국립국어원 2023-01-41

발간등록번호
11-1371028-000969-01

공공언어 감수 자료 구축 방안 연구

연구 책임자

양 명 희

문화체육관광부
국립국어원

국립국어원 2023-01-41

발간등록번호
11-1371028-000969-01

공공언어 감수 자료 구축 방안 연구

연구 책임자

양 명 희

문화체육관광부
국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 “공공언어 감수 자료 구축 방안 연구”에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2023년 11월 ~ 2023년 12월

2023년 12월 20일

연구 책임자: 양 명 희(중앙대학교)

연구 기관 중앙대학교 산학협력단

연구 책임자 양명희

공동 연구원 송상헌, 정유남, 박미은, 김보현, 안예림, 홍승혜

연구 보조원 김혜원, 류다정, 이규민

보 조 원 이혜경, 정민경, 심재란, 박우빈, 조선화, 이혜영, 박예인, 조은비

[국문 초록]

공공언어 감수 자료 구축 방안 연구

본서는 2023년 국립국어원 ‘공공언어 감수 자료 구축 방안 연구’ 사업의 결과물로서, 2010년부터 2022년까지 국립국어원에서 구축된 공공언어 감수 전후 자료를 망라하는 한편, 새로운 공공언어 자료(2023년)를 추가로 수집하여(총 문서 수 5,031건) 정제하고 통합하여 병렬 말뭉치를 구축하고자 한 사업이다. 2010년부터 최근까지의 공공언어 감수 사업의 결과물을 아우름과 동시에 이 분야에 대한 전망과 정책 제안을 담고 있다.

이 사업의 첫째 과업인 공공언어 자료의 수집은 크게 두 가지 과업을 포함한다. 국립국어원의 공공언어 자료 감수와 관련된 기존 사업의 결과물을 통합하는 것과 여기에 더해 새로운 공공언어 자료를 추가로 수집, 감수하는 것이다. 기존 사업의 결과물로는 “공공언어 감수 지원 종합 자료 구축(2017, 김일환 연구책임)” 사업에서 구축한 2010~2016년 공공언어 자료를 대상으로 한 감수 전후 자료와 2017~2022년에 구축된 공공언어 자료를 대상으로 한 감수 전후 자료가 있다.

이 사업의 두 번째 과업은 수집된 자료를 형식과 내용 면에서 통일성 있게 정제하는 것이다. 내용의 통일을 위해 기존의 공공언어 자료 감수 지침을 검토하여 실용적이고 활용 가능한 16개의 오류 유형을 판정하고 이를 기준으로 자료를 수정·보완하였다. 마련된 지침을 바탕으로 기 구축된 감수 전후 자료는 오류를 검토하여 재분류하고, 새로 수집된 자료는 오류를 찾아 분석하고 분류하였다. 형식적으로는 오류를 중심으로 앞뒤 10어절을 포함하는 문장을 단위로 하여 병렬 데이터의 형식으로 통일하여 구축하였다. 이 과정에서 실용성이 떨어지거나 일관성이 결여된 오류의 유형을 수정하고 실제적으로 활용 가능한 방향으로 주석 지침을 보완하는 작업을 진행하였다.

이 사업의 세 번째 과업은 구축된 데이터를 인공지능 자동 윤문 도구 개발을 위한 학습 데이터로 가공하고 학습 데이터로서의 가능성을 검증하는 것이다. 이를 위해 구축한 데이터를 활용하여 간단한 BERT 기반 모델을 훈련하였다. 이 과정을 통해 훈련 과정을 통해서 최종적으로 0.96에서 0.98의 F1 점수 성능을 보임을 확인하여 데이터의 활용 가능성을 파악하였다. 또한, 이러한 형식의 데이터를 효과적으로 구축하기 위해 웹 데이터베이스 기반의 작업 도구를 구성하여 운영하였다.

물론 최종적으로 공공언어 자동 윤문 시스템을 구축하기 위해서는 향후 본 연구의 성과를 보완하는 추가적인 작업과 연구가 수행되어야 할 것이다. 다만 본 연구의 결과물이 공공언어 감수 자료로 현장에서 적극적으로 사용되고 공공언어 감수 내용 연구의 기초 자료로 다양하게 활용되어 수정과 보완이 거듭되기를 기대한다.

[주요어] 공공언어, 공문서, 정확성, 오류 분석, 병렬 말뭉치, 자동 윤문, AI 학습 데이터

[목 차]

[표 목차]

[그림 목차]

제1장 연구 목적 및 범위

1.1. 연구 목적

이 연구의 목적은 각종 공문서의 감수 전후 자료를 통합하여 정제하고, 이를 자동 윤문 도구 개발을 위한 AI 학습 데이터 형식으로 구축하는 것이다. 현재 공공기관에서는 여러 경로를 통해 정책용어, 보도자료, 공고문, 안내문 등 각종 공문서에 대한 감수를 제공하고 있으나 감수 전후의 자료가 산재해 있어 기존의 감수 자료를 효율적으로 활용하기에는 어려운 실정이다. 이에 기존의 감수 자료들을 통합하여 활용할 수 있는 환경을 조성함으로써 감수 전후 자료를 제공하고 활용하는 데 있어서의 비능률을 해소할 필요가 있다.

본 연구의 구체적인 목적은 다음과 같다. 첫째, 기존의 공공언어 감수 전후 자료를 수집하여 통일성 있게 정제한다. 둘째, 새로운 공공언어 자료를 수집하고 감수하여 기존의 감수 자료와 통합하여 구축한다. 셋째, 구축된 자료가 인공지능 자동 윤문 도구 개발을 위한 학습 데이터로 사용될 수 있도록 형식화한다. 넷째, 공공언어의 개선 방안과 향후 발전 방안을 제안한다.

첫 번째와 두 번째 연구 목적은 자료의 수집과 정제, 구축에 관한 것이다. 기존에 제공된 공공언어 자료의 감수 전후 자료 및 기존에 감수되지 않은 새로운 공공언어 자료를 수집하여 동일한 내용과 형식에 맞게 통합 구축한다. 오류의 유형과 분류를 일관성 있게 정제하고, 형식적으로도 통합하여 활용이 가능한 형식으로 가공한다. 이를 위해 기존의 공공언어 감수 지침과 오류의 양상을 망라하여 효율적인 오류 분류와 분석의 틀을 만들고 적용한다.

세 번째 연구 목적은 구축된 데이터가 공공언어 자료 통합 활용 시스템 구축을 위한 기초 자료로 활용될 수 있도록 형식화하는 것이다. 인공지능 자동 윤문 도구 개발을 위해서는 정제된 고품질의 학습 데이터가 필요하다. 이에 본 연구에서 구축한 데이터가 학습 데이터로 사용될 수 있도록 형식의 적절성을 검증하고 데이터의 품질 검증 방안 계획 및 보완 체계를 수립한다.

마지막으로 공공언어의 개선 방안과 공공언어 감수의 발전 방안을 제안하여 관련 업무의 능률을 제고함은 물론 궁극적으로는 사회 전반에 걸친 소통 장벽을 해소하기 위한 토대를 마련하고자 한다.

1.2. 연구 범위

이 연구의 범위는 크게 자료의 수집, 정제, 구축으로 나누어진다. 먼저, 수집은 크게 두 가지 공공언어 자료를 대상으로 한다. 첫째는 국립국어원에서 발주된 공공언어 감수와 관련된 사업의 결과물로 기존에 구축된 공공언어 감수 전후 자료이다. 두 번째는 이번 사업에서 새로이 수집한 공공언어 자료를 감수한 결과이다. 기구축된 공공언어 감수 전후 자료는 “공공언어 감수 지원 종합 자료 구축(2017, 김일환 연구책임)” 사업의 결과물로 구축된 2010~2016년의 공공언어 감수 전후 자료 1,551건과 2017~2022년의 공공언어 자료를 대상으로 구축된 감수 전후 자료 1,375건 중에서 선별된 786건을 더하여 총 2,161건으로 구성된다. 여기에 새롭게 수집된 공공언어 자료 2,839여 건을 포함하여 총 5,000건의 문서를 대상으로 하여 감수 전후 결과를 정제하고 형식과 내용을 통일하여 병렬 데이터의 형식으로 구축한다.

자료의 정제를 위해 기존의 오류 유형과 분류 결과 및 오류 양상을 재검토하여 실제적이고 활용 가능한 세 개의 대분류, 16개의 하위 분류로 구성되는 오류 유형으로 구성된 지침을 마련하였다. 오류 유형은 어문 규범, 문법, 어휘의 세 개의 대분류로 구성되고, 어문 규범은 다섯 개, 문법은 아홉 개, 어휘는 두 개의 하위 분류를 포함한다. 새롭게 마련된 지침을 토대로 기구축된 감수 자료 오류의 분류와 분석을 재조정하고, 새로 수집된 공공언어 자료는 오류를 분석하여 주석하였다. 이 과정에서 과거에 오분석된 것과 일관되지 않은 일부 분석 결과는 수정·보완하여 정제된 데이터가 되도록 하였다.

구축은 대상 자료의 형식과 관련된 것이다. 본 사업은 기존에 산재된 공공언어 감수 자료를 통합하여 활용할 수 있는 시스템을 마련하는 첫 번째 단계로서의 성격을 갖는다. 이를 위해 본 사업의 결과물을 향후 인공지능 자동 윤문 도구 개발을 위한 학습 데이터로 사용될 수 있는 JSON 형식으로 가공한 다음, 그 형식의 적절성을 검증한다.

본 연구에서는 기구축된 공공언어 감수 전후 자료와 새로 수집된 공공언어 감수 전후 자료를 통합하여 정제하고 이를 인공지능 자동 윤문 도구 개발을 위한 학습 데이터 형식으로 구축하고 그 적절성을 검증하는 것을 목적으로 한다.

제2장 자료의 수집과 오류 주석

2.1. 수집 대상 자료의 특성

이 연구의 수집 대상 자료는 크게 두 가지로 나누어진다. 첫 번째는 국립국어원에서 진행한 공공언어와 관련된 사업의 결과물이고, 두 번째는 이 사업에서 새로이 수집된 공공언어 자료이다.

첫 번째 자료는 국립국어원에서 2010~2022년에 진행한 공공언어와 관련된 사업의 결과물인 공공언어 감수 전후 자료이다. 여기에는 2010~2016년의 공공언어 자료를 대상으로 한 "공공언어 감수 지원 종합 자료 구축(2017, 김일환 연구책임)"의 결과물 1,551건과 2017~2022년의 공공언어 자료를 대상으로 한 감수 전후 자료 1,375건 중에서 선별한 786건이 포함된다.

먼저, 2010~2016년의 공공언어 자료를 대상으로 한 "공공언어 감수 지원 종합 자료 구축(2017, 김일환 연구책임)"의 결과물 1,551건은 [그림 1]과 같이 감수 전후가 병렬화되어 있고, 오류 유형이 당시 사업의 지침에 따라 분류되어 있다. 이 자료는 기존의 감수 결과를 재검토하고 본 사업의 목적에 맞게 마련된 지침에 맞게 오류 유형을 재분류하는 과정이 요구된다.

A	B	C	D	E	F
연도	파일명	이전 10어절	감수 대상/결과	이후 10어절	감수 유형 분류 결과
2016	20160104_달성군청_문화재안내	대구 용연사 목조아미타여래삼존좌상과	복장유물	1용연사 목조아미타삼존좌상은 용연사 극락	A2
			복장 유물		
2016	20160104_달성군청_문화재안내	도와 전라도 지역에서 활약한 조각승이다.1	복장유물로는	조성 발원문과 개금 발원문, 묘법연화경과	A2
			복장 유물로는		
2016	20160104_달성군청_문화재안내	역에서 활약한 조각승이다.1복장 유물로는	조성발원문과	개금 발원문, 묘법연화경과 화엄경 등의 전	A2
			조성 발원문과		
2016	20160104_달성군청_문화재안내	조각승이다.1복장 유물로는 조성 발원문과	개금발원문,	묘법연화경과 화엄경 등의 전적(책), 다라니	A2
			개금 발원문,		

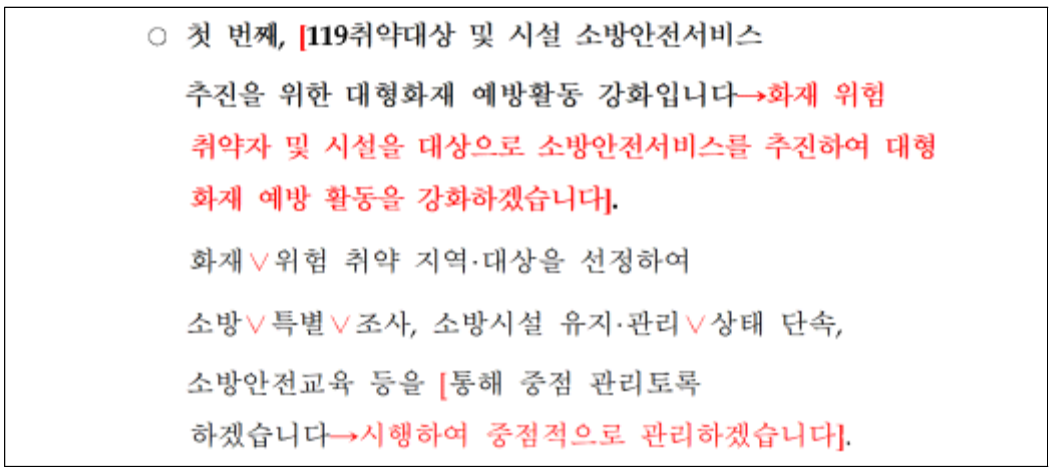
[그림 1] 2010~2016년 공공언어 자료 감수 전후

또한 이 과정에서 원문 자료가 잘못된 경우, 수정하거나 삭제하여 자료의 품질을 높였다. 아래의 [그림 2]는 2010~2016년에 구축된 데이터의 일부로, '장애인 사회 복귀 프로그램'이 '장애인 사회 귀 프로그램'으로 오입력된 사례이다. 오류의 유형과 오류 전후를 재확인하는 과정에서 이와 같은 오류 데이터를 수정 및 삭제하여 전체적으로 고품질의 데이터를 구축하였다.

[그림 2] 2010~2016년 공공언어 자료 오입력 사례

다음으로, 2017~2022년의 공공언어 자료를 대상으로 한 감수 전후 자료 1,375건 중에서는 786건을 선별하였다. 자료의 선별은 자료의 기존 감수 결과의 표시 여부, 자료의 형식적인 유형과 사용역 등을 고려하여 이루어졌다. 2017~2022년의 공공언어 자료는 [그림 3]과 같이 한글 파일에 감수 전과 후를 대괄호에 넣어 색을 달리 표시하는 방식으로 감수 전후가 표기되어 있고 오류의 유형은 분류되어 있지 않다는 특성이 있다.

