

국립국어원
맞춤법
교정
말뭉치 연구 분석

2021년 맞춤법 교정 말뭉치 연구 분석

연구 책임자 | 남 길 임



국립국어원

국립국어원 2021-01-10

발 간 등 록 번 호
11-1371028-000862-01

2021년 맞춤법 교정 말뭉치 연구 분석

사업 책임자

남 길 임



국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2021년 맞춤법 교정 말뭉치 연구 분석’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2021년 07월 ~ 2021년 11월

2021년 11월 19일

사업 책임자: 남길임(경북대학교)

사업 기관 경북대학교 산학협력단
 주식회사 이르테크

사업 책임자 남길임

사업 참여자 광용진, 안미애, 김진웅, 송현주, 안의정, 황은하,
 심난희, 이후영, 최지선, 강신아, 강윤희, 이갑진,
 백미경, 강현아, 안진산, 황지윤, 고예린, 성민규,
 장희선, 이지혜, 김수지, 정나현, 전현진, 박정혁

<사업 수행자> 경북대학교 산학협력단 · 주식회사 이르테크

사업 책임자	남길임(경북대학교 국어국문학과 교수)
사업 참여자	곽용진((주)이르테크)
	안미애(경북대학교 국어국문학과 교수)
	김진웅(경북대학교 국어국문학과 교수)
	송현주(경북대학교 국어교육과 교수)
	안의정(연세대학교 문과대학 강사)
	황은하(배재대학교 국어국문·한국어교육학과 교수)
	심난희(배재대학교 주시경교양대학 교수)
	이후영((주)이르테크)
	최지선((주)이르테크)
	강신아(연세대 국어국문학과 박사 수료)
	강윤희(경북대 국어교육과 박사 과정)
	이갑진(경북대학교 국제교류처 강사)
	백미경(경북대학교 국제교류처 강사)
	강현아(경북대학교 국어국문학과 강사)
	안진산(경북대학교 국어국문학과 석사 과정)
	황지윤(경북대학교 국어국문학과 석사 과정)
	고예린(경북대학교 국어국문학과 석사 과정)
	성민규(경북대학교 국어국문학과 학부 과정)
	장희선(경북대학교 국어국문학과 학부 과정)
	이지혜(경북대학교 국어국문학과 학부 과정)
	김수지(배재대학교 한국어교육학과 석사 과정)
	정나현(배재대학교 한국어교육학과 석사 과정)
	전현진(배재대학교 한국어교육학과 석사 과정)
	박정혁(배재대학교 한국어교육학과 석사 과정)

<국문 초록>

2021년 맞춤법 교정 말뭉치 연구 분석

이 사업의 목적은 메신저 및 웹 말뭉치 300만 어절을 자동 형태소 분석, 기계 번역 등 한국어 처리 도구가 분석할 수 있는 수준으로 교정하고, 메신저 및 웹 언어의 특수성을 살린 교정 병렬 말뭉치의 구축 방안을 연구하고 구축하는 데 있다. 이를 위한 사업의 범위는 다음 세 가지이다.

첫째, 맞춤법 교정을 위한 지침을 수립한다.

둘째, 개인정보 및 부적절한 표현 등에 대한 처리 방안을 마련한다.

셋째, 인공지능 학습 데이터로서의 효용성을 고려한 맞춤법 교정 병렬 말뭉치를 구축한다.

각 하위 사업을 요약하여 제시하면 다음과 같다.

(1) 맞춤법 교정을 위한 지침의 수립

이 사업의 대상인 웹과 메신저 말뭉치는 사용자 생성 콘텐츠(User Generated Content, UGC)로 기호나 철자의 변형을 활용한 감정 표현, 구어체, 비규범적 표현, 오자와 탈자가 많은 외에, 혐오 및 차별적 표현을 포함하고 있다는 특성을 가지고 있다. 이 특성들은 문어·구어를 중심으로 학습된 기존의 형태소 분석기 등 NLP 도구의 적용을 어렵게 하며, 인공지능 데이터의 윤리적 활용에서도 문제를 야기한다. 본 사업에서는 메신저와 웹 말뭉치의 교정 목표 수준을 구어 전사 말뭉치 수준으로 상정하고, 이들의 언어적 특성을 연구, 분석함으로써, 맞춤법 교정 지침을 수립하였다. 교정 지침은 교정 유형별 지침으로 구성되며, 표준형과 비표준형의 판별은 <우리말샘>을 주요 기준으로 하되, 유형에 따라 별도의 지침을 수립하여 목록을 관리하였다.

(2) 개인정보 및 부적절한 표현 등에 대한 처리 방안 마련

이 사업에서는 민간에서 변환 및 호환이 용이한 공공재로서의 말뭉치 구축을 목표로 한다. 이를 위해 웹과 메신저 말뭉치에서 나타나는 개인정보 및 부적절한 표현을 파악하고, 비식별화하는 기준과 방안을 마련하였다. 개인정보의 경우, 기존 사업과의 유기성을 고려해 2019년에 추진한 국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업에서 제시한 범주에 의거해 비식별화 방안을 마련하여 적용하였다. 또 혐오, 차별 표현 및 부적절한 표현의 범주는 메신저와 웹의 특수성을 고려하여 혐오와 차별, 욕설, 성적 표현 등의 별도 범주를 분류함으로써 비식별화 지침을 수립하고 실제 말뭉치 구축에 적용하였다.

(3) 맞춤법 교정 병렬 말뭉치의 구축

이 작업은 자동 검사기 처리를 거친 후 수작업으로 맞춤법과 띄어쓰기를 전수 교정하는 방식으로 이루어진다. 또한 교정 작업의 효율화를 위해 교정 병렬 말뭉치 구축 도구인 Kronoth와 마이크로소프트사의 엑셀(Excel)을 사용하였으며, (주)이르테크의 말뭉치 검증 시스템을 활용해 분석 결과의 정확도를 확보하였다. 맞춤법 교정 말뭉치의 구축은 (1) 텍스트 전처리를 통한 맞춤법 교정용 말뭉치 변환 (2) 자동 맞춤법 교정 도구를 이용한 1차 자동 교정 (3) 수작업 전수 교정 (4) 개인정보와 부적절한 표현의 비식별화 (5) 세 차례의 품질 검수 (6) JSON 구조화 (7) 최종 형식 검수의 과정으로 구축되었다.

주요어: 맞춤법 교정 말뭉치, 메신저 말뭉치, 웹 말뭉치, 구어 말뭉치, 병렬 말뭉치, 맞춤법 검사기

차 례

제1장 서론

1. 사업의 목적	1
2. 사업의 범위	2
2.1. 맞춤법 교정 지침 수립	2
2.2. 개인정보 및 부적절한 표현 등에 대한 처리 방안 제시	2
2.3. 맞춤법 교정 병렬 말뭉치 구축	3

제2장 맞춤법 교정 말뭉치의 구축

1. 맞춤법 교정 말뭉치의 유형과 특성	7
1.1. 교정 대상 말뭉치의 유형과 특성	7
1.2. 맞춤법 교정 말뭉치의 구축 방향	7
2. 맞춤법 교정 말뭉치의 구축 단계	8
2.1. 맞춤법 교정용 말뭉치 변환	9
2.2. 자동 교정 및 후처리	10
2.3. 작업자 교육	12
2.4. 수작업 전수 교정	15
2.5. 개인정보와 부적절한 표현의 비식별화	20
2.6. 품질 검수	21
2.7. 최종 결과물 산출	23
2.8. 결과물 납품	26

차 례

제3장 교정 말뭉치 교정 지침 수립

1. 기본 지침과 지침 연구	29
1.1. 기본 지침	29
1.2. 지침 연구	30
2. 맞춤법 교정 말뭉치 교정 지침	32

제4장 결론

결론	77
----------	----

차 례

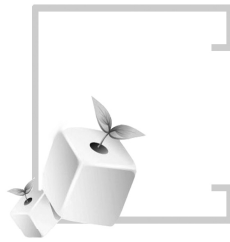
<표 차례>

<표 1> 교정 대상 말뭉치의 유형과 규모	7
<표 2> 맞춤법 자동 교정 도구 성능 비교	11
<표 3> 맞춤법 검사기로 과교정 또는 오교정된 예	12
<표 4> 맞춤법 교정 말뭉치의 JSON 형식 기본 구조	25
<표 5> 메신저 맞춤법 교정 말뭉치의 JSON 양식	26
<표 6> 교정 단계와 단계별 교정 내용	35
<표 7> 널리 통용되는 비규범 표기에 대한 규범형 목록	69

차 례

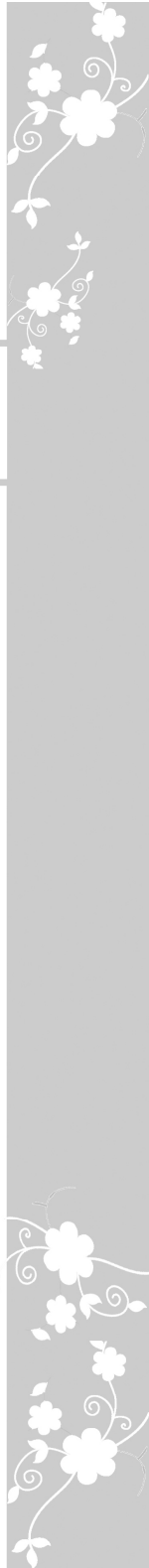
<그림 차례>

<그림 1> 맞춤법 교정 말뭉치 구축 단계	9
<그림 2> 메신저 원시 말뭉치(좌)와 웹 원시 말뭉치(우)의 JSON 구조	10
<그림 3> 구글 문서로 공유한 교정 지침 예시	13
<그림 4> 작업자 대상 1차 교육 자료 일부(좌)와 Kronoth 작업 도구 사용 방법 안내 영상 (우)	13
<그림 5> 작업자 대상 2차 교육 자료 일부	14
<그림 6> 검토자의 피드백 예시	14
<그림 7> 질의응답용 구글 시트 예시	15
<그림 8> Kronoth(v1.0)를 이용한 작업자의 작업 화면	16
<그림 9> Kronoth의 작업 현황 관리 화면	18
<그림 10> Kronoth의 작업자별 작업 현황	18
<그림 11> Kronoth의 전체 현황 통계 화면	19
<그림 12> 엑셀(Excel)을 이용한 작업자 화면	20
<그림 13> 엑셀(Excel)을 이용한 작업자 화면	20
<그림 14> 고빈도 통용 방언 검색 예시	59
<그림 15> 비교정 대상 구어체 방언형 수집 구글 시트	60
<그림 16> 방언:표준어 대응쌍 수집 구글 시트 예시	60
<그림 17> 웹 말뭉치 원 문장:교정 문장 예시	67
<그림 18> 외래어/외국어 관련 원 문장:교정 문장 예시	69
<그림 19> 기호, 문장부호 관련 원 문장:교정 문장 예시	71



제 1 장

서 론



1. 사업의 목적

이 사업은 인공지능 기술 개발 및 활용을 위해 기구축된 국립국어원의 메신저 및 웹 말뭉치를 대상으로, 맞춤법 교정을 위한 지침과 개인정보 및 부적절한 표현 등에 대한 처리 방안을 연구하고, 이를 반영하여 300만 어절 규모의 맞춤법 교정 병렬 말뭉치를 구축하는 것을 목적으로 한다. 이 사업의 구체적인 결과물은 다음과 같다.

