열린다
IT조선	http://it.chosun.com/site/data/html_dir/2020/10/	국립국어원, 자연어처리 전문가 모여

	12/2020101201012.html	'빅데이터 토론회' 연다
우리문화신문	https://www.koya-culture.com/news/article.html?no=126598	우리말 빅데이터의 활용, 언어 기술 개발 이끈다
경향	http://news.khan.co.kr/kh_news/khan_art_view.html?artid=202010121224001&code=960100	국립국어원, '인공지능 시대를 향한 우리말 빅데이터의 활용' 토론회 14일 개최
한겨레	http://www.hani.co.kr/arti/culture/religion/965474.html	'인공지능시대 우리말 빅데이터' 14일 토론회
보안뉴스	https://www.boannews.com/media/view.asp?idx=91713&kind=	국립국어원, 인공지능과 우리말 자료 활용 온라인 토론회 개최

구글 폼을 통해 취합된 참가 희망자를 대상으로 행사 접속 URL 및 행사 안내문을 메일과 문자로 발송하였다.



<그림 37> 발송 메일 및 문자

4.3. 참가자 구분 및 자유토론, Q&A 운영

1) 참가자 구분

모든 참가자에게는 사전 안내 및 Zoom내 공지를 통해 회의실 입장 시 참여자명을 작성하도록 하였다.

예) 기자 참여자명 : [Press] 소속_이름

일반 참여자명 : [일반] 소속_이름

2) 자유토론 운영

ZOOM을 통해 현장 강연자 3인 외 사회자 1인으로 진행되는 토론을 송출하였다. 토론 주제는 다음과 같았다.

① 원시 말뭉치와 분석 말뭉치의 바람직한 발전 방향

② 국가경쟁력 증진의 측면에서 특히 가치 있는 언어자원은 어떠한 것들이 있는가?

③ 국가 차원(국립국어원 등)에서 실물적 측면으로 어떠한 역할과 도움을 주어야 하는가?

3) Q&A 세션 운영

ZOOM 플랫폼에서 지원하는 Q&A 기능을 통해 질문을 접수하고, 강연자가 전면에 설치된 모니터링 스크린을 통해 질문을 확인한 후 실시간 영상중계를 통해 답변하였다. 질의응답은 공통질문, 기자 질문, 일반 참가자 질문 순서로 진행하였다.

<그림 38> 토론 및 Q&A

<질문 리스트>

공통질문 : GPT-3 관련 각 회사의 구체적 대응 계획이 있는가? 있다면 그 계획과 국립국어원의 말뭉치 구축과는 어떤 관계가 있을 것인가?

기자 질문 :

1) AI학습과 자연어 처리기술에 있어 자본력이 충실한 회사가 유리한 것이 현실이고 이러한 추세는 점점 커질 것으로 보이는데 어떻게 생각하시는지 또한 대처 방안이 있으신지?

3. 일반 참가자 질문 :

1) 요약문 생성에 있어 원문과 사람이 쓴 요약문을 쌍으로 학습해야 하는가? 아니면 언어학습이 충분히 이루어졌을 때 기계가 요약문을 자동생성할 수 있는가?
2) 한국 내에서 외국어 자연어처리(중국어 등)에 대한 수요가 있는가? 만약 있다면 어떤 수요들이 있는가?
3) 현제 NAVER가 데이터 수집에 가장 유리해 보이는데 말뭉치나 데이터셋을 공개할 계획이 있는가?
4) AI연구도 그렇지만 언어학 내에서도 좀더 정확한 구문 분석을 하기 위한 또는 어떠한 문제를 해결하기 위한 연구가 계속되고 있다. 언어학적인 부분과 공학적인 부분이 앞으로 어떻게 함께 나아갈 수 있는가?
5) 데이터3법이 이루어지면 내 정보들이 AI에 활용될 가능성이 있는가?
6) 기계가 자율의지를 가지는 것이 가능할까? 만약 그렇다면 향후 미래 전망은?
7) 각 기업의 기계번역 연구의 방향성 및 계획은?