본 사업팀에서는 기존의 감수 전후가 명백히 드러나는 문서를 위주로 선별하고, 일부 감수 전후의 표시가 분명하지 않은 자료는 제외하였다. 또한 일부 txt 형식 등 형식적으로 이질적인 문서도 처리를 고려하여 제외하였다. 이렇게 수집된 786건의 자료를 대상으로 하여 감수 전후를 재검토하고 본 사업팀에서 마련한 지침에 맞게 오류의 유형을 분류하였다.



[그림 3] 2017~2022년 공공언어 자료 감수 전후

두 번째는 본 사업팀에서 새로이 수집한 공공언어 자료 안내문 1,000건, 보도자료 2,000건이다. 모두 저작권과 무관한 자료(공공저작물에 특화된 자유 이용 허락표시)로 제한하여 수집하였다.

이렇게 수집된 기구축 자료와 새로 수집된 자료는 시기적으로 크게 두 단계로 구분된다는 특성이 있다. 먼저 2010~2016년의 자료이고 다음으로 2017~2022년의 자료와 새로이 수집된 자료이다. 가장 초기의 자료와 가장 최근의 자료는 약 10여 년의 시기적인 차이가 나게 되는데, 시기에 따라 공공언어의 변화도 일부 보여준다. 즉, 전반부의 자료는 어문 규범의 원칙을 주로 따르는 모습을 보이고, 후반부의 자료는 동일한 대상에 대해 허용 규정을 따르는 경향을 확인할 수 있다.

예컨대, 2010~2016년의 자료에서는 큰따옴표(“ ”) 안에 마침표를 찍는 원칙을 지키는 양상

이 주로 나타나는 반면, 2017~2022년의 자료에서는 마침표를 찍지 않는 허용의 양상이 높은 비율로 혼재되어 나타나는 경향을 보이는 것을 들 수 있다. “하여-” 형과 “해-” 형도 유사한 양상을 보인다. 앞선 시기의 자료에서는 “하여-” 형을 선호하는 경향이 있어서 일부 “해-” 형을 “하여-” 형으로 교정하여 자료의 통일성을 꾀한 모습을 보였으나, 뒤 시기의 자료는 원문에서 이미 “해-” 형식이 선호되는 양상이 나타난다. 또한, 최근의 신문기사도 “해-” 형이 주로 쓰여 변화의 양상을 방증한다.

2.2. 오류 분류와 주석

2.2.1. 오류 유형 분류

2.2.1.1. 오류 유형 분류의 기준

본 사업팀은 크게 두 가지를 고려하여 지침을 구성하였다. 첫째, 지침의 활용 가능성과 둘째, 공공언어 자료의 시대적인 흐름을 고려한 통일성이다.

첫째, 본고는 지침의 실용성을 고려하였다. 본서의 지침은 “공공언어 감수 전문가 양성을 위한 지침서(국립국어원, 2020)”와 앞서 언급한 기구축 공공언어 자료의 결과물 중에서 “공공언어 감수 지원 종합 자료 구축(2017, 김일환 연구책임)” 사업의 지침을 참고하되, 오류의 출현 빈도 및 지침의 적용 가능성과 인공지능 자동 윤문 도구 개발을 고려하여 마련하였다.

국어학적으로 지나치게 세분화된 지침은 이론적으로 정확할 수는 있으나 실제 이용의 측면에서는 활용도가 떨어질 수 있다. 이용자의 입장을 고려하면 지나치게 낮은 빈도로 출현하는 오류 유형과 구분이 어려운 것을 세세하게 구분하는 것보다는 실제적으로 활용이 가능한 지침이 효과적일 것이다. 가령, 어떤 말의 표기를 잘못한 경우 그 오류의 유형이 한글 맞춤법을 위배한 것인지, 표준어 규정을 위배한 것인지 구분하는 것은 일반 언중들에게는 크게 유의미하지 않다. 결국, 어문 규범에 맞지 않는다는 것은 동일하기 때문이다. 또한, 명사와 명사 띄어쓰기도 규범적으로 취급하기가 어려운 경우이다. 특히 공문서의 특성상 많이 사용되는 고유 명사와 전문 용어는 한글 맞춤법의 제5장 띄어쓰기 규정 제49항과 제50항에 단어별로 띄어 씀을 원칙으로 하되, 각각 단위별로 띄어 쓰거나 붙여 쓸 수 있도록 허용하고 있다. 고유 명사와 전문 용어는 명사와 명사가 결합되어 구성된 예가 많으며 실제 사용자들은 이를 붙여 쓰는 경우가 많다. 가령, 아래 (1ㄱ)의 ‘세계 태권도 연맹’은 2017~2022년 자료에서는 띄어 써야 하는 것으로 일부 지적되어 있으나, (1ㄴ)과 같이 해당 단체에서는 붙여 쓰는 표기를 사용하고 있다. 이는 고유 명사나 전문 용어에만 해당되지 않으며 일반 명

사구에서도 자주 지적되는 오류이다. 그러나 본 연구에서는 허용 규정을 따라 의미 해석상 크게 문제가 되지 않는 경우를 제외하고는 따로 명사와 명사의 띄어쓰기를 오류로 판단하여 수정하지 않았다. 허용 규정이 존재하는 고유 명사나 전문 용어의 띄어쓰기 오류를 걸러내는 것보다는 모국어 화자가 보기에 어색한 문장과 표현을 찾아 자연스러운 표현으로 고치는 것이 인공지능 자동 윤문 도구 개발에 더 필요한 작업이라고 판단하였기 때문이다.

(1) ㄱ. '세계태권도연맹' → '세계 태권도 연맹'

ㄴ. 해당 기관의 표기

또한, 지나치게 낮은 빈도로 출현한 오류의 유형을 세세하게 구분하는 것도 실제성의 측면에서는 유용하지 못하다. 예컨대, 2010~2016년의 공공언어 자료를 대상으로 오류의 유형을 분류한 "공공언어 감수 지원 종합 자료 구축(2017, 김일환 연구책임)" 사업의 결과를 분석한 결과에서 부적절한 사동은 0.2%, 부적절한 피동은 0.6%만이 출현하였다. 언어학적으로 유사한 범주의 오류이면서 출현 빈도가 낮기 때문에 이런 경우의 구분은 크게 유의미하지 않다고 판단하여 둘을 묶어 하나의 오류 유형으로 처리하였다. 이러한 판단에 따라 한글 맞춤법과 표준어 규정, 부적절한 피동과 사동 등과 같은 경우를 하나의 오류 유형으로 파악하여 기존 연구에서 분류되었던 오류 유형을 크게 16개의 유형으로 재분류하였다.(재분류한 오류 유형 2.2.2.2. 참고)

둘째, 공공언어 자료의 시대적인 특성을 반영하고자 하였다. 앞서 2.1절에서 소개한 바와 같이 본 사업의 대상 자료는 크게 2010~2016년 자료와 2017~2022년 자료의 두 시기로 구분되고, 시기의 구분에 따라 앞선 자료는 원칙 규범을 따르는 모습을, 뒤의 자료는 허용 규범을 따르는 양상을 보인다. 본 사업에서는 이러한 공공언어 자료의 시대적인 변화를 포착하기 위하여 지침을 두 가지로 나누어 마련하였다. 원칙 규범과 허용 규범이 모두 가능한 경우에 2010~2016년 대상 자료는 기존 자료의 고빈도 감수 전후 결과를 따라 원칙 규범 쪽으로 통일하는 지침을 마련하였다. 이와 달리, 2017~2022 자료는 이와 반대의 경향을 보여 허용 규범 쪽으로 통일하여 주석하는 지침을 구상하였다. 이는 공공언어 감수의 시대적 변화를 그대로 보존하고자 하는 태도에 따른 것이다.

더불어, 앞선 두 자료를 본 연구의 목적에 맞게 재검토하는 과정에서 일부 누락된 오류를 추가하고, 잘못 분석된 것은 바로 잡았다. 이를 통해 전체적으로 통일성을 기하여 데이터의

품질을 높이고자 하였다.

2.2.2.2. 오류 유형 분류

오류의 유형은 대분류 3개, 소분류 16개로 분류하였다. 대분류는 어문 규범 오류와 문법 오류, 어휘 오류로 구분된다. 어문 규범 오류에는 한글 맞춤법과 표준어, 띄어쓰기, 외래어 표기법, 로마자 표기법, 문장 부호 사용법 오류가 포함된다. 문법 오류는 부적절한 호응과 시제, 어순, 높임, 피/사동, 접속, 조사, 어미, 생략, 표현 오류로 구분된다. 어휘 오류는 외래어와 외국어, 한자어의 오남용과 잘못된 어휘 선택으로 분류하였다. 이를 간단히 정리하면 다음과 같다.

<표 1> 오류 유형 분류

대분류	소분류
어문 규범 오류	한글 맞춤법과 표준어 규정 오류
	띄어쓰기 오류
	외래어 표기법 오류
	로마자 표기법 오류
	문장 부호 사용법 오류
문법 오류	부적절한 호응(주-술, 목-술, 수식-피수식)과 시제
	부적절한 어순
	부적절한 높임
	부적절한 피/사동
	부적절한 접속(연결어미 오류)
	부적절한 조사
	부적절한 어미(종결어미, 전성어미 오류)
	부적절한 생략
	부적절한 표현
어휘 오류	외래어·외국어·한자어 오남용
	어휘 선택 오류

2.2.2. 오류 주석

2.2.2.1. 오류 주석 지침의 기본 원칙

여러 명의 연구원이 작업을 하는 만큼 오류 주석 지침의 기본 원칙을 다음과 같이 세워 교육 훈련을 하였다. 세부적인 지침은 문제가 발생할 때마다 공동 연구원의 회의를 통해 결정하였다.

첫째, 오류 주석의 대상(수집 자료에서 잘못된 부분)을 기준으로 유형을 분류한다. '오류'의 유형이므로 잘못된 부분의 유형을 나누는 것이다.

둘째, 각각의 항목에 해당하는 오류를 먼저 분류하고, 그 어느 범주에도 속하지 않는 것이나 있는 것을 삭제한 경우, 그리고 문장의 여러 부분을 함께 수정해야 바른 표현이 되어 오류를 특정하기 어려운 경우에만 부적절한 표현으로 분류한다. 이는 '부적절한 표현'이라는 다소 모호한 영역 안에 많은 유형이 뒤섞여 포함되는 것을 지양하기 위함이다.

셋째, 중복되는 오류는 다음과 같이 주석한다.

(1) 기구축 자료는 이미 오류의 주석이 되어 있으므로 문법성이나 의미에 큰 영향을 미치는 오류를 우선으로 주석한다.

(2) 신규로 수집된 자료는 하나의 표현에 여러 개의 오류가 포함되어 있는 경우 각각의 유형으로 모두 주석한다.

넷째, 어문 규범에 원칙 규정과 허용 규정이 공존하는 경우, 2010~2016년 자료는 '원칙' 규정 쪽으로 통일하고, 2017~2022년 자료와 새로 수집된 자료는 '허용'을 인정한다.

2.2.2.2. 오류 주석 지침과 사례

이 절에서는 본 사업팀에서 사용한 오류 주석 지침과 사례를 들어 보인다. 기본적으로 어문 규범과 한국어 문법에 준하되, 원칙 규정과 허용 규정은 자료의 시기와 출현 빈도, 언중들의 인식 등을 고려하여 융통성 있게 결정하였다. 또한, 많이 혼동하는 규범은 따로 밝히고 특별히 주의가 요구되는 것도 개별적으로 추가하여 통일된 작업을 할 수 있도록 안내하였다.

[1] 어문 규범 오류

(1) 한글 맞춤법, 표준어 규정 오류

① 명백한 오타

(예) 사용하느 → 사용하는

서울시민들게 → 서울시민들에게

② 비속어, 은어, 방언 어휘 등 사용한 경우

(예) 대궁이 → 줄기가

곱은자 → 곱자 형태의/ 'ㄱ'자 형태의

※ 부적절한 어휘2와 혼동하지 않도록 주의한다. 원문에 출연한 어휘가 비속어, 은어, 방언이면 "표준어 규정"을 위배한 것으로 판단한다.

(2) 띄어쓰기 오류

① 명백한 띄어쓰기 오류

(예) 네 모서리에서는지하의 → 네 모서리에서는∨지하의

홍길동 입니다 → 홍길동입니다

② -되다, -하다, -받다, -시키다 띄어쓰기

- 접미사로 쓰인 경우는 붙여 쓴다.

(예) 공부하다, 마련되다, 발전시키다, 무시당하다, 강요받다, 사랑받다, 용서받다, 말씀드리다, 불공드리다

- 단, 명사를 꾸미는 관형어가 명사 앞에 오면 '하다'를 띄어 쓴다.

(예) 말하다 → 필요 없는 말 하지 마라

- 본동사로 쓰인 경우는 띄어 쓴다.

(예) 편지 받다, 월급 받다, 선물 드리다, 용돈 드리다

※ 조사나 어미, 의존명사와 같은 문법요소의 띄어쓰기가 잘못된 것만 교정한다.

- 명사 + 명사 띄어쓰기는 교정하지 않음.

(예) 임용^시험, 자격^요건

- 규범상 허용된 것은 교정하지 않음.('-아/어 보조용언, 아라비아 숫자 + 단위명사 등)

① 본용언+--'아/--어'+보조 용언, 관형사형 어미+ 의존 명사+-하다/싶다

(예) 먹어보았다 → 먹어 보았다(교정 X)

적어둘만하다 → 적어둘 만하다('만하다'만 교정)

되어가는듯싶다 → 되어가는 듯싶다('듯싶다'만 교정)

② 아리비아 숫자 + 단위(원 등)

(예) 20000원 cf) 이만∨원

(3) 외래어 표기법

(예) 프랭카드 → 플래카드

카톨릭 → 가톨릭

팔로워 → 폴로어

※ 영문으로 표기하는 단위 명사(m, km 등)는 관례에 따라 한글 표기를 병기하지 않는 것으로 통일한다.