- 맞춤법 교정을 위한 지침
- 개인정보 및 부적절한 표현 등에 대한 처리 방안
- 인공지능 학습 데이터로서의 효용성을 고려한 맞춤법 교정 병렬 말뭉치

최근 국립국어원에서는 ‘일상 대화 말뭉치, 구어 말뭉치, 형태 분석 말뭉치, 개체 분석 말뭉치’ 등 다양한 말뭉치를 구축하고 ‘모두의 말뭉치’를 통해 배포하고 있다. 이들 말뭉치는 21세기 세종계획을 통해 구축된 말뭉치의 한계를 극복하고 새로운 시대적 요구에 따라 구축된 대규모 언어 자원이다. 특히, 메신저 및 웹 말뭉치의 경우에 빅데이터로서 그 가치와 중요성을 인정받아 학계 및 산업계로부터 다양한 분석이 시도되어 왔으나 신어 및 미등재어, 띄어쓰기, 비표준형 등의 문제 등에 부딪혀 언어 자원으로서의 활용에 어려움을 겪어왔다.

메신저 및 웹 말뭉치에서는 생산 과정에 있어 전문가의 개입이 부재할 뿐만 아니라 텍스트 입력의 편의성을 높이기 위한 띄어쓰기의 무시, 오자와 탈자, 표음적 표기, 기호나 철자의 변형 등으로 전통적인 텍스트에서 나타나지 않는 다양한 현상들이 등장한다. 기존의 형태소 분석기나 구문 분석기와 같은 대부분의 자연언어처리 도구는 현대 표준어를 기준으로 개발되었기 때문에, 메신저 및 웹 말뭉치를 분석하는 경우에 오류율이 급격히 늘어나는 결과를 피하기 어렵다. 맞춤법 교정 병렬 말뭉치는 비표준형의 문제를 해결하기 위한 말뭉치로, 형태 분석과 구문 분석 등 상위 분석으로 가기 위한 필수 단계인 동시에, 공공재로서의 말뭉치, 초기 인공지능 학습 데이터로서의 의의를 가진다.

이 사업을 통하여 기구축된 메신저 및 웹 원시 말뭉치를 대상으로 맞춤법 교정 병렬 말뭉치를 구축하는 데 필요한 지침을 수립하고 표준화된 말뭉치를 제공함으로써, 국내 인공지능 산업에 활용할 수 있는 기반을 마련하고, 인공지능 기술의 국가 경쟁력에 기여할 수 있다. 또한 이 사업의 결과물이 메신저 및 웹이라는 새로운 언어 환경에 적합한 다양한 인공지능 기술 개발의 기반 자료로 활용되어 대국민 서비스 강화에 이바지하리라 기대한다.

2. 사업의 범위

이 사업은 크게 세 부분으로 구성되어 있다. 첫째, 언어학적 정밀성과 공학적 활용도를 고려한 맞춤법 교정 지침을 수립한다. 둘째, 개인정보 및 부적절한 표현 등에 대한 처리 방안을 마련한다. 셋째, 메신저 대화 원시 말뭉치(200만 어절)와 웹 원시 말뭉치(100만 어절)를 대상으로 인공지능 학습 데이터로서의 효용성을 고려한 맞춤법 교정 병렬 말뭉치를 구축한다.

2.1. 맞춤법 교정 지침 수립

이 사업의 목적은 자동 형태소 분석, 기계 번역 등 한국어 처리 도구가 메신저 및 웹 원시 말뭉치를 분석할 수 있는 수준으로 교정하되, 언어의 특수성을 살린 교정 병렬 말뭉치를 구축하는 데 있다. 사업의 목적이 이상적인 교육용 규범 말뭉치를 구축하는 데 있기 보다는 일종의 이개어(원 문장과 교정 문장) 병렬 말뭉치, 기계를 위한 학습 데이터로서의 교정 말뭉치를 구축하는 데 있는 만큼, 교정의 수준은 오타자, 비표준형, 띄어쓰기 등을 구어 전사 말뭉치 수준을 목표로 한다. 또한, 맞춤법 검사기, 자동 형태소 분석기 등 기계가 처리할 수 있는 수준에 한해 엄격한 규범을 추구하기보다는 지침에서 명시된 허용 규정을 적용한다. 마지막으로 개인정보를 비롯하여, 욕설, 혐오 및 차별 표현 등 부적절한 표현을 비식별화하기 위한 세부 지침을 제공한다. 이 사업에서 수립한 맞춤법 교정 지침의 주요 내용은 다음과 같다.

- 한글 맞춤법에 따른 띄어쓰기, 오타자 등 교정 방안
- <우리말샘> 미등재어(외래어, 신어 등) 및 비표준어 처리 방안
- 온라인 환경에서 나타나는 특수 표현 등의 처리 방안
- 개인정보 및 부적절한 표현(욕설, 혐오 표현 등)의 비식별화 방안

2.2. 개인정보 및 부적절한 표현 등에 대한 처리 방안 제시

이 사업에서는 민간에서 변환 및 호환이 용이한 공공재로서의 말뭉치 구축을 목표로 한다. 이를 위해 웹과 메신저 말뭉치에서 나타나는 개인정보 및 부적절한 표현을 파악하고, 비식별화하는 기준과 방안을 마련하였다. 개인정보의 경우, 기존 사업과의 유기성을 고려해 2019년에 추진한 국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업에서 제시한

범주에 의거해 비식별화 방안을 마련하여 적용하였다. 또 혐오, 차별 표현 및 부적절한 표현의 범주는 메신저와 웹의 특수성을 고려하여 혐오와 차별, 욕설, 성적 표현 등의 별도 범주를 분류함으로써 비식별화 지침을 수립하고 실제 말뭉치 구축에 적용하였다. 이 사업에서 수립한 비식별화 지침의 주요 내용은 다음과 같다.

- 이름, 온라인 아이디, 이메일, 각종 번호, 장소, 출신 및 소속 등에 대한 철저한 비식별화
- 웹 말뭉치의 경우 민감한 개인정보인 ‘고유 식별 번호(주민 등록 번호, 사번, 학번 등), 전화 번호, 금융 번호, 주소’에 한해 비식별화
- 욕설과 성적인 표현은 비식별화하고, 혐오 및 차별 표현은 ‘형태’가 아닌 ‘맥락’을 기반으로하여 비식별화 여부를 결정
- 비속한 표현 가운데에도 강조의 의미이거나 욕설로 분류하기 어려운 경우에 비식별화 대상에서 제외

2.3. 맞춤법 교정 병렬 말뭉치 구축

이 작업은 자동 검사기 처리를 거친 후 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식으로 이루어진다. 수작업 시에는 교정 작업의 효율화를 위해 교정 병렬 말뭉치 구축 도구인 Kronoth와 마이크로소프트사의 엑셀(Excel) 프로그램을 사용하였다. 또한 (주)이르테크의 말뭉치 검증 시스템을 활용해 자동 맞춤법 분석 결과의 정확도를 확보하였다. 맞춤법 교정 말뭉치는 (1) 텍스트 전처리를 통한 맞춤법 교정용 말뭉치 변환 (2) 자동 맞춤법 교정 도구를 이용한 1차 자동 교정 (3) 수작업 전수 교정 (4) 개인정보와 부적절한 표현의 비식별화 (5) 세 차례의 품질 검수 (6) JSON 구조화 (7) 최종 형식 검수의 과정으로 구축되었다.

- 사업 목표: ‘원문-교정문’ 형식의 병렬 말뭉치로 정제 및 가공

- 대상 자료:

- 2019년 국립국어원 메신저 대화 원시 말뭉치(200만 어절(100만 어절은 2020년 국립국어원에서 구축한 어휘 의미 분석 말뭉치))
- 2019년 국립국어원 웹 원시 말뭉치(100만 어절)

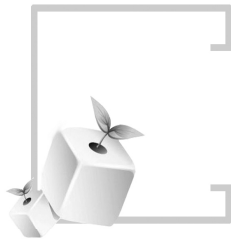
- 사업 수행 방식과 단계: 3단계 검수와 구조 형식 검수를 통한 품질 관리

- 1차: 샘플(10%) 검수

2차: 동일 어절의 불일치 교정 결과 추출 및 검수

3차: 대규모 말뭉치와의 어절 비교 및 불일치 교정 결과 추출

최종: 구조 형식 검수



제 2 장

맞춤법 교정 말뭉치 의 구축



1. 맞춤법 교정 말뭉치의 유형과 특성

1.1. 교정 대상 말뭉치의 유형과 특성

이 사업의 교정 대상 말뭉치는 메신저 대화 원시 말뭉치와 웹 원시 말뭉치로 총 300만 어절이다. 각 말뭉치별 규모는 아래 표와 같다.

교정 대상 말뭉치 유형	규모
국립국어원 메신저 대화 원시 말뭉치	200만 어절
국립국어원 웹 원시 말뭉치	100만 어절
총계	300만 어절

<표 1> 교정 대상 말뭉치의 유형과 규모

위의 두 말뭉치는 국립국어원에서 2019년에 구축한 말뭉치이다. 먼저 메신저 원시 말뭉치는 메신저를 통해 수집한 대화문으로 구어의 형식을 띤 문어 텍스트 원시 말뭉치이며, 이 사업의 대상은 이 원시 말뭉치에서 별도로 추출한 200만 어절 정도의 말뭉치이다. 메신저 대화 말뭉치는 기존의 대화 자료를 제출하거나 별도의 대화 수집창에서 실시간으로 나눈 대화를 제공하는 방식으로 수집된 대화문이다. 대화의 유형은 2인 대화와 다자 대화로 구분된다. 따라서 이 말뭉치는 텍스트로 된 대화문이기에 문어지만 구어의 특징도 가진다.

웹 원시 말뭉치 또한 메신저 말뭉치와 유사한 특징을 공유한다. 이 사업의 교정 대상인 웹 원시 말뭉치는 누리 소통망(SNS)에서 수집한 내용으로 구성되었으며 이 사업의 대상은 이 원시 말뭉치에서 별도로 추출한 100만 어절 정도의 말뭉치이다. 다양한 웹 도메인에서 생산되므로 웹 도메인의 특성에 따라 문어와 구어의 특성을 동시에 가질 수도 있다. 이 두 말뭉치 모두 수집 과정에서 전문가의 중재가 없기에 사용자의 어문규범이나 표현 규약이 사전에 정돈되지 않고 예외적인 표기 등이 다수 나타난다. 이러한 점은 기존의 형태소 분석기나 맞춤법 교정기 등과 같은 NLP 도구의 자동 처리 정확도를 떨어뜨린다.