제 5 장

결론과 제언

1. 결론

국립국어원에서는 4차 산업혁명에 대비한 국어 빅데이터(말뭉치) 구축 사업을 지속적으로 수행중이다. 그러한 노력이 성과가 있어, 대규모 말뭉치 구축의 결과물이 순차적으로 공개되고 있으며 이용자들의 호평도 들려오고 있다. 그런데 진정으로 학계와 산업계에서 중요 자료로 활용될 수 있는 말뭉치는 한 번 구축해 두는 것으로 완성된다 할 수 없다. 구축되어 있는 데이터에 대해 지속적으로 보완 의견을 수렴하고 후속 서비스를 유지하였을 때 그것이 비로소 현실에 활용될 수 있는 데이터로서 생명력을 갖는 것이다. 본 사업에서는 정교하게 구축된 대규모 말뭉치 자료의 후속 관리라는 대의에 기여하기 위하여, 말뭉치의 통합과 관리라는 측면에 주목하며 과업들을 수행하였다.

음성 말뭉치의 통합 및 정비 과업은 고품질의 구어 말뭉치 구축을 위하여, 녹음과 전사 텍스트의 정교한 대응을 성립시키는 통합 작업이었다. 다만 이 과업에 속한 두 가지 세부 과업의 방향은 달랐다. 2018년 일상 대화 음성 말뭉치 정비 작업은 음성 데이터를 가공하는 작업이 대부분이었고, 서울말 낭독체 말뭉치 정비 작업은 전사 텍스트를 가공하는 작업이 대부분이었기 때문이다. 이렇게 과업 성격이 달랐기 때문에 각각에 어울리는 과업 수행 방식을 택하여 진행하였다. 아울러 2019년도 일상 대화 말뭉치 구축 사업에 참여한 전문 인력이 본 사업에도 참여하였기 때문에, 비교적 적은 시행착오를 겪으며 과업 수행을 할 수 있었던 것으로 평가한다.

7개 층위 분석 말뭉치의 통합 및 정비 과업은 말뭉치의 규모에 비해 짧은 시간에 집약적으로 이루어졌다고 할 수 있다. 이 때문에 가능한 효율적인 과업 구성을 할 필요가 있었다. 말뭉치 층위 통합 단계에서는 말뭉치의 내적 비일관성과 말뭉치들 사이의 불일치를 여러 단계에 거쳐 검증하였고, 국립국어원과 긴밀히 연락하며 원칙을 세워 일관되게 바로잡았다. 전문가 주석 검증 단계는 사업단 내 전문가들에게 일률적인 검증 환경을 사전 결정하여 제공할 경우 검증 환경 설계 과정에서 고려하지 못한 요인 때문에 비효율이 초래될 것이 우려되었다. 따라서 검증에 참여한 전문가들에게 지속적으로 피드백을 받고, 작업 과정에서 생겨난 요구와 제안에 유연하게 대응하여 필요한 데이터와 작업 환경을 신속하게 제공하는 체제를 갖추어 운영하였다. 이 과업에도 전년도의 말뭉치 통합 검증 사업에 참여한 전문 인력이 참여하여 지침에 대한 이해와 오류 유형 파악이 높은 수준에서 이루어졌다고 평가할 수 있다.

전문가 토론회는 말뭉치 관련 산업에 종사하는 각계 전문가들을 섭외하여, 약 150명의 청중이 지켜보는 가운데 개최하였다. 순수 웨비나 방식과 사전녹화본 송출을 결합한 방식을 시도하였는데, 전반적으로 원활하게 행사가 진행되고 청중의 호응도가 높았다.

코로나 시대의 한 모범이 되는 토론 행사가 되었다고 생각된다.

종합하면, 국가 주도로 구축한 다층위 말뭉치들의 응용 가능성을 극대화하기 위해 말뭉치의 통합과 정비를 수행하고, 전문가 의견 수렴을 진행하여 앞으로 이루어질 말뭉치의 구축과 활용의 비전을 제시한 데에 본 사업의 의의를 둘 수 있다.

2. 제언

본 사업을 진행하는 동안 사업단이 마주한 어려움과 시사점을 간략히 기술한다. 이것이 앞으로의 말뭉치 구축과 관리에 참고가 되길 바란다.