(예) 10킬로미터(km) → 10km

100m(미터) → 100m

(4) 로마자 표기법

(예) yangnyeom kejang → yangnyeom gejang

(5) 문장 부호 사용법

① 쓰여야 할 문장 부호가 누락된 것

(예) 등본초본 → 등본 · 초본

② 문장 부호가 잘못 쓰인 것

(예) OTT; → OTT: OTT 서비스(over-the-top media service)는~

③ 고유어나 한자어에 대응하는 외래어나 외국어 표기의 준말을 나타낼 경우에는 대괄호 [ ]로 표기한다.(어문 규범 묶음표의 대괄호 규정 참조)

(예) WHO → 세계보건기구[WHO]

FTA → 자유 무역 협정[FTA]

④ 쌍점 앞은 붙이고 뒤는 띄어 쓴다.

(예) 쌍점: 쌍점은~

※ 다음은 자료의 시기에 따라 허용 규정과 원칙 규정을 구분한다.

① 2010~2016년 자료는 “ ” 안에 마침표를 찍는 것으로 통일한다.

(예) “기쁘다”고 → “기쁘다.”라고

[부적절한 조사]와, 큰따옴표(“ ”)안 마침표 생략을 모두 오류로 판단하고 원칙 규범에 준하여 수정한다.

② 2017~2022년 및 신규 수집 자료는

(예) “기쁘다”고 → “기쁘다”라고

‘부적절한 조사’만 오류로 주석. 큰따옴표(“ ”)안 마침표 생략은 허용 규정이므로 오류로 주석하지 않는다.

※ 문장 부호에 포함된 띄어쓰기 규칙이 틀린 경우는 띄어쓰기 규칙으로 분류한다. 이는 다른 어문 규정도 기본적으로 띄어쓰기를 포함하고 있는 것과 동일하다. “단어와 단어는 띄어 쓴다, 조사와 어미는 붙여 쓴다” 등에서 조사를 띄어 썼다고 조사의 오류로 분류하지 않는 것과 같은 원리이다.

※ 문장 부호가 있어야 하는데 누락된 것은 부적절한 생략이 아니라 문장 부호 오류로 본다.(생략 오류는 문장 성분 누락이나 문법 요소의 누락을 주로 포함한다.)

[2] 문법 오류

※ 문법 오류 판단의 순서

먼저 ① 비문 여부를 판단한다 → ② 비문의 판단 근거가 되는 유형을 찾아 분류하고 수정한다 → ③ 어떤 유형에도 포함되지 않는 경우에 (9) 부적절한 표현으로 분류한다.

※ 감수된 문장 전체의 호응이 문제가 아니라 해당 문법 요소 부분만 고치면 각각의 해당 부분의 오류(조사, 어미, 피사동 등)로 처리한다.

문법 오류 판단의 예를 보이면 다음과 같다. (예 1)은 ①과 ②의 순서에 따라 조사 오류로 판단하여 수정한 사례이고, (예 2)는 ①과 ②에서 판단한 근거가 여덟 가지 문법 오류의 유형에 해당하지 않아 부적절한 표현으로 분류한 사례이다.

(예 1) 제시문: 저출산으로 인해 신생아 수가 급격하게 줄어들고 있음에도 가족제대혈 보관은 줄어들지 않고 있어 신생아수 대비 가족제대혈 보관 비율을 증가하고 있다.

① 제시문의 비문 여부를 판단한다.

② 비문의 판단 근거를 찾고, 그 유형을 분류한다.

'보관 비율을 증가하고 있다 → 보관 비율이 증가하고 있다': 조사 오류로 판단하여 분류한다.

(예 2) 제시문: 장기요양기관 217개 중 36개에 대해 건강보험공단이 현지조사한 결과, 34개 기관(94.4%)에서 약 30억 원의 요양급여를 부풀려 청구한 것이 사실이 드러났다.

① 제시문의 비문 여부를 판단한다.

② 비문의 판단 근거를 찾고, 그 유형을 분류한다.

'요양급여를 부풀려 청구한 것이 사실이 드러났다 → 요양급여를 부풀려 청구했다는 사실이 드러났다': 오류로 판단한 '-는 것이'는 조사, 어미, 피사동 등의 문법 오류의 유형에 속하지 않으므로 (9) 부적절한 표현으로 분류한다.

본 사업팀에서 분류한 문법 오류의 유형을 간단히 보이면 다음과 같다.

(1) 호응(주어-술어, 목적어-술어, 수식어-피수식어 등)과 시제 오류

① 주어와 서술어의 호응 문제

(예) 초가지붕은 여름에는 뜨거운 태양열을 막고, 겨울에는 효율적으로 열을 가두는 아주 실용적인 집이었습니다. → 초가지붕은 여름에는 뜨거운 태양열을 막고, 겨울에는 효율적으로 열을 가둡니다.

② 목적어와 서술어의 호응 문제

(예) 고양이가 아이와 똑같은 몸짓과 표정을 하는 까닭을 친구들과 이야기하여 봅시다. → 몸짓을 하고 표정을 짓는

휴대품에 대해서는 엄격한 휴대 수화물에 대한 규정을 한다. → 적용한다.

③ 수식어와 피수식어의 호응 문제

(예) 베트남 어린이들은 설에 우리나라처럼 세뱃돈을 받는데 행운과 부를 상징하는 붉은색 봉투에 받습니다. → 베트남 어린이들은 설에 우리나라 어린이들처럼 세뱃돈을 받는데 행운과 부를 상징하는 붉은색 봉투에 받습니다.

④ 시제 오류

※ 시제 선어말어미 '-었-', '-느(는)-', '-겠-' 이 관련된 경우가 포함된다.

※ '-고 있-' 사용 오류도 시제 오류에 포함하여 다룬다.

(예) 옛날이야기에 나오는 주인공들은 어려운 문제를 어떻게 풀 수 있을지 생각해 봅시다. → 옛날이야기에 나오는 주인공들은 어려운 문제를 어떻게 풀 수 있었는지 생각해 봅시다.

(예) 정월 대보름날 밤에 하던 놀이로, 이날 다리를 밟으면 일 년간 다릿병을 앓지 않았다고 하여 많이 함. → 정월 대보름날 밤에 하던 놀이로, 이날 다리를 밟으면 일 년간 다릿병을 앓지 않는다고 하여 많이 함.

(2) 부적절한 어순

※ 형식은 거의 바뀌지 않고 어순만 바뀐 경우

(예) 이번 주말에 하고 싶은 일을 연필을 바르게 잡고 바른 자세로 앉아 써 봅시다.
→ 연필을 바르게 잡고 바른 자세로 앉아, 이번 주말에 하고 싶은 일을 써 봅시다.

(예) 동글이는 예쁜 나의 친구입니다. → 동글이는 나의 예쁜 친구입니다.

(3) 부적절한 높임

※ 높임과 관련되는 조사, 어미가 나타난 경우

※ 높임과 관련된 어휘 오류는 어휘2 오류(부적절한 어휘 선택) 오류로 처리한다.

(예) 어떤 분들이 생명을 구하기 위해서 애쓰시고 계실까요? → 어떤 분들이 생명을 구하기 위해서 애쓰고 계실까요?

(4) 부적절한 피동과 사동

※ 사동의 '-이/히/리/기/우/구/추', '-시키다', '-게 하다', '-게 만들다', '-게 시키다'가 관련된 경우만 해당하고, 피동의 '-이/히/리/기-', '-어 지다', '-게 되다', '되다'가 관련된 경우만 해당한다. 이는 피사동의 형태를 규정하여 1차 감수에서 혼란을 줄이기 위함이다.

※ 특히, '○○하다'가 타동사인 경우 '○○시키다'를 쓰지 않는 것이 원칙이므로 주의한다.

(예) 금지의 말, 도움을 요청하는 말, 주의를 환기시키는 말을 듣고 말할 수 있다.
→ 금지의 말, 도움을 요청하는 말, 주의를 환기하는 말을 듣고 말할 수 있다.

(예) 자신의 현재 체력 수준을 알고 부족한 체력을 향상시키기 위하여 노력해야 합니다.
→ 자신의 현재 체력 수준을 알고 부족한 체력을 향상하기 위해 노력해야 합니다.

(예) 인천과 경기도로 인구가 분산되면서 서울의 통근권과 상권을 확대시켰다.
→ 인천과 경기도로 인구가 분산되면서 서울의 통근권과 상권이 확대되었다.

(예) 이처럼 전체 내용을 체계적으로 정리하면 효과적인 메모가 될 수 있다.

→ 이처럼 전체 내용을 체계적으로 정리하면 효과적으로 메모를 할 수 있다.

(5) 부적절한 조사

※ '께'와 '께서' 등 높임의 조사(문법 3)의 경우를 제외한 나머지 경우들이 여기에 해당한다. 높임과 관련된 조사는 부적절한 높임 오류로 포함한다.

(예) 단 하나의 위상으로 공급되는 교류. 전류의 크기와 흐르는 방향이 시간과 함께 주기적으로 변화하는 보통의 교류로써 일반 가정의 전등에 공급되는….

→ 단 하나의 위상으로 공급되는 교류. 전류의 크기와 흐르는 방향이 시간과 함께 주기적으로 변화하는 보통의 교류로서

(예) 말판에 제시한 내용대로 하지 못하면 말을 뒤로 한 칸 옮깁니다.

→ 말판에서 제시한 내용대로 하지 못하면 말을 뒤로 한 칸 옮깁니다.

(예) 그중에 한 가지를 정하여 사랑하고 보호하는 방법을 생각해 보고 실천해 봅시다.

→ 그중의(또는 그중에서) 한 가지를 정하여 사랑하고 보호하는 방법을 생각해 보고 실천해 봅시다.

(예) 어떤 사실을 분명하게 할 때에 '○을 박았다'라고 합니다.

→ 어떤 사실을 분명하게 할 때에 '○을 박았다'고 합니다.(※ 작은따옴표 뒤에 '고' 사용에 주의한다.)

(예) 어떤 사실을 분명하게 할 때에 "○을 박았다"고 합니다.

→ 어떤 사실을 분명하게 할 때에 "○을 박았다"라고 합니다.(※ 큰따옴표 뒤에 '라고' 사용에 주의한다.)

(예) 중앙대와 또는 고려대

→ 중앙대 또는 고려대(※ 원문에 잘못 쓰인 조사가 출현한 경우도 경우도 원문을 기준으로 하므로 '부적절한 조사'에 해당한다.)

(6) 부적절한 어미

※ 종결어미, 전성어미가 문제인 경우만 '부적절한 어미' 오류에 해당한다. 연결어미는 '부적절한 접속' 오류에서 다룬다.

(예) 그리고 부탁하는 내용에 알맞지 않는 까닭도 보이는데 고칠 부분을 같이 찾아보세요.

→ 그리고 부탁하는 내용에 알맞지 않은 까닭도 보이는데 고칠 부분을 같이 찾아보세요.

(예) 오늘날 이 제품을 쓰지 않은 곳이 거의 없다.

→ 오늘날 이 제품을 쓰지 않는 곳이 거의 없다.

(7) 부적절한 접속

※ 연결어미, 접속 부사를 수정하는 경우, 나열되는 항목들을 일관되게 하는 경우가 여기에 포함된다.

(예) 글쓴이의 나의 경험을 비교하여 글을 읽습니다. → 비교하며

(예) 행복 마을을 갖기 위한 방법과 도덕 공부의 방법을 비교하면서 그 내용을 공부해 봅시다.

→ 도덕 공부를 하는 방법(나열되는 항목을 일관된 형식으로 바꾼 경우)

(예) 중앙대 및 고려대

→ 중앙대와 고려대(※ '및'도 접속 기능을 하는 것으로 보아 부적절한 접속으로 분류한다.)

(예) 의논하거나와 협의를 통해 해결한다.

→ 의논과 협의를 통해 해결한다.(※부적절한 어미를 삭제한 경우에도 원문을 기준으로 '부적절한 접속'으로 처리한다.)

(8) 부적절한 생략

(예) 주변 인물을 면담하고 목적에 맞게 정리할 수 있는가?

→ 주변 인물을 면담하고 면담 결과를 목적에 맞게 정리할 수 있는가?

※ 줄임말도 부적절한 생략으로 포함한다. 이는 잘못된 어휘를 바른 어휘로 1:1 교체한 "어휘 오류 2"와 구분하기 위해서이다. 준말과 본말의 관계는 형태만 다르고 의미는 동일한 경우이다.

(예) "지자체" → 지방자치단체"

※ <표준>에 줄임말이 등재되어 있더라도 "공공언어 감수 전가 양성을 위한 지침서(2020)"을 우선하여 본말로 바꾼다.

※ "하여야"와 "해야"는 모두 허용한다. 2010-2016년 기구축 자료에서는 "하여야"를 사용하는 것을 규범으로 삼았으나, 최근 신문 기사 등에도 "해야"형이 많기 때문에 2017~2022년 자료와 신규 수집 자료에서는 모두 허용한다.

※ "허용 → 허용함"과 같이 어근만 남기고 어미를 생략한 경우도 부적절한 생략으로 본다.

※ 한글 → 한글(한자)도 "한자"의 부적절한 생략으로 본다(한자어 병기는 한글 표기만으로 의미가 정확하지 않거나 중의적인 경우에는 한자어를 병기하여 의미를 정확하게 전달하는 데 필요한 경우에 해당된다).

(예) 쇠를 초련할 때에는… → 쇠를 초련(初鍊)할 때에는

(9) 부적절한 표현

※ 문법 8가지가 아닌 경우, 있던 표현을 삭제하는 경우도 문법 9(부적절한 표현)로 처리한다.

※ 가능한 한 개별 오류로 우선 분류하고 분류가 어려운 경우에만 부적절한 표현으로 처리한다.

(예) 내역에 대한 기록은 해당 농산물의 출하 시점을 기준으로 1년 이상 기록·관리하여야 한다.

→ 내역에 대한 기록은 해당 농산물의 출하 시점을 기준으로 1년 이상 관리하여야 한다.

[3] 어휘 오류

(1)외국어 · 외래어 · 한자어 오남용

※ 외국어 · 외래어를 순화어로 고친 경우

※ 언중들의 인식이나 현실 언어의 사용 빈도보다 국어원의 다듬은 말을 우선한다.

(예) 지역 소개 팸플릿

→ 지역 소개 소책자

(예) 따라서 인터넷으로 만들어진 사이버 공간에서도 우리가 살아가고 있는 현실 공간에서처럼 바르게 행동하고 예절을 지켜야 합니다.

→ 따라서 인터넷으로 만들어진 가상 공간에서도 우리가 살아가고 있는 현실 공간에서처럼 바르게 행동하고 예절을 지켜야 합니다.

(예) 그림의 메뉴판에 토머스가 좋아하는 음식은 O표, 싫어하는 음식은 X표 해 봅시다.

→ 그림의 차림표에서 토머스가 좋아하는 음식에는 O표, 싫어하는 음식에는 X표 해 봅시다.