그럼에도 불구하고 구어와 문어의 특성을 공유하는 메신저와 웹 원시 말뭉치는 향후의 자연어 처리 연구에서 핵심적인 역할을 기대할 수 있다. 그러나 이 두 말뭉치의 활용도를 높이기 위해서는 형태소 분석기와 같은 NLP 도구의 적용이 가능한 수준으로의 교정이 필요하다.

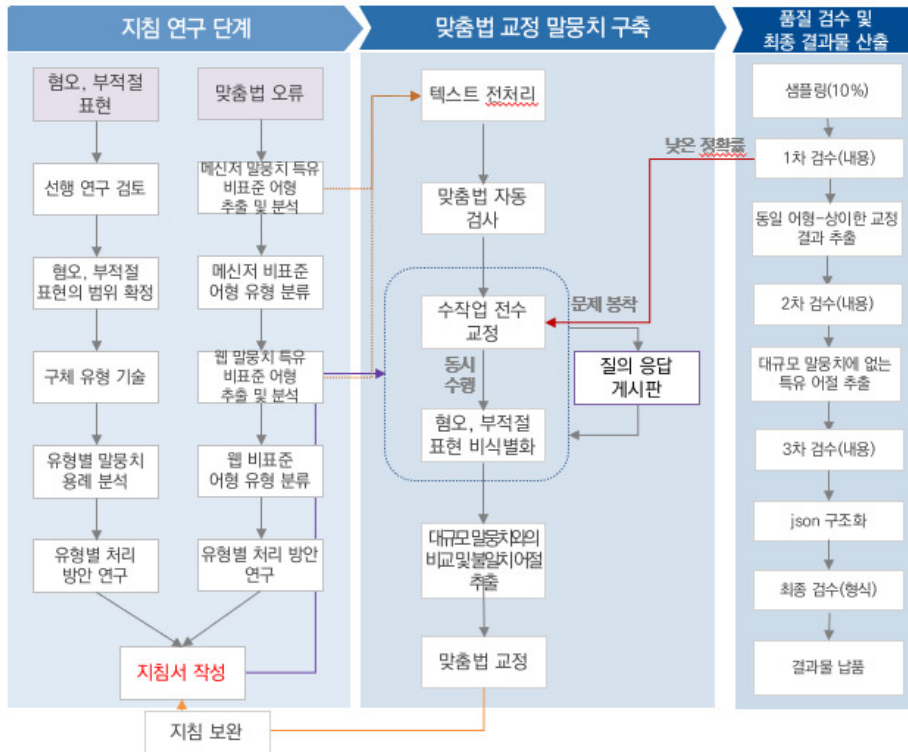
1.2. 맞춤법 교정 말뭉치의 구축 방향

1.1에서 언급한 바와 같이 메신저와 웹 원시 말뭉치 300만 어절은 문어와 구어의 특징을 동시에 가지며, 말뭉치의 장르 특징상의 비규범적 형태가 다수 나타나 자연어 처리 도구로 일정 수준의 교정을 기대하기 어렵다. 이에 이 사업은 이 메신저와 웹 원시 말뭉치의 비규범적 형태를 규범적 형태로 전환하여, 원문-교정문 형식을 갖춘 맞춤법 교정 병렬 말뭉치를 구축하는 데 사업의 기본 목적을 두고 있다.

맞춤법 교정 말뭉치는 말뭉치의 형태 분석이나 구문 분석 등의 심층적 분석을 시도하기 위한 필수 단계이다. 기존 출판물 기반의 문어 원시 말뭉치나 인터뷰 기반의 구어 원시 말뭉치는 예측 가능한 비규범적 변이형이 주로 나타났기에 맞춤법 교정기만으로도 어느 정도 수준의 어문규범 교정을 기대할 수 있다. 그러나 메신저 말뭉치나 웹 말뭉치는 예측 범위 외의 비규범 변이형이 다수 존재하므로 맞춤법 교정기만으로 말뭉치의 품질을 제고하기 어렵다. 이에 이 사업은 ‘표준화를 위한 초기 학습 데이터로서의 맞춤법 교정 병렬 말뭉치’를 메신저와 웹 말뭉치 교정의 기본 구축 방향으로 삼는다. 즉, 완벽한 수준의 맞춤법 교정이 아니라 한국어 형태소 분석과 언어학·공학에서 실용적인 목적으로 활용이 가능한 수준의 맞춤법 교정이다. 언어학적 정밀성과 공학적 활용도를 고려한, 맞춤법 교정 말뭉치의 구축이 이 사업이 지향하는 바이다. 다음으로 교정한 말뭉치의 가치와 활용도 수준은 개인정보와 부적절한 표현이 적절하게 비식별화되어 공공재로 적절한 수준이다.

2. 맞춤법 교정 말뭉치의 구축 단계

맞춤법 교정 말뭉치 사업은 첫째, 언어학적 정밀성과 공학적 활용도를 고려한 맞춤법 교정 지침의 수립, 둘째, 개인정보 및 부적절한 표현 등에 대한 처리 방안 마련, 셋째, 실제 맞춤법 교정 병렬 말뭉치의 구축의 세 가지 목적을 가지고 있다. 이 목적을 실현하기 위해 본 사업은 아래와 같은 단계로 시행되었다.



<그림 1> 맞춤법 교정 말뭉치 구축 단계

1장에서 제시한 맞춤법 교정 말뭉치의 구축 방향에 따라 본 사업의 맞춤법 교정 말뭉치는 위의 세 단계를 거쳐 구축되었다. 2장에서는 위의 <그림 1>의 구축 단계별 세부 공정 중 맞춤법 교정 말뭉치 구축 단계와 품질 검수 및 최종 결과물 산출 단계에 대해 설명하고자 한다.

2.1. 맞춤법 교정용 말뭉치 변환

이 단계에서는 교정 말뭉치의 구축을 위한 텍스트 전처리 작업을 시행하였다. 텍스트 전처리 작업은 JSON 구조 분석, 노이즈 제거, 작업 도구 탑재의 순서로 진행되었다. 이를 위해 국립국어원에서 제공한 메신저와 웹 말뭉치의 JSON 구조를 분석하였다. 원시 JSON 구조는 다음과 같다.

<pre> "utterance": [{ "id": "MDRW1900000012.1.1.1", "form": "얼마~", "original_form": "얼마~", "speaker_id": "1", "time": "20191128 16:55" }, { "id": "MDRW1900000012.1.1.2", "form": "name3했나?", "original_form": "&name3&했나?", "speaker_id": "2", "time": "20191128 16:55" }, { "id": "MDRW1900000012.1.1.3", "form": "네 name3 자고일어나서 기분좋아요", "original_form": "네 &name3& 자고일어나서 기분좋아요", "speaker_id": "1", "time": "20191128 16:56" }], </pre>	<pre> "paragraph": [{ "id": "ERRW1905000308.26.1", "form": "#이마트 트레이더스 가서 장만해온..", "original_form": "#이마트 트레이더스 가서 장만해온.." }, { "id": "ERRW1905000308.26.2", "form": ".", "original_form": "." }, { "id": "ERRW1905000308.26.3", "form": "#비포착로 #유신군 넘나 잘먹는다 우리집 꼬맹들..", "original_form": "#비포착로 #유신군 넘나 잘먹는다 우리집 꼬맹들.." }, { "id": "ERRW1905000308.26.4", "form": "세상 꼭 떨어 날고 아직 어린 둘째도..", "original_form": "세상 꼭 떨어 날고 아직 어린 둘째도.." }], </pre>
---	---

<그림 2> 메신저 원시 말뭉치(좌)와 웹 원시 말뭉치(우)의 JSON 구조

다음으로 메신저와 웹 원시 말뭉치에 나타나는 노이즈 유형을 분석하였다. 원시 말뭉치에 나타나는 텍스트 노이즈의 주요 유형은 중복 파일과 비정상적인 공백 등으로 분석되었다. 노이즈 유형에 대한 분석을 바탕으로, 국어원과 협의하여 중복 파일을 제거하고, 비정상적인 공백을 제거하는 등의 노이즈 제거 작업을 진행하였다.

2.2. 자동 교정 및 후처리

(1) 맞춤법 자동 검사기 비교 및 선정

본 사업팀은 단기간에 300만 어절의 말뭉치에 대한 맞춤법 교정을 완수하기 위해 수작업 전수 교정 작업 이전에 1차 자동 교정을 수행하였다. 자동 교정의 정확률이 높을수록 수작업 교정의 부담을 줄일 수 있으므로 맞춤법 교정 도구를 메신저와 웹 텍스트의 교정에 최적화하는 작업이 필요하다.

이를 위해 우선, 메신저와 웹 말뭉치에서 추출한 문장으로 공개된 맞춤법 자동 교정 도구를 비교 분석하여 정확도가 높은 도구를 선별한다. 다음은 세 가지의 서로 다른 맞춤법 교정기의 자동 교정 결과이다.

	원문	자동 교정 결과		
		부산대 맞춤법 교정기	도구 B	도구 C
1	일하면서 자격증 따놔~~~ㅋㅋ	일하면서 자격증 따놔~~~ <u>ㅋㅋ</u>	일하면서 자격증 따놔~~~ <u>ㅋㅋ</u>	일하면서 자격증 따놔~~~ <u>ㅋㅋ</u>
2	응응. 힘들어서 안 되겠더라고. ㅋㅋ	응응. 힘들어서 안 되겠더라고. ㅋㅋ	응응. 힘들어 <u>서</u> <u>안</u> 되겠더라고. ㅋㅋ	응응. 힘들어서 안 되겠더라고. ㅋㅋ
3	ㅋ나도 공인중개사 맛만 보고 접음. ㅋㅋ	ㅋ. 나도 공인중개사 맛만 보고 접음. ㅋㅋ	<u>ㅋ나도공인중개사맛만보고접음.</u> <u>ㅋㅋ</u>	<u>ㅋ나</u> <u>도공인</u> <u>중개사</u> 맛만 보 고 접음. <u>ㅋㅋ</u>
4	근데. 공부이제 하기 싫으네요. 언니도 진급 공부했조 —	근데. 공부 이제 하기 싫네요. 언니도 진급 공부했조. —	근데. <u>공부이제</u> <u>하기 싫으네요.</u> <u>언니도진급공부</u> <u>했조—</u>	근데. 공부 이제 하기 싫네요. 언 <u>니</u> <u>도</u> 진급 공부 했조—
5	&name3&이도내 년이문초당인다. 적응 잘하려는지 거정되죽겠다.	&name 3&이도 내 년이면 초등학교생인 데. 적응 잘하려 <u>는</u> <u>자기 전</u> <u>되</u> 죽 겠다.	&name 3&이 <u>도</u> <u>내</u> 년이 <u>문초</u> 당인 <u>디.</u> <u>적응</u> <u>잘</u> 하 <u>려</u> 는 <u>지</u> <u>거정</u> <u>되</u> 죽 겠다.	&name3&이 <u>도내</u> 년이 <u>문초</u> 당인다. <u>적응</u> <u>잘</u> 하 <u>려</u> 는 <u>지</u> <u>거정</u> <u>되</u> 죽 겠다.