첫째, 과업의 중요 참고 데이터와 관련한 문제이다. 2018년도 일상 대화 말뭉치 통합 및 정비 과업은 녹음된 음성 파일과 구축된 전사 텍스트를 기본 자료로 하여 전사 텍스트와 잘 대응되는 정제된 음성 파일을 산출하는 것이 주 작업이었다. 그러나 전사 지침에 어긋나게 전사되거나 음성과 다르게 전사된 텍스트가 적지 않아, 전사 텍스트를 참조하며 진행하던 음성 정제 작업에서 미처 예견하지 못했던 어려움을 겪었다. 이를 2019년도 일상 대화 말뭉치 구축(국립국어원, 2019)과 견주어 보면, 2019년도 사업의 경우 전사 과업을 수행한 사업단이 음성 정제 과업도 수행하였기 때문에 과업의 연속성이 유지되었던 반면, 본 사업의 경우 전사 과업을 수행한 사업자와 음성 정제 과업을 수행한 사업자가 달라 다소간의 혼란이 유발된 것으로 생각된다. 향후 이와 유사한 성격의 과업이 수행된다면, 과업 수행자의 연속성이 유지된다면 좋을 것이고, 과업 수행자의 변경을 피할 수 없는 경우 기구축 자료 검토를 일정 계획에 포함하는 것이 좋을 것이다.

둘째, 과업 일정과 관련한 문제이다. 이 사업은 이미 구축된 말뭉치의 통합과 정비를 지원하는 사업이기 때문에, 구축 사업단의 유지 보수 업무 일정 등 외부 요인의 영향으로 세부사항이 변동될 가능성이 있었다. 그러나 본 사업의 진행 일정은 세부 과업 조율에 투입되는 노력과 시간을 고려하지 않은 채 계획되었다는 아쉬움이 있었다. 사업의 높은 변동성이라는 특성이 사업 착수 전에 공유되고, 세부 과업 조율까지 고려한 사업 추진 일정이 수립되었다면 더 좋았을 것이다. 나아가, 의미있는 세부 과업 조율을 위해서는 사업단이 자료의 구조에 대해 잘 이해하고 있어야 하고 말뭉치 관리의 지향점이 모든 참여자 사이에서 깊은 수준까지 공유되고 있어야 할 것이다.

셋째, 구축 지침의 신뢰성과 효력에 대한 문제이다. 사업단은 사업의 성격상 과업 수행 과정 중에 구축 지침이 변경되는 일은 일어나지 않는 것으로 상정하였다. 이미 주어진 지침에 의해 구축된 말뭉치를 검증하는 것이 핵심 과업이었기 때문이다. 그런데 사

업 과정에서 지침에 명시된 일부 의미역 격틀 프레임셋 자료를 말뭉치에서 배제하는 등 지침 변경에 해당하는 업무 지시가 있었으며, 이 때문에 과업 수행 과정에서 다소간의 혼란이 있었다. 구축 지침의 변경을 최소화하거나, 지침 변동의 가능성이 사전에 고지된다면 이러한 혼란을 예방할 수 있으리라 생각된다.

이와 같은 점들을 고려한다면 향후의 말뭉치 구축 및 후속 관리가 더욱 성공적으로 진행될 수 있을 것이다.

참고문헌

국립국어원(2019ㄱ), 『개체명 분석 말뭉치 구축』, 국립국어원 연구보고서.
국립국어원(2019ㄴ), 『구어 자료 수집 및 원시 말뭉치 구축』, 국립국어원 연구보고서.
국립국어원(2019ㄷ), 『상호 참조 해결 말뭉치 구축』, 국립국어원 연구보고서.
국립국어원(2019ㄹ), 『일상 대화 말뭉치 구축』, 국립국어원 연구보고서.
국립국어원(2019ㅁ), 『형태 분석 말뭉치 구축』, 국립국어원 연구보고서.
국립국어원(2020ㄱ), 『말뭉치 통합 검증』, 국립국어원 연구보고서.
국립국어원(2020ㄴ), 『어휘의미 분석 말뭉치 구축』, 국립국어원 연구보고서.

<Abstract>

2020 NIKL Corpus Integration and Maintenance Support

This project aims to seek integrated management and operation plans for state-led multi-level corpus, support continuous quality assurance in response to increasing demand for high-quality corpus, establish an accurate and quick response system to meet the needs of corpus users, and contribute to managing the needs of professional consumers related to corpus such as academia, industry. Three tasks were carried out to achieve the goal. i) Integration and maintenance of the spoken corpus, ii) Integration and maintenance of seven-level tagged corpus, and iii) Holding experts forum.

In the spoken corpus' integration and maintenance, the segmentation of two hundred twenty-one voice files by a speech unit for the sound to match the transcription unit and personal information de-identification, i.e., speech refinement. Besides, eighty-eight thousand recording files that various Seoul speakers by age and gender group read and recorded given scripts consisting of approximately eight thousand seven hundred words by space were transcribed.