(예) 모토(motto)로 → 기치로

(예) 홈페이지를 → 누리집을

(2) 부적절한 어휘

※ 어휘를 1:1로 교체한 경우만 이 부류에 포함한다.

(예) 공을 피하기 위해 공이 있는 곳으로 몸의 방향을 바꾸어 봅시다.

→ 공을 피하기 위해 공이 있는 쪽으로 몸의 방향을 바꾸어 봅시다.

(예) 세계적으로 눈이 많이 분포하는 곳은 겨울 기온이 낮은 냉대와 한대 기후 지역이다.

→ 세계적으로 눈이 많이 내리는 곳은 겨울 기온이 낮은 냉대와 한대 기후 지역이다.

(예) 학교에 도착하자 수진이는 짝이었던 심술쟁이 종수를 아는 체도 하지 않습니다.

→ 학교에 도착하자 수진이는 짝이었던 심술쟁이 종수를 알은체도 하지 않습니다.

(예) 민들레씨의 얇은 실 끝에는 털이 여러 개 달려 있습니다.

→ 민들레씨의 가는 실 끝에는 털이 여러 개 달려 있습니다.

※ -들 부착/삭제 관계도 부적절한 어휘로 파악한다.

(예) 민들레씨들 → 민들레씨; 민들레씨 → 민들레씨들

※ “에 의해”류를 “에 따라”로 고치는 것도 부적절한 어휘에 포함한다.

[기타]

(1) 중복 오류는 모두 별도의 유형으로 주석한다.

(예) 제시문: → → 이태원 참사와 더불어 광주시 겪었던 동구 학동 참사, 화정 아이파크 붕괴 사고를 교훈삼아

① 부적절한 생략

이태원 참사와 더불어 광주시 겪었던 동구 학동 참사 및 화정 아이파크 붕괴 사고를 교훈삼아

→ 이태원 참사와 더불어 광주시가 겪었던 동구 학동 참사 및 화정 아이파크 붕괴 사고를 교훈삼아

② 부적절한 접속

이태원 참사와 더불어 광주시 겪었던 동구 학동 참사 및 화정 아이파크 붕괴 사고를 교훈삼아

→ 이태원 참사와 더불어 광주시 겪었던 동구 학동 참사와 화정 아이파크 붕괴 사고를 교훈삼아

(2) 비식별화는 개인정보와 관련이 있는 것만 아래와 같이 처리한다.

* 이름(살아 있는) → $실명$, 전화번호 → $전화번호$, 주소 → $주소$, 계좌번호 → $계좌번호$

다만, 공무원의 이메일은 공개된 정보로 판단한다.

작업 중에 추가로 비식별화할 항목이 나오면 추가한다. 예) 주민번호

2.3. 오류 주석의 결과

2.3.1. 오류 주석 작업의 진행

실제 오류 주석 작업은 앞서 서술한 오류 분류 및 주석 작업 지침을 기준으로 교육받은 작업자가 직접 검토하는 방식으로 진행하였다. 작업 과정은 '입력 - 검토 - 확인'의 3단계로 진행되었으며, 2단계 '검토' 과정에서는 1단계 '입력' 과정의 결과물 전체를 검토하였고, 3단계 '확인' 과정에서는 작업 차수와 오류 유형을 기준으로 2단계 '검토' 결과물의 일부(10% 이상)에 대한 최종 검수를 진행하였다. 각 과정에는 해당 과정을 효율적으로 진행하기 위해 별도의 온라인 데이터베이스 기반의 워크벤치가 개발되어 사용되었으며(3.2절과 4.1.1절에서 상술), 1단계 '입력'의 경우 작업물의 유형에 맞추어 총 세 개의 워크벤치가 구축되어 사용되었다. 이 과정에서 1단계 '입력' 작업자의 경우 작업에 부적절하다고 판단한 문서에 대한 제외 작업을 진행할 수 있었으며, 2단계 '검토' 및 3단계 '확인' 작업자의 경우 이전 작업자의 결과물이 부적절한 경우 해당 사례를 삭제, 처리할 수 있었다. 아래에서 제시하는 통계와 과제의 결과물인 최종 결과 데이터 등에서는 해당 삭제 사례 및 제외 문서들을 배제하였다.

구체적인 주석 작업 중 1단계 입력 작업은 수집 자료의 특성에 맞추어 진행되었다. 수집 자료 중 기관에서 제공한, 2010년에서 2016년의 자료에 해당하는 검수용 자료의 경우에는 이미 사례 단위로 정리되어 있기 때문에, 해당 사례에 대한 검수 및 이번 사업에서 구성한 오류 유형으로 변환하고 확인하는 순서로 검수를 진행하였다. 반면 기관에서 제공한 2017년~2022년 자료에 해당하는 기구축 자료는 기관에서 검수한 내용이 비정형으로 hwp 등의 문서에 기재되어 있어, 해당 문서 파일에서 검수한 내용에 대한 사례 단위 데이터화 및 오류 유형 부착 작업이 진행되었다. 이외에 본 사업팀에서 구축한 신규 자료들은 2010~20116, 2017~2022년 기구축된 자료에 준하여 본 사업의 작업 지침을 바탕으로 오류 및 교정 사례를 작성하고 유형을 부착하였다.

2.3.2. 오류 주석 작업의 결과

오류 주석 작업의 결과 총 5,031개의 문서에 대해 304,046개의 사례가 구축되었다. 해당 문서 중 기관에서 제공한 검수용 데이터는 1,536건(약 30.53%), 기관에서 제공한 처리용 데이터는 783건(약 15.56%)에 해당했고, 나머지 2,712건(약 53.91%)의 문서는 본 사업의 연구팀에서 지방자치단체의 안내문, 보도자료 등 공공기관에서 작성하여 민간에 배포한 대민 문서를 수집한 것이다.

전체 사례 304,046개에 대해서는 2.2.1.에서 정리한 바와 같이 정의한 오류 유형이 부착되었으며, 이러한 유형별 개수는 <표 2>와 같다. 이러한 사례 전체에 대해서 별도의 검수 작

업자들이 검수 작업을 진행하였으며, 이 결과로 구축된 사례들 중 문서 수를 기준으로 28.28%에 해당하는 1,423건의 문서에 대한 사례 17,635건에 대해 추가 확인 작업이 이루어졌다.

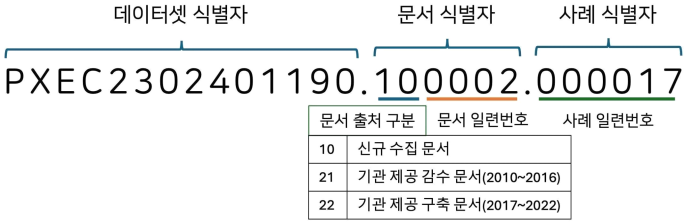
한편, 모든 사례에 대해서 형식 검증용 Python 스크립트를 구성하여 검증을 진행하였다. 저장된 값들의 자료형, 오류영역 값의 범위, 인코딩 및 유니코드 정규화 여부 등을 검증한 결과 전체 사례들에서 이러한 형식 오류가 발생하지 않았음을 확인하였다.

2.3.3. 데이터 셋의 구축

구축된 사례들은 향후 인공지능 학습에 유용하게 활용될 수 있는 형태임을 주안점으로 삼아 데이터 셋으로 구축하였다. 데이터 셋은 총 세 가지 버전으로 구축하였는데, 기준이 되는 JSON 데이터 셋과 더불어, 다량의 데이터 셋을 시스템 자원의 부하를 최소화하며 사용할 수 있도록 JSON Lines 버전의 파일을 보조로 구축하였다.

<표 2> 오류 유형별 사례 빈도

유형	개수	비율
표기 오류	**154,127**	**50.692%**
한글 맞춤법, 표준어 규정	3,000	0.987%
띄어쓰기	125,769	41.365%
외래어 표기법	1,497	0.492%
로마자 표기법	89	0.029%
문장 부호 사용법	23,772	7.819%
문법 오류	**110,755**	**36.427%**
호응 및 시제	2,143	0.705%
어순	1,985	0.653%
높임	588	0.193%
피동과 사동	3,011	0.990%
조사	9,124	3.001%
어미	2,556	0.841%
접속	18,338	6.031%
생략	31,327	10.303%
문법상의 표현 오류	41,683	13.709%
어휘 오류	**39,164**	**12.881%**
외국어 및 외래어 오남용	20,859	6.861%
어휘상의 선택 오류	18,305	6.021%
총계	**304,046**	**100%**



[그림 4] 사례 식별자의 구성

데이터의 식별자는 [그림 4]와 같이 문서와 사례에 대해 각각 부여되었다. 문서의 아이디는 총 6자리의 자연수로, 앞 두 자리를 통해 출처 정보를 표시하였고, 네 자리의 일련번호를 문서별로 부여하였다. 이때 출처 정보는 신규 수집 문서의 경우 10, 기관에서 제공한, 2010년에서 2016년 자료에 해당하는 감수 문서는 21, 2017년에서 2022년 자료에 해당하는 구축 문서는 22를 부여하였다. 이후 네 자리 일련번호는 삭제 여부와 무관하게 출처별로 부여되었다. 이후 사례에 대해서는 문서 및 사례 단위의 삭제 여부를 반영한 뒤, 각 문서에 해당하는 사례들에 대해 총 6자리의 자연수 일련번호를 부여하였다.

JSON 형식의 데이터는 기준이 되는 데이터로, 문서 당 하나의 파일로 구성하여 총 5,031개의 파일로 이루어졌다. 각 파일은 NFC 정규화를 적용한 UTF-8 인코딩으로 저장하여 범용성을 유지하였고, 관리를 위해 문서 식별자를 파일 이름으로 사용하였다. 각 파일은 <표 3>과 같은 구조로 정리하였다. 해당 데이터들에 대해서는 별도의 Python 스크립트를 통한 형식 검증을 진행하였고, 데이터의 구조 및 자료형에 대한 오류가 없음을 확인하였다.

<표 3> 개별 JSON 파일의 구조

1수준	2수준	자료형	내용	JSON Lines 포함 여부
docId		Text	문서 식별자	
docSource		Text	문서 출처 (<표 4>)	포함
docFName		Text	전달 문서명	
entity		Array	문서 내 사례	
	entityId	Text	사례 식별자	포함
	expr	Text	전체 텍스트	포함
	exprWithFix	Text	교정 적용 텍스트	
	errType	Text	오류 유형 (<표 5>)	포함
	errExpr	Text	오류 표현	
	errRangeBegin	Number	오류 영역 시작지점	포함
	errRangeEnd	Number	오류 영역 끝지점	포함
	fixExpr	Text	교정 표현	포함

<표 4> docSource 데이터 값 정의

값	정의
NEW	사업 중 신규 구축 자료
NIKL-1016	기관 제공 검수용 자료 (2010년~2016년 자료)
NIKL-1722	기관 제공 구축용 자료 (2017년~2022년 자료)

<표 5> errType 데이터 값 정의

분류	값	정의
NA	NA_1	한글 맞춤법, 표준어 규정
	NA_2	띄어쓰기
	NA_3	외래어 표기법
	NA_4	로마자 표기법
	NA_5	문장 부호 사용법
NB	NB_1	호응 및 시제
	NB_2	어순
	NB_3	높임
	NB_4	피동과 사동
	NB_5	조사
	NB_6	어미
	NB_7	접속
	NB_8	생략
	NB_9	문법상의 표현 오류
NC	NC_1	외국어 및 외래어 오남용
	NC_2	어휘상의 표현 오류

이와 별개로 인공지능 학습 등에 활용이 용이하도록 JSON Lines 형식의 보조 파일을 구축하였다. JSON Lines 파일은 파일 속 한 줄이 하나의 JSON object가 되도록 구성한 파일로, 동일한 구조로 정리되는 JSON object 자료형의 데이터가 나열된 형태의 데이터에 적용할 수 있는 형식이다. 자료 구조를 파악하기 위해 파일 전체를 파싱해야 하는 JSON과 달리

줄바꿈 문자를 기준으로 순차적 파싱이 가능하기에 제한된 메모리 자원으로 대규모의 데이터를 활용하기에 용이하다. 이러한 성질로 인해 유사한 구조의 대규모 데이터를 활용하여 훈련을 진행하는 상황에서 자주 활용되는 형식으로, Pandas, Arrow 등의 라이브러리 등과의 연동도 원할한 상황이다. 이와 함께 줄바꿈 단위가 데이터 단위와 일치하기에 bash 등의 도구를 통한 데이터 조작 등도 용이한 형식이다.

본 사업에서의 JSON Lines 파일은 JSON으로 구성한 데이터의 보조 데이터로, JSON과 동일하게 UTF-8 형식으로 구성되었으며, 사례 단위의 데이터 전체를 하나의 파일로 구성하였다. 하지만 모델 학습 용도의 사용을 목적으로 작성된 만큼 JSON 형식의 내용 전체를 포함하는 대신, <표 3>에서 별도로 표기한 최소한의 정보만을 포함하였다. 이러한 정보를 활용하여 모델의 훈련 방식에 맞춘 활용이 가능하며, 이를 실증하기 위하여 BERT 기반의 표현 오류 범위 지정 모델을 개발하였다.(3.3절 참고)

제3장 작업 시스템 구축 및 데이터 유용성 검증

3.1. 문서의 처리

3.1.1. 작업용 문서 형식의 정의

본 사업에는 공공기관에서 작성된 문건에 대한 교정 결과를 딥러닝을 비롯한 인공지능 훈련에 활용할 수 있는 형태의 데이터로 구축하고, 이와 유사한 방식으로 신규 문서에 대한 교정 데이터를 추가하는 과정을 포함한다. 이러한 목적을 효율적으로 달성하기 위해서는 작업 과정 전반에 걸친 파일 형식 및 작업 도구를 통일하고, 그 작업 과정을 실시간으로 판단할 방법의 구상이 필요하다.

사업 진행을 위해 기관에서 전달받은 파일은 기존 사업에서 정리한 엑셀 형식의 파일 7건과 교정 내용이 정형화되어 정리되지 않은 hwp 등의 파일들로 구성되었다. 이중 엑셀 형식의 파일을 제외한 파일들에 대해서는 파일의 형식을 제한하여 작업자가 유사한 형식의 파일들을 처리하도록 하는 것이 제한적이다. 이에 따라 기관에서 제공한 파일 중 hwp 및 hwpx 형식의 파일을 우선 작업하기로 하였다.