<표 2> 맞춤법 자동 교정 도구 성능 비교

<표 2>에서 밑줄로 표시한 부분이 자동 교정 결과에도 고쳐지지 않은 띄어쓰기, 표기 등의 한국어 어문규범에 어긋나는 부분이다. 위와 같은 비교 작업의 결과, 부산대 맞춤법 교정기가 교정 후 교정 정확률이 가장 높다고 판단하여 이 교정기를 본 사업의 자동 교정 단계에 활용하였다.

(2) 자동 교정

이 단계에서는 2.1에서 노이즈를 제거하고 형식을 변환한 메신저 말뭉치 200만 어절과 웹 말뭉치 100만 어절을 부산대 맞춤법 교정기를 활용하여 자동 교정을 진행하였다.

다만, 다른 맞춤법 교정 도구들과 마찬가지로 부산대 맞춤법 교정기 또한 일반 텍스트의 교정을 위해 개발된 것으로, 메신저 텍스트와 웹 텍스트의 교정 결과에 과교정 또는 오교정된 유형의 예가 발견되었는데, 일부를 보이면 다음과 같다.

원어절	과교정/오교정 어절
가성비	구성비
π π	bb 또는 Bb
스카이스캐너	스카의 스캐너
ㄴ ㄷ ㄴ ㄷ	sese
아니예요	아니예요
겁나	많이
달라고	달이라고
똥똥이	똥똥이
돼요	돼 요
화나다	빠치다
오오	오와

<표 3> 맞춤법 검사기로 과교정 또는 오교정된 예

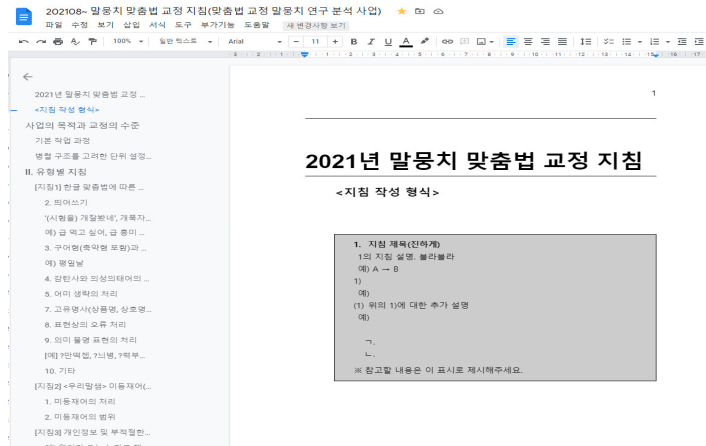
따라서 자동 교정 결과에 대한 샘플링 검수를 통해 주요 오교정과 과교정 양상을 분석하여 일괄 교정이 가능한 오류 유형을 정리하였다. 오교정에 대해서는 일괄 교정을 하였다. 과교정에 대해서는 과도 교정과 과소 교정으로 나누어 전자는 원시 형태로 복원하고, 후자에 대해서는 추가 교정하는 작업을 수행하였다. 이와 같은 텍스트 후처리는 다음 단계의 수작업 전수 검수의 부담을 줄이고 교정의 일관성을 높이는 데 효과적이었다.

2.3. 작업자 교육

작업자 교육은 크게 (1) 사전 지침 교육, (2) 상위 검수자에 의한 샘플링 검수 및 피드백을 통한 재교육, (3) 구글 스프레드 시트를 이용한 즉각적인 질의 및 응답 공유를 통한 수시 교육의 방식으로 진행되었다.

(1) 사전 지침 교육

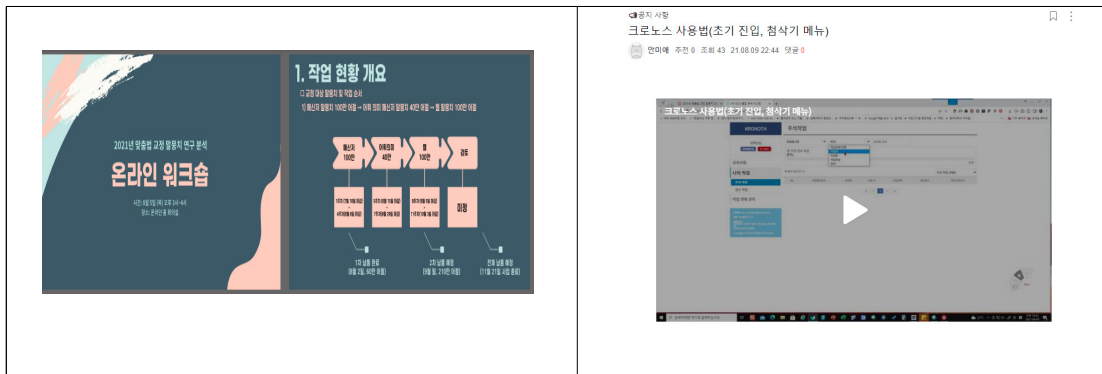
작업과 관련된 사항은 웹 커뮤니티와 메신저를 통해 수시로 소통할 수 있도록 하였으며 교정 지침은 아래와 같이 구글 문서로 공유하여 수정 사항이 발생할 경우, 즉각적으로 대응할 수 있도록 하였다.



<그림 3> 구글 문서로 공유한 교정 지침 예시

또한 작업자의 지침 및 작업 도구 사용 방법 숙지를 위해 작업자를 대상으로 한 작업 교육이 2회 시행되었으며 작업 도구 사용 방법은 동영상으로도 제공하였다.

-1차 교육: 작업 현황 개요, 작업자별 분담 내역, 지침 및 작업 도구 사용 방법 안내



<그림 4> 작업자 대상 1차 교육 자료 일부(좌)와 Kronoth 작업 도구 사용 방법 안내 영상(우)

1. 작업 현황 개요

□ 공정 대상 장비 및 작업 순서

1) 용선저 발생지 100만 어를 -여유 여하 용선저 발생지 40만 어를 -필 발생지 100만 어를

용선저 100만	여유여하 40만	필 100만	마정
15주 (7월 19일 ~ 8월 2일)	55주 (7월 19일 ~ 8월 2일)	95주 (7월 19일 ~ 8월 2일)	135주 (7월 19일 ~ 8월 2일)
4500여척 (용선저 100만)	4500여척 (용선저 40만)	4500여척 (용선저 100만)	4500여척 (용선저 100만)
1차 15일 범위 (8월 2일, 80만 어를)	2차 15일 범위 (9월 1일, 140만 어를)	3차 15일 범위 (11월 21일 사업 종료)	

2. 작업자별 진척율

□ 월 발생지 작업 세부 계획

1. 일일

- 9주 차: 09. 06. ~ 09. 12.
- 10주 차: 09. 13. ~ 09. 19.
- 추석 연휴: 09. 20. ~ 09. 26.
- 11주 차: 09. 27. ~ 10. 03.
- 12주 차: 10. 04. ~ 10. 10.
- 13주 차: 10. 11. ~ 10. 17.