In the maintenance and integrated management of the seven-level tagged corpus, JSON format verification, level integration and annotation error verification, guideline supplementation, example expansion, and comparison with the maintenance project group corpus were performed. The details are as follows.

The JSON format consistency is examined for the seven-level tagged corpus constructed in 2019 (morpheme analysis, lexical meaning, named entity,

cross-reference resolution, dependency construction analysis, semantic role, zero subject). After that, the level formats are integrated, and annotation errors in the designated number of documents were examined and corrected. Furthermore, investigating the experts' opinions within the project group, we renovated corpus construction guidelines and suggested the sample data for corpus expansion. At the same time, co-verification with the maintenance project group was carried out. The level-integrated corpus was sent to the maintenance project group. The revision was made respectively, and both project groups' results were compared at the end of the project.

For hosting an experts forum, a panel discussion entitled "Utilization of Korean Big Data for the Age of Artificial Intelligence" was held. One moderator and three experts participated in a parallel online/offline forum, gave lectures, had a discussion with each other, and answered the audience's questions. An audience of about one hundred fifty people participated.

The significance of this project is to carry out the integration and refinement of the corpus to maximize the applicability of multi-level corpus constructed by the government and, based on the specialists' opinion, to present a vision for the future construction and utilization of corpus.

Keywords: integrated corpus management support, spoken corpus management, sound refinement, level integration, tagged corpus verification

Project Director: Lee Euijong (Nara Information, Co., LTD)

사업 책임자 이의종 (㈜나라지식정보)

사업 참여자 고동현 (㈜나라지식정보)

길혜빈 (경희대학교 국어국문학과 박사수료)

김선영 (서울대학교 언어교육원 대우전임강사)

김은수 (한양대학교 한국언어문학과 학사)

김지원 (㈜나라지식정보)

김태경 (한양대학교 ERICA 창의융합교육원 부교수)

김태우 (부산대학교 국어국문학과 조교수)

김한나 (경희대학교 국어국문학과 박사과정)

김희숙 (㈜나라지식정보)

박승희 (㈜나라지식정보 전무이사)

박영훈 (㈜나라지식정보 부장)

박융배 (이화여자대학교 뇌융합과학연구소 빅데이터 경영연구소 연구원)

박지용 (서울대학교 국어국문학과 강사)

박진호 (서울대학교 국어국문학과 교수)

박하선 (㈜나라지식정보)

박혜승 (서울대학교 국어국문학과 강사)

배준호 (㈜나라지식정보)

손지은 (고려대학교 국어국문학과 박사수료)

송상헌 (고려대학교 언어학과 조교수)

신용남 (서울대학교 국어국문학과 박사수료)

신희원 (한양대학교 한국언어문학과 학사)

안대섭 (고려대학교 노어노문학과 강사)

안의정 (연세대학교 학부대학 강사)

유현조 (서울대학교 인문데이터과학연계전공 초빙부교수)

유혜선 (㈜나라지식정보)

윤기현 (바이칼AI 대표)

윤예진 (서울대학교 국어국문학과 박사수료)

이강혁 (서울대학교 국어국문학과 박사수료)

이경원 (㈜나라지식정보)

이규환 (㈜나라지식정보)

이민우 (사이버한국외국어대학교 한국어학부 조교수)

이용규 (서울대학교 국어국문학과 박사과정)

이재혁 (㈜나라지식정보 수석연구원)

이주연 (㈜나라지식정보)

장원철 (㈜언어과학 상무이사)

장하연 (부산외국어대학교 영어학부 조교수)

정규상 (㈜나라지식정보 부장)

정우헌 (㈜나라지식정보)

정유남 (고려대학교 국어국문학과 강사)

정혜주 (㈜나라지식정보)

조혜미 (서울대학교 영어영문학과 학사과정)

최준호 (서울대학교 국어국문학과 박사수료)

최 진 (㈜나라지식정보)

황은하 (배재대학교 국어국문한국어교육학과 조교수)

담당 연구원 이승재(국립국어원 언어정보과장)

김소희(국립국어원 언어정보과 학예연구사)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2021년 2월 1일
발행일: 2021년 2월 1일
인 쇄: ㈜유성프린팅

※ 이 책은 국립국어원의 용역비로 수행한 '2020년도 국립국어원 말뭉치 통합 관리 지원' 사업의 결과물을 발간한 것입니다.