한편, 신규로 구성하여 작업할 파일에 대해서는 신규 구축 문서의 특성에 따라 형식이 정의되었다. 사업 수행 중 수집 가능한 공공기관의 작성문 가운데 수집 대상에 부합하는 분량

의 문서는 대다수 hwp(x)와 이를 보조하는 pdf 형식으로 확인되었다. 이중 pdf 형식은 어절 사이에서 줄이 바뀔 수 있는 한국어 표기법상 파일 안의 텍스트를 복사하였을 때 어절 단위가 끊어질 우려가 있고, 문서의 머리말 등이 복사 텍스트에 포함될 수 있는 만큼 수집 문서는 hwp 및 hwpx 형식으로 한정하였다.

3.1.2. 작업용 문서의 처리

본 사업의 작성 대상 중 엑셀 형식으로 전달받은 7개의 파일을 제외한 파일은 1) 폴더 구조 정리, 2) 원본-검수 파일 매핑, 3) 파일 식별자 부여, 4) 파일 변환의 과정을 거쳐 처리되었다. 이 과정은 기존 검수 자료에서도 적용되었으며, 원본-검수 파일 매핑을 제외한 과정은 신규 구축 자료에서도 동일하게 적용되었다.

폴더 구조 정리: 기관에서 전달받은 파일은 다소 복잡한 폴더 구조를 가졌으나, 파일의 일괄 처리에서는 폴더 구조가 단순한 것이 유리하다. 이를 위해 Python 및 Bash 스크립트를 구성하여 전달받은 파일 중 압축되어 있던 파일의 압축을 모두 해제한 뒤, 하나의 경로에 모두 저장하였다. 이때 복수의 경로 속에 같은 이름의 파일이 존재할 가능성이 있으므로, 파일들에는 임의의 경로 식별자를 부여하여 파일 처리상 오류를 방지하였다. 이 과정의 결과 발견된 1건의 ini 파일을 제외한 파일을 정리하여 검수가 완료된 문서 1,703건, 검수 이전 원본 문서 1,646건을 확인하였다.

원본-검수 파일 매핑: 기관에서 전달받은 검수 파일 중에는 화살표 등의 기호로 검수 및 교정 이전 단계의 원문을 유추할 수 있는 사례도 있었으나, 원문을 직접 수정하여 더 이상 유추할 수 없는 사례도 다소 발견되었다. 또한, 작업 진행에 검수본의 본문을 복사한 뒤 화살표 등의 기호를 삭제하는 것보다 검수본의 작업 대상 부분을 원문에서 찾아 복사하기를 선호한다는 작업자의 사전 의견도 있었다. 이에 원본 문서와 검수본 문서 사이의 파일 매핑 과정이 필요하였다. 이를 위해 원본 파일들의 파일명과 검수 파일들의 파일명을 육안으로 살펴보아 파일 이름 등을 기준으로 하는 매핑 휴리스틱을 설정하였고, 이를 활용하여 전달받은 문서에서 원본 문서의 89%에 해당하는 1,465개의 원본-검수본 쌍을 구성하였다.

파일 식별자 부여: 기관에서 전달받은 파일과 신규로 수집한 파일들은 파일 이름에 작성 기관 등의 정보가 담겨있어 길이가 상당히 긴 편이다. 한글 조합자로 이루어진 파일 이름이 길어질 경우 맥OS 등의 시스템에서 처리 장애가 발생할 수 있으며, 파일 이름을 통한 매칭에서도 효율성이 떨어지거나 예기치 못한 오류가 발생할 가능성이 있다. 또한, 파일에 UNIX 및 GNU/LINUX 시스템에서 시스템 용도로 사용하는 기호가 사용된 파일도 있어 시스템 범용성을 고려하였을 때 파일 이름을 일괄적으로 조정할 필요가 있다. 이에 “<분류>-<처리일자>-<처리순번>” 형식의 파일 식별자 구축 방식을 정의하였으며, 중복 식별자

를 방지하기 위하여 일자 단위로 파일 식별자 부여를 한 사람만 진행하도록 하였다. 가령, 2023년 11월 12일에 처리된 10번째 원본 파일은 “orig-20231112-00000010” 식별자를 부여받는다.

파일 변환: 파일에 식별자를 부여한 다음에는 파일의 사본을 생성한 뒤, 사본의 이름을 해당 식별자명으로 설정하였다. 이때 파일의 확장자는 전달받은 파일의 확장자를 모두 소문자로 변환하여 유지하였다. 이와 동시에 파일의 전달 당시 이름과 확장자, 경로명, byte 단위의 용량 등의 정보를 식별자와 함께 별도로 저장하여 식별자로부터 전달 파일을 추적할 수 있도록 하였다. 한편, 이러한 파일 중 hwp 및 hwpx 파일을 우선으로 모바일 등에서의 파일 열람이 유용하도록 파일에 대한 pdf 변환을 진행하였다.

한편, 7개의 엑셀 파일은 구체적인 문서가 존재하지 않으므로 위와 다른 형식의 처리가 필요하였다. 전달받은 엑셀 파일은 두 개의 행이 셀 병합을 활용해 하나의 사례를 구성하는 형태이다. 이처럼 셀 병합을 통해 구성된 파일은 화면을 통해 사람이 확인할 때에는 파악이 쉽지만, 컴퓨터 스크립트 등을 통해 처리한다면 csv 형식 등 하나의 행이 하나의 데이터인 경우에 비해 그 처리가 모호해진다. 이에 전체 엑셀 파일에 대해 Python 스크립트를 통해 데이터 추출을 진행하였다. 이 중 파일 이름에 해당하는 정보는 기타 파일들과 마찬가지로 식별 번호를 부여하여 향후 작업 도구에서의 일관성을 꾀하였다.

3.2. 웹 기반 작업 시스템 구축

3.2.1. 웹 기반 작업 시스템의 필요성

사업의 효율적 진행을 위해서 본 사업팀에서는 다음과 같은 필요성으로 인해 웹 기반 작업 시스템을 직접 구축하여 운영하였다.

중복 작업 방지: 여러 작업자가 동시에 작업을 진행할 경우, 동일한 내용을 중복하여 작업하는 경우가 발생할 수 있다. 이 경우 향후 작업 결과물을 정리할 때 중복 여부 판단과 데이터 일원화 과정 및 정책이 추가로 필요하게 된다. 이에 따라 온라인 데이터베이스 기반의 작업 시스템을 구성하여 작업 대상 및 결과를 실시간으로 공유하도록 함으로써 중복 작업을 방지할 필요가 있다.

작업 대상의 효율적 전달: 본 사업 과정에서는 다량의 문서가 작업자 사이에 오고 가야

한다. 이 과정에서 압축 등으로 파일을 전달할 경우 전달 과정의 오류가 발생할 수 있으며, 서로 다른 시스템 사이에서 여러 개의 사본이 발생하므로 파일의 형식, 버전 및 보안 관리의 문제가 발생한다. 이에 따라 작업 대상 파일을 웹 서버에 저장한 뒤, 작업 도구를 통해 이를 열람하거나 내려받게 함으로써 작업자가 필요한 경우에 항상 최신 버전의 파일을 내려받을 수 있게 할 필요가 있다.

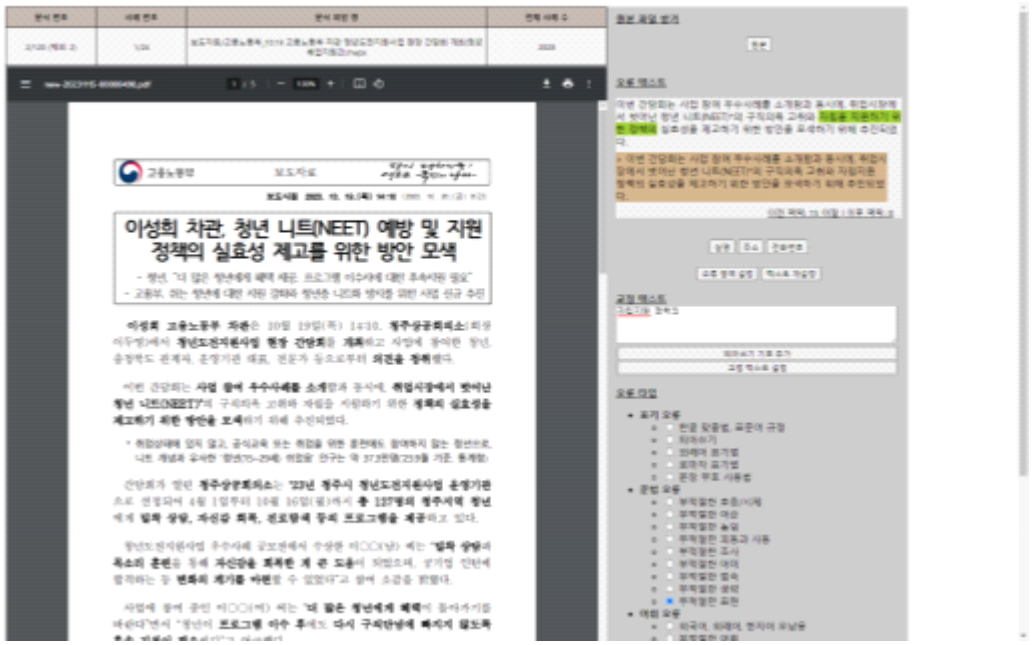
작업 형식의 정리: 작업자가 각자의 장비를 통해 엑셀 등의 상용 문서 프로그램 등으로 작업을 진행할 경우 프로그램 및 운영체제의 버전과 종류에 따라 인코딩과 파일 저장 형식에 관련된 문제 등 형식상의 이슈가 발생할 수 있다. 이러한 문제는 작업물의 양이 많아진 뒤에는 추적하여 해결하기 어려운 경우가 많아 작업 단계 자체에서 예방하는 것이 바람직하다. 하지만 사업 상황을 고려하였을 때 작업 장비의 통일 등은 불가능한 상황이다. 이에 웹 브라우저를 통해 접속하는 도구를 제공할 필요가 있다. 이러한 도구에서는 HTTP 프로토콜 안에서 인코딩을 UTF-8 유니코드(NFC 정규화) 등으로 표준화할 수 있었으며, 도구 내 웹 스크립트를 통해 작업 텍스트의 누락 등의 상황에 결과를 저장하는 대신 경고를 띄워 오류를 방지할 수 있다. 또한, 문자의 위치 정보와 같이 사람이 판단하기 어려운 정보 또한 웹 스크립트를 통하여 획득할 수 있으며, 작업자에게는 입력 정보를 렌더링한 미리보기 등을 제시하여 작업 저장 전, 후 확인 과정의 편의성을 제고할 수 있다.

작업 현황의 관리: 작업자가 각자의 장비에서 작업을 진행하다 보면 관리자가 작업 현황을 확인하기 위해 개인 작업자의 파일을 전달받거나, 개개인이 별도의 보고 과정을 거치는 등의 과정이 필요하다. 온라인 데이터베이스에 작업 결과물이 실시간으로 저장하게 하여 결과물에 대한 실시간 질적, 양적 검토가 가능하다.

3.2.2. 웹 기반 작업 시스템의 구성

작업 시스템은 MySQL 기반의 온라인 데이터베이스와, 데이터베이스에 대한 작업을 진행하는 PHP 내부 API, 그리고 작업 문서나 기존 작업물을 확인하며 jQuery 기반의 AJAX로 내부 API를 호출하는 HTML 인터페이스로 구성하였다.

[그림 5 표 형식의 작업 도구 인터페이스



[그림 6] 하나의 문서를 작업하는 작업 도구 인터페이스

온라인 데이터베이스: 문서 데이터베이스와 사례 데이터베이스로 구성하였다. 문서 데이터베이스에는 문서에 대한 정보와 함께 문서를 할당받은 작업자 정보가 저장된다. 작업자가 작업 도구에 접속하였을 때 이 문서 데이터베이스 속 정보를 통해 자신에게 할당된 문서를 자동으로 불러오게 된다.

작업 인터페이스: 총 네 가지 버전으로 준비하였다. 두 개의 문서를 표시하는 인터페이스에서는 작업자가 검수 완료본과 원본 문서를 보며 오류가 포함된 텍스트와 수정 표현 정보를 입력한다. 그 후 오류 유형을 선택한 뒤 사례 단위로 이동하거나, 문서 단위로 이동하여 작업을 지속한다. 신규 구축되는 문서들은 검수본 문서가 별도로 없기에 하나의 문서를 표시하는 인터페이스를 통해 유사한 작업을 진행하게 된다. 한편, 엑셀 파일로 전달받은 데이터 및 작업자가 이번 사업에서 구축한 데이터를 검수하기 위한 작업에서는 문서가 아닌 사례에 대한 작업이 이루어지므로, 사례들을 표 형식으로 표시하는 인터페이스를 사용한다. 이 인터페이스에서는 한 페이지에 다량의 사례를 표시한 뒤, 수정이 필요한 경우 더블클릭 등을 통해 직접 수정을 진행할 수 있게 하였다. 마지막으로 작업물들에 대한 최종 확인 목적으로 오류 유형 등의 조건을 적용하여 검수 완료 내용을 선별, 확인하는 도구를 구축하였다.

PHP 기반의 내부 API: 작업 인터페이스와 온라인 데이터베이스 사이의 통신을 담당한다. 즉, 데이터베이스 접속에 필요한 정보와 과업에 필요한 SQL 쿼리를 하나의 API로 구성하여 데이터베이스에 대한 접속과 작업 경로를 한정지었다. 이러한 API는 작업 인터페이스 주소 이외의 서버에서의 접근을 제한하여 작업 인터페이스를 통해서만 실행되게 하였다. 이를 통해 데이터베이스 접속에 필요한 정보를 웹페이지에서 숨겨 보안 측면에서의 이점을 얻을 수 있었다.

3.3. 데이터 유용성 검증

3.3.1. 구성 데이터와 검증의 필요성

본 사업을 진행하면서 사업 결과물이 딥러닝을 비롯한 방법으로 인공지능 시스템을 구축할 때 실제로 사용될 수 있도록 하는 것에 주안점을 두었다. 이를 위해 본 사업 결과물의 데이터는 메타정보와 함께, 1) 오류 표현을 포함한 20어절 가량의 텍스트, 2) 오류 표현의 수정 표현, 3) 텍스트 속에서의 오류 표현의 위치, 4) 오류의 유형으로 구성된다. 이를 통해 딥러닝 등의 훈련에 사용하기에 충분한 정보를 제공하면서, 오류 표현 전후 텍스트를 별도로 주는 등의 방식에서 발생할 수 있는 데이터 사용 장벽을 방지하였다. 즉, 오류 표현 전후 데이터 필드를 특정한 뒤, 공백 등의 조정과 함께 이를 덧붙이는 과정이 필요한 방식이 아닌, 오류 표현이 담긴 텍스트와 함께 Python과 JavaScript 등의 프로그래밍 언어에서 사용하는 방식으로 오류 표현을 쉽게 찾아 교정 텍스트로 변환할 수 있도록 하였다.