2. 분량

- 주별 전체 작업량: 약 200,000 어를
- 주별 개인 작업량: 약 17,500 어를

작업자	1주 (9월 19일 ~ 9월 25일)	2주 (9월 26일 ~ 10월 2일)	3주 (10월 3일 ~ 10월 9일)	4주 (10월 10일 ~ 10월 16일)	5주 (10월 17일 ~ 10월 23일)	6주 (10월 24일 ~ 10월 30일)	7주 (10월 31일 ~ 11월 6일)	8주 (11월 7일 ~ 11월 13일)	9주 (11월 14일 ~ 11월 20일)	10주 (11월 21일 ~ 11월 27일)	11주 (11월 28일 ~ 12월 4일)	12주 (12월 5일 ~ 12월 11일)	13주 (12월 12일 ~ 12월 18일)	14주 (12월 19일 ~ 12월 25일)	15주 (12월 26일 ~ 1월 1일)	16주 (1월 2일 ~ 1월 8일)	17주 (1월 9일 ~ 1월 15일)	18주 (1월 16일 ~ 1월 22일)	19주 (1월 23일 ~ 1월 29일)	20주 (1월 30일 ~ 2월 5일)	21주 (2월 6일 ~ 2월 12일)	22주 (2월 13일 ~ 2월 19일)	23주 (2월 20일 ~ 2월 26일)	24주 (2월 27일 ~ 3월 3일)	25주 (3월 4일 ~ 3월 10일)	26주 (3월 11일 ~ 3월 17일)	27주 (3월 18일 ~ 3월 24일)	28주 (3월 25일 ~ 3월 31일)	29주 (4월 1일 ~ 4월 7일)	30주 (4월 8일 ~ 4월 14일)	31주 (4월 15일 ~ 4월 21일)	32주 (4월 22일 ~ 4월 28일)	33주 (4월 29일 ~ 5월 5일)	34주 (5월 6일 ~ 5월 12일)	35주 (5월 13일 ~ 5월 19일)	36주 (5월 20일 ~ 5월 26일)	37주 (5월 27일 ~ 6월 2일)	38주 (6월 3일 ~ 6월 9일)	39주 (6월 10일 ~ 6월 16일)	40주 (6월 17일 ~ 6월 23일)	41주 (6월 24일 ~ 6월 30일)	42주 (7월 1일 ~ 7월 7일)	43주 (7월 8일 ~ 7월 14일)	44주 (7월 15일 ~ 7월 21일)	45주 (7월 22일 ~ 7월 28일)	46주 (7월 29일 ~ 8월 4일)	47주 (8월 5일 ~ 8월 11일)	48주 (8월 12일 ~ 8월 18일)	49주 (8월 19일 ~ 8월 25일)	50주 (8월 26일 ~ 9월 1일)	51주 (9월 2일 ~ 9월 8일)	52주 (9월 9일 ~ 9월 15일)	53주 (9월 16일 ~ 9월 22일)	54주 (9월 23일 ~ 9월 29일)	55주 (9월 30일 ~ 10월 6일)	56주 (10월 7일 ~ 10월 13일)	57주 (10월 14일 ~ 10월 20일)	58주 (10월 21일 ~ 10월 27일)	59주 (10월 28일 ~ 11월 3일)	60주 (11월 4일 ~ 11월 10일)	61주 (11월 11일 ~ 11월 17일)	62주 (11월 18일 ~ 11월 24일)	63주 (11월 25일 ~ 12월 1일)	64주 (12월 2일 ~ 12월 8일)	65주 (12월 9일 ~ 12월 15일)	66주 (12월 16일 ~ 12월 22일)	67주 (12월 23일 ~ 12월 29일)	68주 (12월 30일 ~ 1월 5일)	69주 (1월 6일 ~ 1월 12일)	70주 (1월 13일 ~ 1월 19일)	71주 (1월 20일 ~ 1월 26일)	72주 (1월 27일 ~ 2월 2일)	73주 (2월 3일 ~ 2월 9일)	74주 (2월 10일 ~ 2월 16일)	75주 (2월 17일 ~ 2월 23일)	76주 (2월 24일 ~ 2월 30일)	77주 (3월 1일 ~ 3월 7일)	78주 (3월 8일 ~ 3월 14일)	79주 (3월 15일 ~ 3월 21일)	80주 (3월 22일 ~ 3월 28일)	81주 (3월 29일 ~ 4월 4일)	82주 (4월 5일 ~ 4월 11일)	83주 (4월 12일 ~ 4월 18일)	84주 (4월 19일 ~ 4월 25일)	85주 (4월 26일 ~ 5월 2일)	86주 (5월 3일 ~ 5월 9일)	87주 (5월 10일 ~ 5월 16일)	88주 (5월 17일 ~ 5월 23일)	89주 (5월 24일 ~ 5월 30일)	90주 (5월 31일 ~ 6월 6일)	91주 (6월 7일 ~ 6월 13일)	92주 (6월 14일 ~ 6월 20일)	93주 (6월 21일 ~ 6월 27일)	94주 (6월 28일 ~ 7월 4일)	95주 (7월 5일 ~ 7월 11일)	96주 (7월 12일 ~ 7월 18일)	97주 (7월 19일 ~ 7월 25일)	98주 (7월 26일 ~ 8월 1일)	99주 (8월 2일 ~ 8월 8일)	100주 (8월 9일 ~ 8월 15일)	101주 (8월 16일 ~ 8월 22일)	102주 (8월 23일 ~ 8월 29일)	103주 (8월 30일 ~ 9월 5일)	104주 (9월 6일 ~ 9월 12일)	105주 (9월 13일 ~ 9월 19일)	106주 (9월 20일 ~ 9월 26일)	107주 (9월 27일 ~ 10월 3일)	108주 (10월 4일 ~ 10월 10일)	109주 (10월 11일 ~ 10월 17일)	110주 (10월 18일 ~ 10월 24일)	111주 (10월 25일 ~ 10월 31일)	112주 (11월 1일 ~ 11월 7일)	113주 (11월 8일 ~ 11월 14일)	114주 (11월 15일 ~ 11월 21일)	115주 (11월 22일 ~ 11월 28일)	116주 (11월 29일 ~ 12월 5일)	117주 (12월 6일 ~ 12월 12일)	118주 (12월 13일 ~ 12월 19일)	119주 (12월 20일 ~ 12월 26일)	120주 (12월 27일 ~ 1월 2일)	121주 (1월 3일 ~ 1월 9일)	122주 (1월 10일 ~ 1월 16일)	123주 (1월 17일 ~ 1월 23일)	124주 (1월 24일 ~ 1월 30일)	125주 (1월 31일 ~ 2월 6일)	126주 (2월 7일 ~ 2월 13일)	127주 (2월 14일 ~ 2월 20일)	128주 (2월 21일 ~ 2월 27일)	129주 (2월 28일 ~ 3월 4일)	130주 (3월 5일 ~ 3월 11일)	131주 (3월 12일 ~ 3월 18일)	132주 (3월 19일 ~ 3월 25일)	133주 (3월 26일 ~ 4월 1일)	134주 (4월 2일 ~ 4월 8일)	135주 (4월 9일 ~ 4월 15일)
-----	----------------------	----------------------	----------------------	------------------------	------------------------	------------------------	-----------------------	-----------------------	------------------------	-------------------------	------------------------	------------------------	-------------------------	-------------------------	-----------------------	---------------------	----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	-----------------------	---------------------	----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	----------------------	---------------------	-----------------------	-----------------------	-----------------------	---------------------	----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	----------------------	---------------------	----------------------	-----------------------	-----------------------	-----------------------	------------------------	-------------------------	-------------------------	------------------------	------------------------	-------------------------	-------------------------	------------------------	-----------------------	------------------------	-------------------------	-------------------------	-----------------------	----------------------	-----------------------	-----------------------	----------------------	---------------------	-----------------------	-----------------------	-----------------------	---------------------	----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	----------------------	---------------------	-----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	----------------------	----------------------	-----------------------	-----------------------	----------------------	---------------------	-----------------------	------------------------	------------------------	-----------------------	-----------------------	------------------------	------------------------	------------------------	-------------------------	--------------------------	--------------------------	--------------------------	------------------------	-------------------------	--------------------------	--------------------------	-------------------------	-------------------------	--------------------------	--------------------------	------------------------	----------------------	------------------------	------------------------	------------------------	-----------------------	-----------------------	------------------------	------------------------	-----------------------	-----------------------	------------------------	------------------------	-----------------------	----------------------	-----------------------

(2) 상위 검수자에 의한 샘플링 검수 및 피드백을 통한 재교육

[illegible]

(3) 구글 스프레드 시트를 이용한 수시 질의 및 응답

14

다. 이에 본 사업팀에서는 지침을 벗어나는 변이형에 대해 즉각적으로 대처하기 위해 구글 스프레드 시트를 활용한 질의응답을 진행하였다. 구글 스프레드 방식의 질의응답은 실시간으로 질문과 답변을 공유할 수 있으며 기록이 남으므로 추후 같은 유형에 대해 검토자와 작업자가 대응하는 데 용이하다는 장점을 가진다. 아래는 질의응답을 위해 공유한 본 사업팀의 구글 스프레드 시트 예시이다.

[맞춤법 교정사업] 질의응답 스프레드시트

일련번호	일시	작성자	질문 내용	답변 내용	문의 여부	교정 여부	답변 대안 설명	답변자
1	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
2	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
3	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
4	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
5	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
6	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
7	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
8	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
9	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
10	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
11	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
12	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
13	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
14	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
15	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
16	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
17	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
18	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
19	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
20	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
21	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
22	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
23	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
24	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
25	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
26	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수
27	2021-07-12	홍수	문장 부호(지침 4)	문장 부호(지침 4)	교정 X	교정 X	문장 부호(지침 4)	홍수

<그림 7> 질의응답용 구글 시트 예시

2.4. 수작업 전수 교정

맞춤법 교정 말뭉치의 구축에는 교정 작업을 효율적으로 할 수 있는 도구가 필요하다. 본 연구에서는 Kronoth라고 하는 웹 기반 전용 맞춤법 교정 및 주석 처리 도구와 마이크로소프트사의 엑셀(Excel) 프로그램 두 가지를 작업 도구로 사용하였다.

(1) Kronoth를 이용한 어휘 의미 분석 말뭉치 교정

Kronoth 도구는 병렬 말뭉치의 구축에 최적화된 ㈜이르테크의 교정 말뭉치 구축 전용 시스템으로, 교정 작업의 작업 정확성을 확보하고 작업 진도를 관리하는 등의 기능을 갖춘 도구이다.

메신저 어휘 의미 분석 말뭉치 100만 어절은 아래의 그림과 같은 Kronoth 도구에 탑재하여 맞춤법 교정 말뭉치의 교정 작업의 효율성과 정확도를 높이는 데 활용하였다.

다음으로 자동 교정 및 후처리가 끝난 말뭉치를 대화 단위와 발화 단위별로 시스템에 적재 및 변환 작업을 진행하였다. 다음은 Kronoth 도구에 말뭉치로 탑재되어 전수 수작업 교정 중인 작업자의 작업 화면이다.

맞춤법 교정 및 주석 처리 도구 v1.0



<그림 8> Kronoth(v1.0) 작업 화면

화면에서 보이는 교정 작업의 기능을 설명하면 다음과 같다.

첫째, 상단은 교정 전 원시 상태의 말뭉치를 보이는 ‘원 문장’ 영역으로, 원시 텍스트가 제시되어 교정 작업에 참고 자료가 된다.

둘째, ‘첨삭 영역’, 즉 하단의 좌측 영역은 교정 전후 내역이 제시되며 작업자가 교정 작업을 수행하는 영역이다. 여기에서 마우스로 교정이 필요한 어형을 선택한 후, 오른쪽의 ‘첨삭 기능’에서 교정 유형에 따라 버튼을 눌러서 교정 작업을 수행한다.

셋째, ‘첨삭 기능’은 교정 유형에 따라 삭제, 교체, 삽입, 찢어쓰기, 붙여쓰기 등 버튼이 제공되며, 이를 눌러서 해당되는 교정 작업을 수행한다. 삽입 기능의 경우, 작업 속도를 높이기 위해 기능을 세분화하였는데, 특정 위치(앞/뒤)의 삽입, 고빈도 삽입 내용인 마침표, 마침표와 공백의 동시 삽입 등으로 구분되어 있다. 위의 <그림 8>에서는 비활성화되어 있으나 활성화될 경우, 교정한 내용을 되돌리는(undo) 기능도 세분화되어 있다. 첨삭

삭제, 실행 취소, 다시 실행, 선택 해제 등이 작업의 빈도순으로 제시되어 있다.

넷째, 주석 기능은 하단의 맨 우측에 위치해있다. 이 사업은 교정뿐만 아니라 개인정보 및 혐오 표현 비식별화 작업을 포함하고 있는데, 이는 교정과 는 별도의 주석 작업에 속한다. 비식별화 정보는 그 유형에 따라 결과물에 다르게 마크업되기 때문에(예: 이름의 비식별화는 '&name1~9&', 전화 및 팩스 번호의 비식별화는 '&tel-num&'로 표시), 이를 교정 작업에서 동시에 주석할 수 있도록 혐오 표현, 비속어, 개인정보로 나누어 주석할 수 있도록 하였다. 특히 개인정보는 세부 정보의 유형에 따라 마크업이 다른 점을 감안하여, 개인정보-이름, 개인정보-식별정보, 개인정보-기타, 개인정보-전화번호, 개인정보-계좌번호, 개인정보-상세주소, 개인정보-소속, 개인정보-카드번호' 등의 버튼으로 세분화하였다.