이러한 데이터를 구성하면서, 이를 통한 딥러닝 모델링이 유효한지에 대한 검증을 수행하였다. 이를 위해 상기 방식으로 정리한 데이터를 통하여 실제 오류 탐지 모델을 구축하였으며, 그 과정에서 데이터의 형식과 질에 관련된 문제가 있는지 점검하였다.

3.3.2. 모델링 방법

OpenAI의 생성형 언어모델 기반 상용 서비스의 등장 이후 다양한 작업을 이와 같은 모델로 처리하려는 움직임이 있으나, 공기관의 문서 교정 용도로는 적합하지 않다고 판단하였다. 이러한 텍스트 생성형 모델의 다수는 토큰들이 주어졌을 때 이어질 토큰의 확률을 추론하는 방식을 통해 훈련되며, 추론된 토큰들을 입력 토큰의 후미에 부착하여 추론을 계속 진행하면서 텍스트를 생성하는 별도의 알고리즘을 통해 구성된다. 즉, ChatGPT와 같은 서비스가 프롬프트에 따라 교정 등을 진행하는 것으로 보이나, 이는 입력된 교정 텍스트를 교정된 상태로 재생성 한 것이지, 교정 텍스트 자체를 유지시킨 채 오류 범위를 지정한 것은 아니다. 이처럼 교정 텍스트를 재생산하는 과정에서 교정 전의 텍스트를 그대로 생산한다는 보장은 없다. 한편, ChatGPT의 답변 내용을 시스템화하는 과정에서 필요한 답변 내용의 파싱 등의 문제와 지속적으로 발생하는 사용 비용, 사용에 필요한 교정 대상 데이터의 외부 기관 및 국외 반출 문제 등을 고려하였을 때 작성자가 직접 교정 제안을 판단하여 적용 여부를 결정할 교정 시스템의 목적에서는 기존의 자연어이해 모델 기반의 시스템이 적합하다.

이해 모델의 훈련 결과는 입력된 토큰에 대해, 각 토큰이 보고자 하는 범위에 해당하는지 여부를 판단하는 과정을 수행할 수 있어야 한다. 자연어처리에서 이러한 과정은 주로 토큰 단위 분류[Token Classification]으로 통용되며, 주로 개체명 인식[NER; Named Entity Recognition] 등에 활용된다. 텍스트의 오류 영역 탐지는 개체명 인식과는 내용이 다르지만, 텍스트 속에서 특정 범위를 지정한다는 점에서는 그 방법이 유사하다. 이에 따라 검증을 위한 모델은 일반적으로 사용되는 한국어 BERT 모델을 NER과 유사한 방식의 데이터로 파인튜닝하여 F1 지표로 판단하였을 때 어느 정도의 성능을 보일 수 있는지로 설정하였다.

3.3.3. 데이터 셋의 구성

모델의 훈련 방법과 토큰화 과정에 따라 본 사업에서 제안하는 데이터를 가공할 필요가 있다. 즉, 본 사업에서는 토큰 단위 분류를 통한 모델링을 구상하였으나 이러한 방법이 유일한 모델링 방법은 아니며, 사전 훈련 모델에 대해 훈련을 진행할 경우 모델이 사용하는 토큰화 과정에 따라 텍스트가 토큰화 되어야 한다. 이에 본 사업팀에서는 제안하는 데이터의 형식이 이 과정에 적합한지를 파악하였다.

데이터 셋은 토큰 단위 분류에 주로 사용되는 BIO 방식으로 구성되었다. 이를 위해 1) 텍스트 토큰화, 2) 오류 범위 토큰 설정, 3) 토큰에 따른 BIO 태그 부착의 과정을 진행하였다. 텍스트 토큰화 단계에서는 검증에 사용할 BERT 모델인 KR-BERT의 토큰화 과정을 활용하여 오류가 포함된 텍스트를 모델이 처리할 수 있는 단위로 토큰화하였다. 이때 토큰화는 어절 단위 및 형태소 단위 등을 무시하고 진행되어, 실제 오류 범위를 이 토큰 범위로 재설정할 필요가 있다. 이에 따라 텍스트 위치를 활용한 휴리스틱을 구축하여 오류 표현 범위를

어절 단위로 확장한 뒤, 그 어절에 해당하는 토큰들을 특정하였다. 이러한 토큰들에는 오류 표현의 시작 토큰에 B, 오류 표현이 종료된 다음 토큰에 O, 그 사이 토큰에 I 태그를 부착하였다. 이를 통해 모델은 각각의 토큰에 대해 B, I, O 중 하나의 라벨을 부착하게 되고, 이 태그 정보를 활용하여 텍스트 안에서의 범위를 특정할 수 있다. 한편, 훈련용 데이터의 구성 시, 표현이 새롭게 추가되어야 하는 경우는 뒤쪽을 범위로 잡아서 라벨을 부착하도록 데이터 셋을 구성하였다.

구체적으로 데이터 셋은 검수가 진행된 전체 데이터를 대상으로, 기존의 표현 기반 맞춤법 검사기가 검사 가능한 표기 오류를 비롯하여 놓치기 쉬운 표현 오류 및 문법 오류 데이터를 모두 포함하여 구성하였다. 검증의 주안점은 과정의 확인이어서, 모델 자체의 성능이 아니므로 데이터 셋 역시 표기 오류, 문법 오류, 표현 오류의 세 가지 유형을 각각 별도로 구성하였다. 즉, 하나의 모델은 세 가지 오류 중 하나를 탐지하며, 상세한 유형까지는 제시하지 않는다. 이는 유형별 비율 조정이 필요하고 이를 통해 효과적인 진행이 가능하기 때문이다. 향후 사업 내용을 정리, 발전하여 유형별로 다수의 사례를 비슷한 비중으로 준비할 수 있다면 B, I, O 태그를 유형별로 마련하여 분류함으로써 유형 범위 탐지와 분류를 동시에 진행할 수도 있다.

데이터 셋은 표기 오류에 대해 154,127건, 문법 오류에 대해 110,755건, 표현 오류에 대해 39,164건의 사례로 구성되었으며, 전 과정을 별도의 휴리스틱 없이 스크립트를 통해 자동화할 수 있었다. 이 과정을 통해 구상한 데이터의 형식이 모델링용 데이터 셋 구축에 유용함을 확인하였으며, 인코딩이나 파싱 오류와 같은 형식 문제도 잠재되어 있지 않음을 확인하였다.

3.3.4. 훈련의 진행 및 결과

훈련에 사용할 모델은 토큰 마스킹을 통한 언어 모델링 훈련[Masked Language Modeling]으로 구축된 KR-BERT 모델을 불러온 뒤, 소프트맥스 함수를 통해 입력 토큰 각각에 대해 B, I, O 태그에 대한 확률을 부여하는 레이어를 추가하여 구성하였다. 훈련은 이 모델의 모든 파라미터에 대한 조정을 진행하여 이루어졌으며, 표기 오류, 문법 오류, 표현 오류 각각에 대해 별도의 파라미터 조정이 진행되어 총 세 개의 모델이 구성되었다. 모델링 과정에는 PyTorch 프레임 워크가 사용되었으며, 자체 GPU 장비를 통해 데이터의 외부 전달 없이 훈련 및 평가를 진행하였다. 훈련 과정 중 목적함수에 따른 손실값과 정확도를 주기적으로 확인하였으며, 데이터 셋을 최대 세 번까지 보도록 훈련하였고, 한 번 훈련할 때 훈련에 사용하지 않은 라벨링 데이터 셋에 대한 평가를 진행하여 평가용 손실값을 계산하였다.

<표 6> 훈련 결과 지표의 요약

	표기 오류 탐지 모델	문법 오류 탐지 모델	표현 오류 탐지 모델
최저 손실값†	0.031	0.074	0.055
Precision‡	0.980	0.960	0.960
Recall‡	0.980	0.960	0.960
F1‡	0.980	0.960	0.960

†: 0점에서 무한대; 낮을수록 좋음; 0점은 예측값이 실제와 항상 일치할 경우
‡: 0점에서 1; 높을수록 좋음; 1점은 예측값이 실제와 항상 일치할 경우

훈련 결과는 <표 6>과 같다. 세 모델 모두 손실값이 0에 가까워질 때까지 훈련되었으며, 그 변화 또한 안정적으로 진행되어 훈련 자체가 차질 없이 진행됨을 확인하였다. 각 모델의 Precision, Recall, F1 지표는 B, I, O 라벨 간의 비중을 감안하여, 비율을 고려하여 계산된 수치로 표기 오류 탐지 모델에서 조금 더 높게 나타났다. 표현 오류 탐지 모델은 다른 두 모델보다 데이터의 양 자체가 작았고, O 라벨은 0.98 수준이었지만 B라벨은 0.69, I 라벨은 0.31 정도에 불과했음에도 불구하고, B, I, O 라벨을 전체적으로 보아 비중을 감안한 수치는 다른 두 모델에 근접하는 결과를 보여주었다. I 라벨의 수치가 작은 것은 F1 지표가 가장 높은 표기 오류 탐지 모델에서도 0.55 수준으로 나타나서, 표현 오류 탐지 모델만의 현상은 아니었으며, 이는 오류 유형들마다 데이터 셋 내에 I 라벨 자체의 비중이 떨어진 것이 원인으로 사료된다. 또한 향후 데이터 양이 늘어남에 따라 해결될 여지가 있다.

표기 오류에 대한 예측 결과의 예시는 아래와 같다. 각 사례들은 각각의 모델에서 모델 평가 시에 Validation set에서 무작위로 5개의 예시를 뽑아 추론을 진행한 것이며, 모델의 입력 단위이고, 실제 오류인 부분은 글자 색의 변화를, 모델이 오류라고 예측한 범위는 음영을 주어 구분하였다. 또한, 모델 출력 결과를 기반으로 구성하였으므로, 사용한 파인튜닝 모델의 토큰 목록에 없는, 특히 한자 및 글머리기호 등 특수기호로 구성된 토큰의 경우 미등록토큰(unknown token; 아래 예시에서는 [UNK]로 기재)으로 변환되어 처리되었으며, 결과 보고에서도 이를 반영하였다. $실명$과 $유선번호$ 등은 실제로는 그대로 입력되었으나, 문서 작성용으로 비식별화한 내용이다.

- 서식> 응 시 원 서<LF>재단법인 문화엑스포 이사장 귀하<LF>아래 기재사항은 사실과 다름없으며 만일 시험 결과에 부당한 영향을 끼칠 목적으로 허위[1)]
- 함량은 28~32%이며, 비타민 B군, 칼슘, 인, 소금의 함량이 높다. <LF>3. 고려사항: 비육돈 사료에 5% 이상 사용 시 기호성이 떨어지고 설사를 일으킬
- 1<LF>7<LF>100~200만 원 미만 700~800만 원 미만<LF>2<LF>8<LF>200 ~300만 원

1) 줄바꿈 기호는 데이터 구축 과정에서 CARRIAGE RETURN 등의 기호를 처리해 모두 LINE FEED 단일 기호로 정규화하였고, 모델링 및 테스트에도 UTF-8 영역에서 정의한 해당 기호(0x000A)가 사용되었다. 이곳에서는 지면상의 이유로 줄바꿈 기호가 있던 자리를 <LF>로 대체하여 표기하였다.

미만 800~900만 원 미만<LF>3<LF>9<LF>300~400만 원 미만

- 뾰족한 봉우리와 골짜기가 만들어졌다.(그림② 참조) 이곳에서는 성산 일출봉의 화산 분출, 퇴적과정, 그리고 그 후에 발생한 침식의 전 과정을 한눈에 볼 수
- 쌀은 일반 시장에서 판매할 수 없으니, 시중에 유통되는 것이 발견되면 검사기관(☎$전화번호$)에 신고해 주시기 바랍니다.<LF>생산 연도, 도정 연월일<LF>이 쌀은 우리

문법 오류에 대한 예측 결과는 아래와 같다.

- $실명$ 광주시장은 "시민의 안전한 일상은 광주다움의 완성이다. 이태원 참사와 더불어 광주시 겪었던 동구 학동 참사, 화정 아이파크 붕괴 사고를 교훈삼아 시민의 생명, 안전을 최우선으로 안전행정을 꾸려야 한다"며 "안전을 위한 행정의 책임이 무겁다.
- 비([UNK], 학명: Torreya grandis Fort.)의 씨이다.<LF>성상 비자나무 이 약은 씨로 긴 달걀 모양에서 타원형이다. 길이가 10~25mm이고, 지름이 약 10mm이다. 바깥
- 트래픽 소스와 목적지 사이에 계약된 총 트래픽 양의 상한.<LF>유·무선 인터넷 트래픽 급증, 통신 시장의 성장 정체, 플랫폼을 중심으로
- 눈을 감고 600년 전 한양으로 떠나 볼까? 흥인지문 주변의 거대 쇼핑몰,[2] 대형 시장을 하나하나 지워 본다. 흥인지문 주위를 오가는 온갖
- 당시 사람들이 곰과 같은 맹수를 사냥했다는 것을 알려준다. 또한 사슴의 이빨로 본 계절분석을 통해 구낭굴은 가을에서 봄에 이르는 계절에 사람들이 사냥하면서 생활한 유적이라는 것이

표현 오류에 대한 예측 결과는 아래와 같다.

- [UNK] 이번 $실명$ 국무총리의 카메룬 방문은 대한민국 국무총리로서 최초 방문이며, 양국간 전자정부, 농업·보건의료 등 분야별 협력은 물론, 국제무대 협력까지 강화해나가는 모멘텀을 제공한 것으로 평가된다. <LF><LF>ㅇ $실명$ 총리는 현지 시각 11.1(수) 밤 네 번째 순방지인 노르웨이로 이동할 예정이다.
- 이들 중 기증제대혈과 가족제대혈의 이식 공급 현황을 점검한 결과 기증제대혈은 최근 5년간 419건 치료목적으로 이식되었고, 가족제대혈을 이식한 것은 불과 9건에 불과했다.
- 유산 01-진주의 무형 문화 유산>을 발간합니다. <LF>이번 책자 발간을 위해 장시간에 걸친 조사와 인터뷰, 집필을 위해 노력해 주신 여러 연구자들과 조사에
- 공공심야어린이병원은 현재 광주기독병원에서 시범운영중이며, 9월 본격 운영에 들어가게 되면 기존 응급실 비용보다 저렴하게 이용 가능하다.
- 행동으로 표현할 때 군인답고 지도자다운 태도를 견지하여야 한다. 이를 위해 리더는 군인다움, 육체적 강건함, 주도성, 회복 탄력성 등을 갖추어야 한다.<LF>제

2) '거대' 다음에 '한'이 추가되는 사례에서 훈련용 데이터의 특성을 반영하여 뒤쪽을 올바르게 잡은 예시이다.

결과적으로 모델의 성능은 그 자체로 상용화하기는 어려우나, 데이터의 양과 사전 모델의 성능을 고려하였을 때 향후 더욱 큰 규모의 훈련을 통해 유의미한 모델링이 가능할 것으로 기대할 수 있다. 특히, 최소한의 처리를 통한 모델링의 결과로 1.00에 근접한 F1 점수를 보인다는 점은 입력한 데이터의 형식이 이와 같은 훈련 및 유사한 방식의 훈련에 적합하다는 것을 시사한다. 하지만 구체적인 오류의 분류 및 유형에 대한 분류를 진행하지 못한다는 점, 표기 오류의 예시에서의 복합명사 처리와 같이 본 사업에서 구축한 데이터의 특성을 그대로 반영해 학습하지는 못했다는 점 등을 보아 향후 데이터 속 클래스 사이의 비율 조정 및 증강을 통하여 보다 데이터에 적합한 모델링이 가능할 것으로 보인다. 또한, 토큰나이저의 조정 및 사전훈련 모델의 변경 등을 통해 현재 모델에서 미등록 토큰으로 처리하는 한자 및 특수기호 등의 처리 성능의 고도화를 기대해 볼 수도 있다.

제4장 작업 관리

4.1. 작업 관리 도구

본 용역의 경우 전체 사업 기간이 약 6주에 불과하여 단기간에 과업이 완수될 수 있도록 작업자의 작업 현황 및 작업 결과물을 긴밀하게 관리하는 것이 관건이 되었다. 또한 여러 명의 작업자가 감수 전후 데이터를 산문형으로 변환하고 데이터를 병렬화하는 작업을 수행하기 때문에 형식의 일관성을 보장할 수 있는 도구의 사용이 요구되었다. 이를 위해 본 사업팀은 별도의 워크벤치(Workbench)를 구축하여 작업을 진행하였으며, 워크벤치상의 작업 현황을 실시간으로 파악할 수 있는 별도의 확인용 페이지를 마련하였다. 더불어 작업 관리 전담 연구원을 배치하여, 양적·질적 측면에서 일 단위 작업 현황을 파악하고 일간 보고서를 작성하였다.

4.1.1. 워크벤치 구축

워크벤치는 다수의 사용자가 동시에 작업하고 변경을 공유하는 것을 가능하게 하는 도구이다. 이는 데이터 구축과 관련된 모든 작업을 하나의 통합된 환경에서 수행할 수 있다는 점에서 매우 유용하다. 또한 누가, 언제, 어떤 변경을 가했는지 파악할 수 있다는 점에서 데이터 버전을 관리하고 변경 내역을 추적하는 데에도 효과적이다. 형식화 측면에서 데이터 관계, 문자 인덱싱 등 반복적이고 일괄적인 작업을 효율적으로 수행할 수 있으며, 데이터 구축 프로세스의 오류 가능성을 줄일 수 있다.

이러한 측면을 고려하여 본 사업팀은 데이터 구축과 검수 각각의 작업 특성을 고려하여 작업자용 워크벤치와 검수자용 워크벤치를 구분하여 제공하였다. 각각의 작업자 및 검수자에게는 주차별로 고유한 워크벤치 페이지가 할당되었으며, 각 계정에 작업 수행 계획에 기반하여 작업자별로 할당된 작업 문건이 배포되었다. 더불어 각 워크벤치에서 작업한 내역과 작업된 오류 유형 집계를 확인할 수 있는 작업 종합 현황 페이지를 구축하여 작업 관리에 활용하였다.

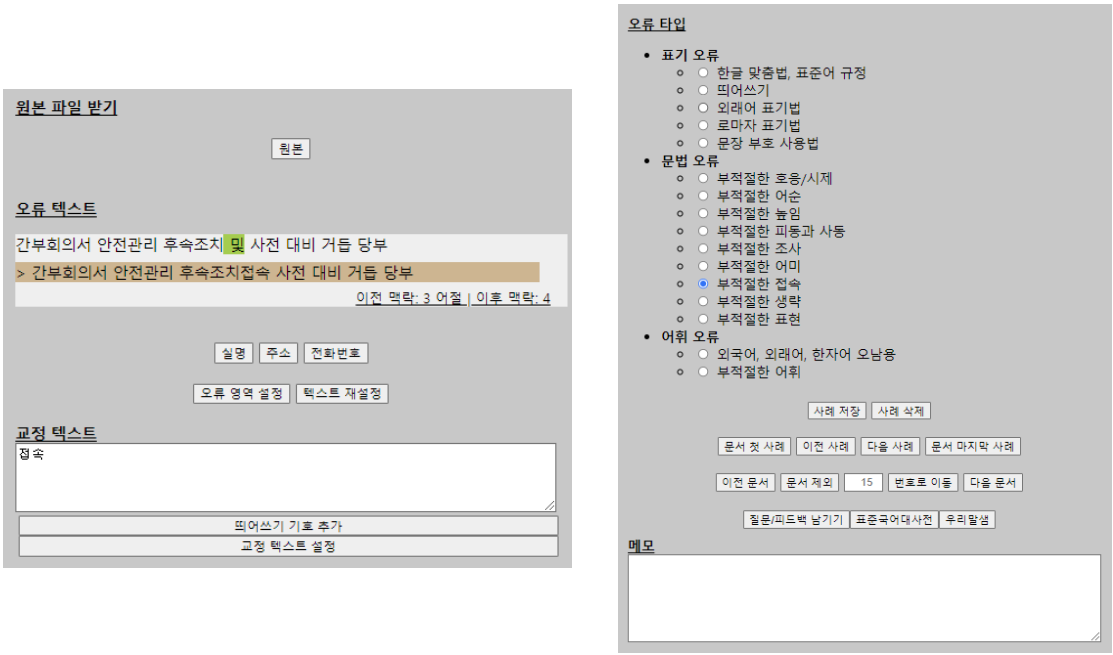
4.1.1.1. 작업자용 워크벤치

[그림 7]은 신규 데이터 구축 작업 및 2017-22년 기구축 자료 검수 등 1단계 작업자들에게 제공된 워크벤치 화면이다. 화면의 왼편에는 작업해야 하는 문서가 나타나고, 화면의 오른편에는 해당 문서에서 발견된 감수 내역을 입력하게 된다. 입력하는 사항으로는 ①감수 전 표현, ②감수 후 표현, ③해당 감수 사례의 오류 유형이다. 오류 유형은 중복으로 입력되지 않으며, 선택 기준은 작업 지침의 내용에 따른다. 구축 데이터에 포함되는 감수 표현 전후 10어절은 작업자가 직접 입력하는 것이 아니라 감수 전후 사례를 병렬화하는 과정에 자동으로 입력될 수 있도록 처리하였다.

[그림 7] 작업자용 워크벤치 화면 구성 1

표기한 오류 유형은 저장 또는 삭제할 수 있으며, 작업하는 과정에서 질문이 생기면 하단 메모란에 질문 및 피드백을 남길 수 있도록 하였다. 더불어 신규 문서 구축 작업자들이 올바른 감수를 수행하는 데 도움이 될 수 있도록, <표준국어대사전> 및 <우리말샘> 홈페이지 화면을 바로 연결하여 볼 수 있는 기능도 제공하였다. 한편 왼쪽 화면에 제시된 검수 대상 문서가 데이터 구축에 적합하지 않은 사례가 있을 수 있다. 일례로 2017년-22년 기구축 자료의 경우 일부 감수 자료가 수기로 입력되어 가공에 어려움이 있다. 따라서 해당 경우 작업자가 제시된 문서를 '제외문서'로 분류할 수 있도록 하였다.

오류의 수정은 원문([그림 7]의 왼쪽 부분)에서 오류를 찾아 해당 영역을 복사하여 붙여넣은 다음 교정 텍스트 부분에 수정된 표현을 입력하고 오류의 유형을 선택하는 방식으로 구성된다([그림 8]). 해당 작업이 끝나면 사례를 저장하고('사례 저장') 다음 문서로 이동하는 방식('다음 문서')이다.



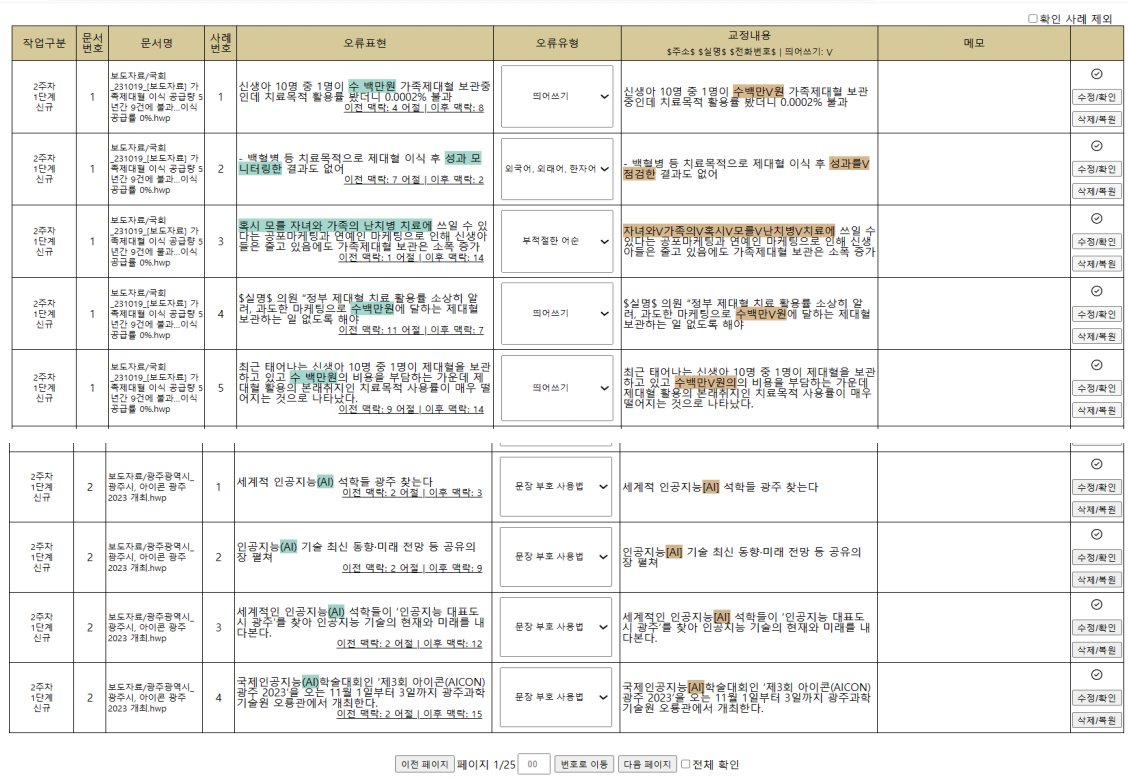
[그림 8] 작업자용 워크벤치 화면 구성 2

4.1.1.2. 검수자용 워크벤치

1단계에서 구축된 데이터는 2단계, 3단계 총 두 차례에 거쳐 검수가 진행되었다. [그림 9]는 2단계 검수자에게 제공된 워크벤치 화면이다. 1단계에서 작업 완료된 문건에 한하여 2단계 검수 작업자들의 페이지에 할당되어 1단계 작업 문서의 작업 내역을 검수할 수 있도록 하였다. 검수자용 워크벤치의 경우, 검수 전후 내역 및 각 사례별로 할당된 오류 유형을 한눈에 볼 수 있도록 데이터 프레임 형태로 화면을 구성하였다. 감수 전 오류 표현은 파란색으로 범위가 표시되며, 감수 후 표현은 주황색으로 범위가 표시된다.

검수자는 검수 전후를 비교하여 오류 유형의 적절성을 평가하고, 각각의 사례마다 '수정/확인' 버튼을 눌러 검수 완료 여부를 저장하도록 하였다. 1단계에 작업된 내역 중 오류 유형이나 감수 후 교정 내용에 오류가 발견되었을 시 검수자 화면에서 즉시 수정이 가능하도록 하였다. 만약 1단계의 작업 내역이 적절하지 않다고 판단되는 경우, 삭제 버튼을 눌러 작업 건수에 포함되지 않도록 하였다. 검수자 역시 마찬가지로 추가 검토가 필요하거나 보고서에 참고할 만한 사항은 메모란에 남기도록 하였다.

2단계 검수 작업이 완료된 데이터들은 최종적으로 별도의 도구([그림 10])를 통해 오류 유형별로 정리하여 3단계에서 최종적으로 확인할 수 있도록 하였다. 해당 도구에서는 [그림 10]과 유사하게 사례 단위로 작업물을 1) 작업 차수, 2) 문서의 구분, 3) 오류의 유형을 기준으로 2단계 작업 완료 데이터에 한하여 선별하여 확인 및 수정할 수 있게 하였다.