본 연구는 메신저와 웹에서 사용되는 OoV 목록 추출 작업도 포함하고 있다. 이 작업은 말뭉치 구축 후 형태소 분석 및 사전 등재어와의 비교를 통해서도 추출이 되지만, OoV 특성상 형태소 분석 실패율이 높은 점을 감안하여, 교정 작업과 동시에 주석할 수 있도록 'OoV' 버튼을 주석 기능에 포함시켰다. 아울러, 전 단계 어휘 의미 분석에서 형태와 의미적 미등재어로 주석된 777, 888과 형태 오류인 999 표지를 그대로 살리고 이를 표시하여 작업자가 교정과 OoV 주석에 참고할 수 있도록 하였다.

특히, 주석 기능은 주석 대상과 기능에 따라 색상을 달리하여, 하단 좌측의 '첨삭 영역'에서도 작업자에게 직관적으로 제시되도록 함으로써 작업의 효율을 높였다.

이 밖에 교정 작업 중에 '임시 저장'하여 교정 중간 결과를 보존할 수 있으며, 해당 대화 파일에 대한 작업이 끝나면 '작업 완료'를 누름으로써 전수 교정 작업이 완료되고 1차 검수 단계로 파일이 이관된다.

Kronoth 도구는 작업 관리 기능 측면에서도 효율적인 기능을 갖추고 있다. 아래 그림의 예시와 같이 전체 표본(대화 파일) 현황, 기간별 작업 현황, 작업자별 작업 현황, 전체 현황 통계가 가능하다. 다음의 그림 예시들은 각각 이러한 작업 관리 화면을 보인 것이다.

KRONOTH

작성하님

아이메이지로그아웃

공지사항

나의 작업

작업 현황 관리

전체 표본 현황

기간별 작업 현황

작업자별 작업 현황

전체 현황 통계

전체 표본 현황

맞춤법교정

검수작업

상태-전체

우선 작업 상태순

표본명 검색

작업자명 검색

총 표본(건) : 2,697

<input type="checkbox"/>	No.	작업할당번호	작업자명	표본명	어절 수	작업상태	할당일	작업 완료일시	주석작업자명	반려
<input type="checkbox"/>	1	8244	송현주	MDRW1900005658.1	229	검수 할당	21-09-08 10:02	-	강윤희	-
<input type="checkbox"/>	2	8245	송현주	MDRW1900005659.1	253	검수 할당	21-09-08 10:02	-	강윤희	-
<input type="checkbox"/>	3	8246	송현주	MDRW1900005660.1	197	검수 할당	21-09-08 10:02	-	강윤희	-
<input type="checkbox"/>	4	8247	송현주	MDRW1900005661.1	173	검수 할당	21-09-08 10:02	-	강윤희	-

<그림 9> Kronoth의 작업 현황 관리 화면

작업자별 작업 현황

맞춤법교정

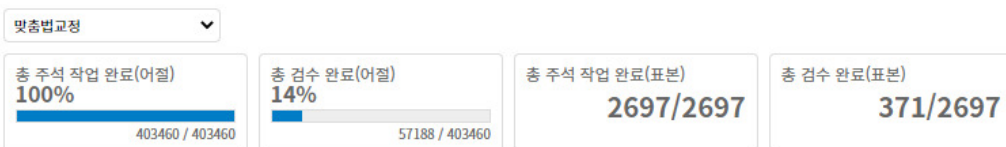
주석 작업

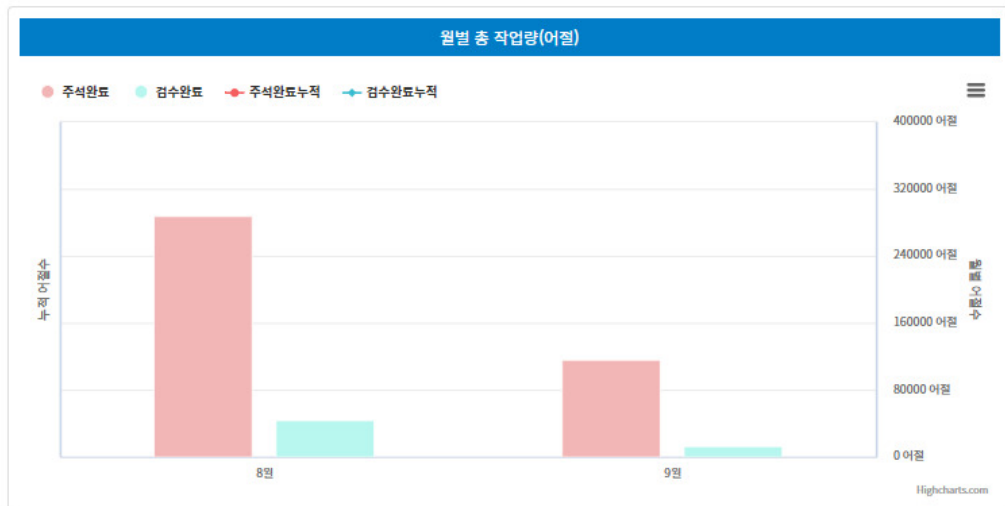
/		할당량		작업완료		미완료		완료율	
작업자	표본건수	총 어절	표본건수	어절	표본건수	어절	표본건수	어절	
심난희	208	31,136	208	31,136	0	0	100%	100%	
강신아	42	7,380	42	7,380	0	0	100%	100%	
강윤희	218	31,253	218	31,253	0	0	100%	100%	
이갑진	220	31,120	220	31,120	0	0	100%	100%	
백미경	163	26,646	163	26,646	0	0	100%	100%	
황지운	245	31,298	245	31,298	0	0	100%	100%	
고예린	48	7,183	48	7,183	0	0	100%	100%	
정나현	189	31,431	189	31,431	0	0	100%	100%	
김수지	188	31,347	188	31,347	0	0	100%	100%	

<그림 10> Kronoth의 작업자별 작업 현황

작업현황

전체현황





<그림 11> Kronoth의 전체 현황 통계 화면

(2) 엑셀(Excel)을 이용한 원시 말뭉치 교정

상술한 Kronoth 도구의 강력한 교정 및 작업 관리 기능에도 불구하고 아쉬운 점이 있다면, 전체 말뭉치에서 특정 오류 어형을 검색하여 일괄 바꾸기가 불가능하다는 점이다. 짧은 사업 기간을 고려할 때, 일괄 교정 기능은 작업 속도 향상과 동일 오류 형태에 대한 일관된 교정에 매우 필요한 기능이다. 이를 위해 본 연구에서는 엑셀 프로그램도 작업 도구로 병행하여 사용하였다.

엑셀을 이용한 교정 작업은 도구에 대한 작업자의 접근성이 높고 작업이 편리하며 일관된 오류 어형을 찾아서 일괄 처리가 가능한 장점이 있다. 맞춤법 오류는 중복도가 높은 오류가 많다. 이를 일일이 읽으면서 교정하는 데는 많은 시간이 필요하며 누락의 위험도 따른다. 이에 따라 본 연구는 메신저와 웹 원시 말뭉치 각 100만 어절에 대해서는 엑셀

	대화 ID	마감일	직업	검토자	최종 검토자	발화 ID	누적 어휘	회자 (C)	원 문장(해명 문서 삭제 X)	최종 직업 검정결과(교정 Ver 6.0)
MDRW1900002972	2021-07-25	장희선	안미애	남길일	MDRW1900002972.1.1.15	41	1	수업 다들 끝났어. 일한 할말은말때에 싶었음.	수업 다들 끝났어. 일한 할말은 말때에 싶었음.	
MDRW1900002972	2021-07-25	장희선	안미애	남길일	MDRW1900002972.1.1.16	45	2	아직 다들 초졸들이고 갠자들.	아직 다들 초졸들이고 갠자들.	
MDRW1900002972	2021-07-25	장희선	안미애	남길일	MDRW1900002972.1.1.17	51	2	나 저번에 일한것들이아니랑 선랑중랑 싸우논.	나 저번에 일한것들이아니랑 선랑중랑 싸우논	
MDRW1900002972	2021-07-25	장희선	안미애	남길일	MDRW1900002972.1.1.18	54	2	선랑중랑 가 일출때는 경찰 부를	선랑중랑 가 일출때는 경찰 부를	
MDRW1900002972	2021-07-25	장희선	안미애	남길일	MDRW1900002972.1.1.19	55	1	선랑중랑 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 존조	선랑중랑 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 존조이다.	
MDRW1900002972	2021-07-25	장희선	안미애	남길일	MDRW1900002972.1.1.20	58	1	근대 선랑중랑들이 아참 수영하고	근대 선랑중랑들이 아참 수영하고	

원 문장(개행 문자 삭제)	최종 작업 문장(원문 개정 Ver 6.0)	OoV(원)	작업자	검토자	협오 및	전화번호	개화번호	이름	주소	소속
수령다닐때 일진 할줄마를때문에 실었음 아직 다들 초보들이라 갸라는데 나 저번에 일진줄마를이랑 선랑줄마랑 싸우는 선랑줄마가 억울해서 경찰부름 선랑줄마ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 근대 선랑줄마들이 아침수영하고	수령 다닐 때 일진 할줄마를 때문에 실었음. 아직 다들 초보들이라 갸라는데, 나 저번에 일진줄마가이랑 선랑줄마랑 싸우는 선랑줄마가 억울해서 경찰 부름. 선랑줄마 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 줄웃기다. 근대 선랑줄마들이 아침 수영하고	할줄마			일진 할줄 마					
					일진줄마, 선랑줄마					
					선랑줄마					
					선랑줄마					
					선랑줄마					

<그림 12>를 보면 좌측의 흰 바탕색 구간은 대화 파일 ID, 발화 ID 등 작업 대상 파일에 대한 정보와 작업자, 검토자, 마감일 등 작업 관리를 위한 정보가 제시되어 있다. <그림 12>의 우측에 바탕색이 표시되어 있는 구간의 첫 칼럼은 해당 발화의 화자 정보를 담고 있다. 화자 교체 여부가 마침표를 부여하는 데 판단 기준의 하나가 되기 때문에 화자에 발화문의 바탕색을 달리하여 시각적으로 구분이 가능하도록 하였다. ‘원 문장’ 칼럼에는 교정 작업 전 발화문이 제시되어 있고, ‘최종 작업 문장’ 칼럼에는 자동 교정 및 텍스트 후처리된 발화문이 들어 있으며, 해당 칼럼에서 작업자가 교정 작업을 진행한다.

2.5. 개인정보와 부적절한 표현의 비식별화

(1) 개인정보의 비식별화

20

등), 각종 번호(고유 식별 번호, 전화번호, 금융 번호 등), 장소(상세 주소, 건물명 등), 출신 및 소속(학교, 직장, 부대 등) 등을 철저히 비식별화한다.

다음으로, 웹 말뭉치의 경우 민감한 개인정보인 ‘고유 식별 번호(주민 등록 번호, 사번, 학번 등), 전화 번호, 금융 번호, 주소’에 한해 비식별화한다.