[그림 9] 검수자용 워크벤치 화면 구성

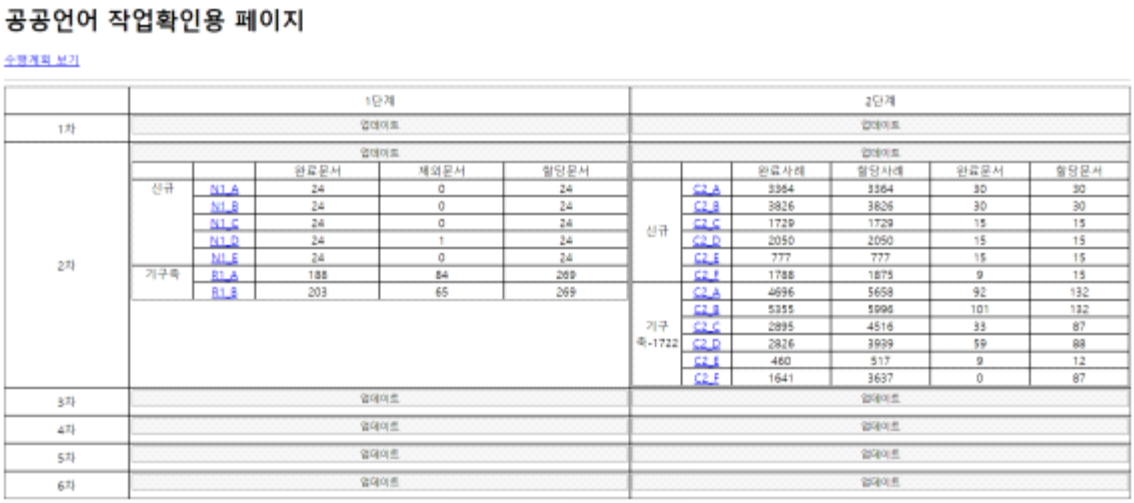
[그림 10] 검수 내용 최종 확인용 워크벤치 사용 예시

4.1.2. 작업 현황 파악 페이지 구축

신규 데이터 구축 및 검수 작업 현황은 워크벤치와 연동한 시스템을 통해 실시간 작업 내역을 확인할 수 있도록 하였다. 1단계로 표시된 왼쪽 면에서는 각 작업자에게 할당된 문서 수, 작업자들이 완료한 문서 수, 그리고 작업 과정에서 제외한 문서 수를 확인할 수 있다. 2단계로 표시된 오른쪽 면에서는 각 검수자에게 할당된 문서 수와 그 안에 포함된 작업 사례 수, 검수자들이 검수를 완료한 문서 수, 각 사례별로 입력된 오류 유형의 수를 확인할 수 있다. 이 수치를 기반으로 개별 작업자의 작업 현황을 파악하고, 해당 내역을 정리하여 일간 보고서에 반영함으로써 일 단위 전체 작업 상황을 참여 연구원들에게 공유하였다.

해당 페이지에서는 작업자 및 검수자의 워크벤치 계정으로 이동할 수 있는 링크가 활성화되어 있어, 작업관리 전담 연구원이 각 작업자의 작업 내역을 직접 확인할 수 있도록 하였

다.



공공언어 작업확인용 페이지

수행계획 보기

	1단계				2단계				
1차	업데이트				업데이트				
2차	업데이트				업데이트				
			완료문서	제외문서	할당문서			완료문서	할당문서
	신규	N1_A	24	0	24	신규	C2_A	3364	3364	30	30
		N1_B	24	0	24		C2_B	3826	3826	30	30
		N1_C	24	0	24		C2_C	1729	1729	15	15
		N1_D	24	1	24		C2_D	2050	2050	15	15
		N1_E	24	0	24		C2_E	777	777	15	15
	기구축	R1_A	188	84	269		C2_F	1788	1875	9	15
		R1_B	203	65	269	기구축-1722	C2_A	4696	5658	92	132
							C2_B	5355	5996	101	132
							C2_C	2895	4516	33	87
							C2_D	2826	3939	59	88
							C2_E	460	517	9	12
							C2_F	1641	3637	0	87
3차	업데이트				업데이트				
4차	업데이트				업데이트				
5차	업데이트				업데이트				
6차	업데이트				업데이트				

[그림 11] 작업 현황 확인용 페이지 화면

4.2. 일간 보고서 작성

본 사업팀은 작업 관리의 일환 중 하나로 월요일부터 토요일까지 매일 저녁 8시까지의 작업 결과를 기준으로 일일 작업 현황을 보고서로 작성하여 연구원들에게 회람하였다. 국립국어원에는 일간 보고서의 핵심적인 내용을 정리하여 주간보고서로 제출하였다. 일간 보고서는 양적·질적 측면으로 구분하여 작성하였다. 일간 보고서의 작업 수행률은 작업 수행 계획에 따라(표 3) 산출된 주간별 작업량을 주간 목표량으로 설정하여 판단하였다. 작업의 수행은 사업의 진행 과정에 따라 탄력적으로 배분하였다. 첫 주차에는 자료를 파악하고 지침을 수정하고 보완하는 단계로서 적은 수의 문서를 할당하고 작업에 익숙해질수록 문서의 수를 늘리는 방식이다.

<표 7> 주차별 작업 수행 계획(단위: 문건 수)

단계	작업	1주	2주	3주	4주	5주	6주	합계
1단계	신규 자료 구축	150	650	650	650	612	오류 수정	2,712
	기구축자료 검수 (2010-2016)	512	512	512				1,536
	기구축자료 검수 (2017-2022)		75	260	260	188	오류	783

2단계	신규 자료 검수		150	650	650	650	612	2,712
	기구축자료 2차 검수 (2010-2016)		512	512	512			1,536
	기구축자료 2차 검수 (2017-2022)			75	260	260	188	783
3단계	3차 검수 (전체 10%)					250	250	500
4단계	형식/품질 검증							5,031

4.2.1. 양적 측면 보고

양적 측면의 작업 보고는 크게 전체 작업 공정률과 작업자별 공정률로 구분하여 제시하였다. 전체 작업 공정률의 경우, '신규 자료 구축', '기구축 자료 검수' 등의 작업 유형별로 1단계와 2단계를 구분하여 작업 완료 건수를 보고하였다. 작업자별 공정률의 경우, 각각의 워크벤치 계정을 기준으로 하루 동안의 추가 작업 문건의 수를 기입하는 방식을 취했다. 다만 주간 작업량을 일 단위로 보고함에 따라, 매주 토요일을 주간 작업 완료일로 잡고 수요일부터 목표량 대비 미진한 작업 현황을 빨간색과 노란색으로 주의 상태를 표기함으로써 작업자가 진행 상태를 한눈에 파악할 수 있도록 하였다. 주의 상태가 표기된 요일별 작업 달성량 기준은 아래와 같다.

<표 8> 요일별 작업 수행 현황

수요일		목요일		금요일		토요일	
50% 미만	30% 미만	65% 미만	45% 미만	80% 미만	60% 미만	95% 미만	75% 미만
노랑	빨강	노랑	빨강	노랑	빨강	노랑	빨강

4.2.2. 질적 측면 보고

질적 측면의 작업 보고의 경우, 2단계, 3단계 검수와는 별개로 작업 관리 전담 연구원이 일간 작업 결과를 무작위로 살펴 이상이 있는 경우에 보고서에 기록하였다. 검수자들이 판단하기에 전반적인 작업 방식에 문제가 있는 경우 담당자에게 통지하도록 하였으며, 그 외에는 필요에 따라 작업자에게 직접 피드백을 제공하도록 하였다. 이를 기반으로 작업자용 지침을 정기적으로 업데이트하였으며, 필요에 따라 작업자 재교육을 실시하였다. 일간 보고서상 질적 측면 작업 보고 항목 아래에는 오류 유형의 토큰 수를 제시하였다. 11월 말까지는 작업자 워크벤치 계정별로 각각의 오류 유형별 토큰 수를 보고하였으나, 오류 유형의 상세 보고는 불필요하다고 판단하여 이후부터는 계정별 오류 유형 작업 수치만 기록하여 보고하였다.

4.2.3. 기타 작업 이슈 보고

그 밖에 일간 보고서에는 당일 발생한 작업 지침 업데이트 내역, 작업 및 워크벤치 관련 이슈, 추가 문건 배포사항, 국어원과의 협의 내용 등을 포함하였다. 더불어 일간 보고서와는 별개로 작업 현황에 대한 실시간 확인이 필요한 경우, 작업 관리 전담 연구원과 직접 소통할 수 있도록 하였다.

제5장 연구의 의의와 과제

5.1. 연구의 의의

이 사업은 공공언어의 질을 향상하고 이를 체계적으로 관리하기 위한 중요한 기초 작업이라 할 수 있다. 본 사업의 의의를 요약하면 다음과 같다. 먼저, 산재해 있던 공공언어 자료를 체계적으로 수집하고 통합하였다. 2010년부터 2022년까지의 공공언어 감수 자료를 망라하고 새로운 자료를 추가하여 병렬 말뭉치를 구축함으로써, 공공언어의 사용 패턴과 변화를 체계적으로 추적하고 분석할 수 있는 기반을 마련하였다.

둘째, 공공언어의 오류를 유형화하고 기존의 감수 지침을 개선하였다. 기존의 오류 지침을 통합하여 16개의 오류 유형을 정의하고 이를 기준으로 공공언어 감수 자료를 수정·보완함으로써, 공공언어의 품질을 높이고 일관된 표준의 형식을 제시하였다.

셋째, 구축된 자료를 바탕으로 인공지능 기술과의 접목을 꾀하였다. 수집된 데이터를 BERT 기반 모델을 이용한 인공지능 자동 윤문 도구 개발에 활용함으로써, 기술의 발전을 공공언어의 질 개선에 적극적으로 활용하였다. 이를 통해 궁극적으로 공공언어 감수 및 수정 과정을 자동화하고 효율화하는 데 기여하고자 하였다.

넷째, 향후 공공언어 감수 사업의 확장성과 연구의 기반을 제공하였다. 공공언어 감수와 관련된 추가 연구와 개선 작업을 위한 기초 자료를 제공함으로써, 이를 통해 공공언어의 품질을 지속적으로 개선하고, 더 나은 정책과 연구 방향을 모색할 수 있는 기회를 제공하고자 하였다.

마지막으로, 공공언어에 대한 대중의 접근성과 이해도를 향상하는 기반을 마련하였다. 공공언어의 명료성과 정확성을 높임으로써, 일반 언중이 정부 및 공공기관의 정보를 보다 쉽게 이해하고 접근할 수 있도록 지원하여 원활한 의사소통에 기여하고자 하였다.

5.2. 정책 제안

연구 결과를 바탕으로 공공언어 사용에 대한 정책적 제언과 미래 발전 방향 설정에 대해 다음과 같은 점들을 고려할 수 있다.

먼저, 지속적인 공공언어 감수 및 개선이 이루어져야 한다는 점이다. 공공언어의 질을 계속하여 향상하기 위해, 감수된 자료를 정기적으로 검토하고 업데이트하는 시스템을 마련해야 한다. 이를 통해 언어 변화와 사회적 요구에 신속하게 대응할 수 있는 기반을 마련할 수 있을 것이다.

다음으로, 수집한 말뭉치와 분석 결과를 활용하여, 공공언어 정책을 보다 데이터 중심으로 수립해야 한다. 이는 가시적인 데이터가 정책 결정 과정에서 실질적이고 구체적인 근거를 제공하기 때문이다.

셋째, 인공지능 기반의 윤문 도구 개발을 지속적으로 추진하여, 공공언어 감수와 수정 작업을 더욱 효율적으로 수행할 수 있도록 해야 한다. 이는 시간과 자원을 절약하는 동시에 공공언어의 일관성과 품질을 유지하는 데 기여할 것이다.

넷째, 공공기관 및 관련 분야 종사자들을 대상으로 정기적인 교육과 워크숍을 실시하여, 공공언어의 중요성과 올바른 사용 방법에 대한 인식을 높이는 한편, 대중으로부터의 피드백을 적극적으로 수집하고, 이를 정책에 반영할 수 있는 체계적인 메커니즘을 구축해야 한다. 이를 통해 공공언어가 꾸준히 개선되고 대중의 요구를 반영할 수 있을 것이다. 또한, 이를 바탕으로 공공언어 정책의 효과를 정기적으로 평가하고, 필요한 경우 정책을 수정하거나 개선하는 유연한 접근 방식을 취해야 한다.

마지막으로, 언어 규범에만 초점을 두지 않고 언어 다양성과 포용성을 강화해야 할 것이다. 다양한 언어 사용자들의 요구를 반영하여 공공언어 정책을 수립하고, 다문화 사회의 언어적 다양성을 존중해야 한다. 이를 위해 다양한 언어로 번역된 공공 자료의 제공과 다문화 배경의 사용자들을 위한 맞춤형 서비스를 강화해야 할 필요가 있다.

Abstract

This book is a result of the 2023 National Institute of the Korean Language (NIKL) project titled "Study on the Construction of Public Language Proofreading Data." This project aims to compile and integrate parallel corpora by incorporating materials from pre- and post-proofreading stages related to public language proofreading, covering the period from 2010 to 2022. Additionally, it involves the collection, refinement, and integration of new public language data. The report encompasses the outcomes of public language proofreading projects conducted by the NIKL from 2010 to the present, offering insights and policy recommendations for the field.

The first task of this project involves the collection of public language data, encompassing two major components. It includes the integration of results from previous projects related to public language proofreading conducted by the NIKL, as well as the collection of new public language data. Previous project outcomes, such as those from the "Comprehensive Data Construction for Public Language Proofreading Support (2017, led by Ilhwan Kim)," cover proofreading data from 2010 to 2016 and 2017 to 2022.

The second task of the project is to standardize and refine the collected data in terms of format and content. To achieve content uniformity, a practical review of existing guidelines for public language proofreading was conducted, identifying 16 error types for practical use. Based on these criteria, data were modified and enhanced. Existing pre- and post-proofreading data were reclassified after error review, and newly collected data underwent error analysis and classification. Format-wise, sentences containing errors, along with the preceding and following 10 words, were standardized into parallel data format. Practical improvements were made by modifying error types lacking utility or consistency and enhancing annotation guidelines for effective utilization.

The third task of the project involves transforming the constructed data into training data for developing artificial intelligence (AI) automatic proofreading tools and validating its potential as training data. To accomplish this, a simple BERT-based model was trained using the constructed data. Through this process, the model demonstrated F1 scores ranging from 0.96 to 0.98, confirming the potential utility of the data. Additionally, a web database-based working tool was established to effectively construct data of this format.

To ultimately build a public language automatic proofreading system, additional work and research are required to complement the outcomes of this study. Nevertheless, it is anticipated that the results of this study will be actively used in the field as public language proofreading materials and serve as foundational data for various studies, undergoing continuous modification and improvement.

[Key Words] Public Language, Official Documents, Accuracy, Error analysis, Parallel Corpus, Automatic proofreading, AI Training data

<기획·연구>
국립국어원 이승재 학예연구관
국립국어원 박 선 학예연구사

<연구 참여자>
연구 책임자 양명희
공동 연구원 송상헌, 정유남, 박미은, 김보현, 안예림, 홍승혜
연구 보조원 김혜원, 류다정, 이규민
보 조 원 이혜경, 정민경, 심재란, 박우빈, 조선화, 이혜영, 박예인, 조은비

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775,
전송 02-2669-9727
인쇄일: 2023년 12월 20일
발행일: 2023년 12월 20일
인 쇄: 유일문화사

※ 이 책은 국립국어원의 용역비로 수행한 '공공언어 감수 자료 구축 방안 연구' 사업의 결과물을 발간한 것입니다.

국립국어원
2023-01-41

공공언어 감수 자료 구축 방안 연구

국립국어원

문화체육관광부
국립국어원