(2) 부적절한 표현의 비식별화

부적절한 표현은 혐오 및 차별 표현, 욕설, 성적 표현 등을 포함하며, 비식별화 대상이다.

혐오 표현은 국가인권위원회의 혐오표현 리포트(2019)에 따르면 “성별, 장애, 종교, 나이, 출신지역, 인종, 성적 지향 등을 이유로 어떤 개인·집단에게 모욕, 비하, 멸시, 위협 또는 차별·폭력의 선전과 선동을 함으로써 차별을 정당화·조장·강화하는 효과를 갖는 표현”을 이른다.

혐오 표현에 대한 판단은 ‘형태’가 아닌 ‘맥락’을 기반으로 판단하였으며, 판단한 기준과 결과에 대해 국내 혐오 표현에 대한 전문가의 자문을 구하여 비식별화 여부를 확정하였다.

공인이나 기관의 경우는 비식별화 대상은 아니지만, 부정적인 내용이 포함될 경우에는 해당 대상을 비식별화하였으며 이 경우에도 상품, 상호, 영화명 등은 비식별화하지 않았다. 예를 들면, 상품이나 영화 등에 대한 부정적 평가는 비식별화 대상에서 제외하였다.

이상의 비식별화 작업은 일괄 교정이 불가능하므로 작업자와 검수자가 작업 및 검수 과정에서 수작업으로 진행하였다.

2.6. 품질 검수

품질 검수 단계에서는 맞춤법 수작업 전수 교정 결과에 대해 4차례에 걸친 품질 검수를 실시함으로써, 말뭉치의 교정 품질을 최대화한다.

(1) 1차 검수: 샘플링 검수

1차 검수는 전수 수작업 교정 결과의 10%에 대한 샘플링 검수로서, 교정 정확도가 낮은 작업자를 관리하는 동시에 교정 결과의 품질을 점검하고 향상시키는 역할을 한다. 이 작업은 주차별로 수행되며, 연구보조원이 교정한 결과물의 10%를 샘플링하여 상위 검수 집단인 공동 연구진들이 검수 및 교정하고, 작업자에게 피드백을 주는 방식으로 진행된다.

1차 샘플링 검수의 결과는 다음과 같은 세 가지로 반영된다. 우선, 작업자에게 주요 오류 유형을 알림으로써, 나머지 90%의 교정에 반영하도록 한다. 다음으로, 주요 오류 유형

을 수집하여 목록화함으로써 2차 검수에 대비한다. 끝으로, 정확률이 낮은 작업자에 대해서는 지침 및 작업에 관한 재교육을 제공하는 방식으로 작업의 품질을 높인다.

(2) 2차 검수: 오류 후보 목록을 이용한 일괄 검수

2차 일괄 검수는 1차에서 추출된 주요 오류 유형을 목록화하고, 이를 토대로 일괄 교정하여 전체적인 정확률과 서로 다른 작업자 간의 교정의 일관성을 높이는 데 기여한다. 이를 위해 오류 후보 목록의 작성과 분석, 이를 반영한 검수 작업을 진행하는데, 구체적으로는 다음과 같다.

첫째, 일괄 검수를 위한 주요 오류 후보 목록을 작성하되, 1차 검수에서 수집한 주요 오류 유형 목록에 일관된 표기와 띄어쓰기 준수가 어려운 외래어, 구 단위 표제어 등을 추가하여 작성한다.

둘째, 이 목록은 다시 자동 일괄 교정이 가능한 유형과 문맥 확인 및 전공 지식을 이용한 판단이 필요한 유형으로 나눈다.

셋째, 자동 일괄 교정이 가능한 유형에 대해서는 일괄 적용을 위한 규칙을 작성하여 반영한다. 이에 해당되는 목록은 다음과 같다.

예: 컴터 → 컴퓨터, 쌤 → 쌤, 라떼 → 라테, 큐알 코드 → 큐아르 코드...

아냐 아냐 → 아냐, 아냐, 가자 가자 → 가자, 가자...

인공눈물 → 인공 눈물, 음주운전 → 음주 운전...

넷째, 문맥 확인이 필요한 오류 후보 목록은 공동 연구진과 작업의 정확률이 높은 검수자로 구성된 검수팀에 의해 검수 및 교정한다. 다음과 같은 목록이 이에 해당된다.

예: 못 하다/못하다, 잘 하다/잘하다, 한번/한 번...

-(으)ㄴ데/-(으)ㄴ 데/(으)ㄴ대...

이 작업은 1, 2, 3차 납품 전에 수행되며, 1차는 전체 말뭉치의 20%, 2차는 70%, 3차는 100%에 대해 상이 어형 추출과 검수가 이루어진다.

(3) 3차 검수: 단발성 특수 어형과 비식별화 표지에 대한 검수

3차 검수는 맞춤법 교정 내용에 대한 검수와 비식별화 표지에 대한 분석으로 나뉜다.

우선, 맞춤법 교정에 대한 내용 검수는 메신저와 웹 텍스트의 특수성에서 착안한 검수 방법으로, 2차 검수가 완료된 말뭉치에서 단발어(hapax legomenon) 내지는 저빈도의 특수 어형을 추출하여 문맥을 확인하는 방식으로 진행되었다. 이러한 검수 절차를 거쳐 1, 2차 검수에서 누락된 저빈도 오류 어형을 찾아서 교정하였다.

다음으로, 비식별화 표지가 붙은 대화만 추출하여 비식별화 표지의 정확 여부를 확인한

다. 나아가 비식별화 표지가 붙은 어형과 같은 어형이 포함된 대화를 추출하여 비식별화 표지 부착 여부를 확인함으로써 비식별화 누락 오류를 바로잡고, 비식별화 작업의 일관성을 확보하였다.

또한, 1~3차 납품 결과에 대한 국립국어원의 피드백을 토대로 오류 유형을 목록화하여 전체 말뭉치에 반영하는 검수 및 작업도 진행되었다.

(4) 최종 검수: 형식 검수

최종 검수는 형식 검수이다. 3차례의 내용 검수가 완료된 말뭉치는 최종 결과물인 JSON 형식으로 변환하는데 이 과정에서 형식적인 오류가 걸러진다. 이를 교정 말뭉치에서 찾아서 수정하면 최종 검수가 완료된다.

2.7. 최종 결과물 산출

최종 결과물은 JSON 형식으로 구조화하는데, 맞춤법 교정 말뭉치의 JSON 구조와 결과물 양식에 대해서 기술하면 다음과 같다.

(1) JSON 구조

맞춤법 교정 말뭉치 구축의 최종 단계는 맞춤법을 교정한 결과를 JSON 형식으로 변환하여 최종 결과물을 산출하는 것이다. JSON 형식은 이 과제의 주관 연구 기관인 국립국어원과 협의해 결정하였다. JSON 구조는 교정 이전의 원시 말뭉치 또는 어휘 의미 분석 말뭉치의 JSON 구조를 계승하되 맞춤법 교정 결과는 문장(메신저의 경우 말풍선) 단위로 원문과 병렬하여 제시한다. <표 4>는 JSON 형식의 기본 구조이다.

1수준	2수준	3수준	4수준	타입	말뭉치 유형	분석 층위	설명
id				str	전체	전체	말뭉치 아이디
metadata				obj	전체	전체	말뭉치 메타 정보
	title			str	전체	전체	-
	creator			str	전체	전체	생성자: 국립국어원
	distributor			str	전체	전체	배포자: 국립국어원
	year			str	전체	전체	생성 년도
	category			arr (str)	전체	전체	분류
	annotation_level			arr (str)	전체	전체	맞춤법 교정
	sampling			str	전체	전체	샘플링 방식
document				arr (obj)	전체	전체	문서 정보
	id			str	전체	전체	문서 ID
	metadata			obj	전체	전체	문서 메타 정보

		title		str	전체	전체	문서 제목
		author		str	전체	전체	작성자
		publisher		str	전체	전체	출판사
		date		str	전체	전체	일시
		topic		str	신문 , 구어(준구어 제외), 메신저	전체	주제
		url		str	웹	-	URL 주소
		speaker		arr (obj)	구어(준구어 제외), 메신저	전체	* 발화자 정보
			id	num	구어(준구어 제외), 메신저	전체	발화자 ID
			age	str	구어(준구어 제외), 메신저	전체	나이. 모를 경우 "NA"
			occupation	str	구어(준구어 제외), 메신저	전체	직업
			sex	str	구어(준구어 제외), 메신저	전체	성별: 남성/여성
			birthplace	str	구어(준구어 제외), 메신저	전체	출생지
			principal_resi dence	str	구어(준구어 제외), 메신저	전체	주 성장지
			current_reside nce	str	구어(준구어 제외), 메신저	전체	현 거주지
			device	str	메신저	전체	메신저 사용 기기: 스마트폰/태블릿/PC
			keyboard	str	메신저	전체	자판 종류: 쿼티/천지인/나랏글/단 모음/기타
		setting		obj	구어(준구어 제외), 메신저	전체	환경 정보
			relation	str	구어(준구어 제외), 메신저	전체	관계: [가족] 부부...
	utterance			arr (obj)	맞춤법 교정 (메신저)	전체 (단, 원시 제외)	발화
		id		str	맞춤법 교정 (메신저)	전체 (단, 원시 제외)	발화 ID
		original_form		str	맞춤법 교정 (메신저)	전체 (단, 원시 제외)	원문 형태
		form		str	맞춤법 교정 (메신저)	전체 (단, 원시 제외)	정제된 형태

		corrected_from		str	맞춤법 교정 (메신저)	전체 (단, 원시 제외)	맞춤법 교정 형태
		speaker_id		str	맞춤법 교정 (메신저)	전체 (단, 원시 제외)	화자 ID
	paragraph			arr (obj)	맞춤법 교정 (웹)	전체 (단, 원시 제외)	문단
		id		str	맞춤법 교정 (웹)	전체 (단, 원시 제외)	문단 ID
		original_form		str	맞춤법 교정 (웹)	전체 (단, 원시 제외)	원문 형태
		form		str	맞춤법 교정 (웹)	전체 (단, 원시 제외)	정제된 형태
		corrected_from		str	맞춤법 교정 (웹)	전체 (단, 원시 제외)	맞춤법 교정 형태

<표 4> 맞춤법 교정 말뭉치의 JSON 형식 기본 구조

(2) JSON 양식

JSON으로 변환한 최종 결과물의 JSON 양식을 메신저 맞춤법 교정 말뭉치의 예를 들어 보이면 다음의 <표 5>와 같다.

```
{
  "id" : "MXSC2102111260",
  "metadata" : {
    "title" : "국립국어원 메신저 말뭉치 추출 MXSC2102112090",
    "creator" : "국립국어원",
    "distributor" : "국립국어원",
    "year" : "2021",
    "category" : [ "메신저 대화 > 2인 대화", "메신저 대화 > 다자 대화" ],
    "annotation_level" : "맞춤법 교정",
    "sampling" : "부분 추출 - 임의 추출"
  },
  "document" : [ {
    "id" : "MDRW1900000012.1",
    "metadata" : {
      "title" : "메신저 대화",
      "author" : "개인 대화 참여자",
      "publisher" : "카카오톡",
      "date" : "20191219",
      "topic" : "주거와 생활 (집안일, 육아, 부동산, 경제 활동, 생활 정보)",
      "speaker" : [ {
        "id" : "1",
        "age" : "30대",
        "occupation" : "전문가 및 관련 종사자",
        "sex" : "여성",
        "birthplace" : "울산",
        "principal_residence" : "부산",
        "current_residence" : "부산",
        "device" : "스마트폰",
        "keyboard" : "천지인"
      }, {
        "id" : "2",
        "age" : "50대",
        "occupation" : "서비스 종사자",
        "sex" : "여성",
```

```

    "birthplace" : "경북",
    "pricipal_residence" : "울산",
    "current_residence" : "울산",
    "device" : "스마트폰",
    "keyboard" : "천지인"
  },
  "setting" : {
    "relation" : "가족 : 부모-자녀",
  }
},
"utterance" : [ {
  "id" : "MDRW1900000012.1.1.1",
  "original_form" : "엄마~",
  "form" : "엄마~",
  "corrected_form" : "엄마~.",
  "speaker_id" : "1",
}, {
  "id" : "MDRW1900000012.1.1.2",
  "original_form" : "&name3&했나?",
  "form" : "name3했나?",
  "corrected_form" : "name3 했나?",
  "speaker_id" : "2",
}
]
}

```

<표 5> 메신저 맞춤법 교정 말뭉치의 JSON 양식

2.8. 결과물 납품

(1) 1차 납품

납품 말뭉치 유형 및 규모: 메신저 맞춤법 교정 원시 말뭉치 60만 어절

납품 시기: 2021년 7월 31일

(2) 2차 납품

납품 말뭉치 유형 및 규모: 1) 메신저 맞춤법 교정 원시 말뭉치 100만 어절, 2) 메신저 맞춤법 교정 어휘 의미 분석 말뭉치 100만 어절

납품 시기: 2021년 10월 1일

(3) 3차 납품

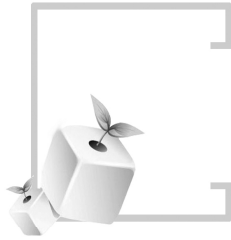
납품 말뭉치 유형 및 규모: 1) 메신저 맞춤법 교정 원시 말뭉치 100만 어절, 2) 메신저 맞춤법 교정 어휘 의미 분석 말뭉치 100만 어절 3) 웹 맞춤법 교정 말뭉치 60만 어절

납품 시기: 2021년 11월 3일

(4) 최종 납품

납품 규모: 1) 메신저 맞춤법 교정 원시 말뭉치 100만 어절, 2) 메신저 맞춤법 교정 어휘 의미 분석 말뭉치 100만 어절 3) 웹 맞춤법 교정 말뭉치 100만 어절

납품 시기: 2021년 11월 21일



제 3 장

교정 말뭉치 교정 지침 수립



1. 기본 지침과 지침 연구

메신저 및 웹 말뭉치의 경우, 신어 및 미등재어, 띄어쓰기, 다양한 형태의 비표준형 등의 문제로 기존의 문어 및 표준어를 학습 데이터로 사용한 맞춤법 교정 도구로는 교정의 정확도를 높이기 어렵다. 이에 본 사업에서는 표준화를 위한 초기 학습 데이터로서 ‘맞춤법 교정 병렬 말뭉치’의 구축을 목표로 문어 및 구어를 넘어서는 메신저 및 웹 언어의 특성을 반영한 교정 지침을 수립하되, 언어학적 정밀성과 공학적 활용도를 고려한 맞춤법 교정 지침의 수립을 목표로 하였다. 교정의 수준은 한국어 형태소 분석기 적용이 가능한 수준을 목표로 한다. 특히 공공재로서의 말뭉치의 활용도를 높이기 위해 개인정보 및 부적절한 표현의 범주를 엄격하게 분석하여 그 결과를 지침에 반영하였다. 맞춤법 교정 말뭉치의 기본 지침과 설계 방향은 다음과 같다.

1.1. 기본 지침

1) 한글 맞춤법 및 오타자 교정

메신저 및 웹 말뭉치에 나타나는 맞춤법 오류와 오타자에 대하여 교정한다. 이 경우, 메신저 및 웹 말뭉치에 특화되어 나타나는 표현과 표기의 처리 부분을 고려할 필요가 있다. 메신저 및 웹 말뭉치의 경우, 사용자 생산 콘텐츠로, 문어와 구어의 특성을 동시에 가지기 때문이다. 메신저와 웹 각각도 표현과 이해의 목적에 차이가 있어 이러한 점을 교정 지침 마련 시 고려할 필요가 있다. 웹의 경우, 웹 페이지의 목적에 따라 표현 언어에 차이가 있으며, 메신저의 경우, 참여자 간의 사회적 관계, 대화 주제와 대상, 대화 환경 등에 따라 다양한 표현과 표기가 나타난다. 이 과정에서 비문법적 표현과 비표준형들이 다수 나타나며 표현의 극대화를 위해 다양한 기호를 이용하는 양상도 관찰된다. 이에 메신저 및 웹 말뭉치의 한글 맞춤법 및 오타자 교정 지침 설계 시에는 이러한 점을 고려하였다.

2) <우리말샘> 미등재어(외래어, 신어) 및 비표준어의 처리

메신저 및 웹 말뭉치에는 다수의 <우리말샘> 미등재어와 비표준어가 나타난다. 미등재어의 경우, 외래어, 신어, 구어, 방언 등으로 나누어 처리하도록 지침을 수립하였다. 미등재어의 경우, OoV(Out of Vocabulary) 목록을 구축하여 관리하도록 하였다.

3) 개인정보 및 부적절한 표현에 대한 처리

개인정보 및 부적절한 표현의 처리는 공공재로서의 말뭉치의 성격을 제고하기 위해 반드시 필요한 작업이다. 앞서 언급한 바와 같이 사업의 전체적인 유기성을 고려해 2019년에 추진한 국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업에서 제시된 범주에 의거하여 비식별화하였으며, 웹과 메신저 말뭉치 교정과 검수 작업을 통해 맥락에 따라 다시 판단하여 비식별화 대상을 추출하였다. 이렇게 추출한 비식별화 결과물을 바탕으로 전문가의 자문을 구하였으며 그 결과를 다시 지침에 반영하여 맞춤법 교정 말뭉치의 공공재로서의 성격을 제고하였다.

1.2. 지침 연구

이 절에서는 지침을 구축하기 위한 본 사업팀의 지침 연구 설계 단계별 세부 내용에 대해 설명하고자 한다. 본 사업팀은 맞춤법 교정 말뭉치의 교정 지침을 수립하기 위해 선행 연구 검토, 교정 대상 말뭉치의 분석과 샘플링 교정을 바탕으로 한 지침 수립, 실제 교정을 통한 지침 보완 및 추가의 단계로 지침 수립을 위한 연구를 진행하였다.

(1) 선행 연구 검토

본 사업의 대상인 웹과 메신저 말뭉치는 사용자 생성 콘텐츠(User Generated Content, UGC)로, 규범성의 결핍이 이러한 말뭉치 유형이 가지고 있는 중요한 특성이다. 기존의 전통적 의미의 텍스트는 일반적으로 생산 과정에서 전문가의 교정과 교열 작업이 선행되지만 UGC는 그렇지 않다. 텍스트 입력의 편의성을 높이기 위해서 텍스트가 갖춰야 할 규범성을 무시하는 경우가 많으며(띄어쓰기 무시) 교정과 교열의 부재로 오자와 탈자, 비문법적 표현이 다수 존재한다. 구어성이 강조된 텍스트기에 구어체나 창의적인 표현, 감정 표현을 위한 텍스트 기반 이모티콘 등이 빈번하게 쓰이기도 한다. 본 사업은 이러한 점을 고려하여 웹과 메신저 말뭉치의 맞춤법 교정을 위한 지침 연구 단계를 사업 수행의 첫 번째 단계로 두었다.

(2) 말뭉치 분석 및 샘플링 교정을 통한 지침 수립

지침 연구 단계에서는 맞춤법 교정을 위한 지침 마련과 혐오 및 부적절한 표현의 비식별화 범위에 대한 지침을 마련하였다. 맞춤법 교정을 위한 지침을 마련하기 위해 먼저 기존의 맞춤법 검사기를 이용해 1차 자동 교정 작업을 시행하였다. 두 번째 단계는 1차 자동 교정 말뭉치를 대상으로, 파일럿 수작업 교정 작업을 통해 맞춤법 교정 말뭉치를 구축하는 동시에 오류 교정 유형과 과도 교정형을 도출하여 목록화하는 작업을 시행하였다. 이를 통해 아래와 같이 지침 마련을 위한 기본 오류 유형이 수립되었다.

1. <맞춤법 교정 말뭉치 기본 오류 유형>
2. [유형1] 한글 맞춤법 미준수형
3. [유형2] <우리말샘> 미등재어형(외래어, 신어 등)
4. [유형3] 개인정보 및 부적절한 표현
5. [유형4] 특수 표현
6. [유형5] 방언형
7. [유형6] 문장부호

(3) 실제 교정 작업을 통한 지침의 수정 및 보완

이 단계에서는 전체 교정 대상 말뭉치의 10%에 대해 시행한 샘플링 교정 작업을 통해 발견된 <우리말샘> 미등재어(외래어, 신어 등)와 비표준어의 처리 방안 모색 작업을 진행하였다. <우리말샘> 미등재어의 유형은 고유명사와 준말(예: 냇이, 겸둥), 신어(예: 깨발랄), 방언(예: 개안타)과 전달 효과를 고려하여 개인적으로 사용하는 표현들이다(예: 응, 응응).

또한 웹과 메신저 말뭉치 특유의 도메인에 따라 나타나는 특수 표현의 처리 방안 마련 작업도 병행하였다. 웹의 경우, 해당 텍스트가 추출된 웹 페이지의 목적과 특성에 따라 표현에 차이가 있을 수 있으며, 메신저의 경우, 대화자 간의 관계나 주제와 대화 환경, 목적 등에 따라 다양한 변이 표현과 표기가 나타날 수 있기 때문이다. 이러한 점을 고려하여 이 사업에서는 웹 말뭉치와 메신저 말뭉치 각각의 교정 지침을 수립하였고, 특히 웹 말뭉치에 나타나는 오류 유형과 교정 대상의 유형을 아래와 같이 추가하였다.

8. <웹 말뭉치의 오류 및 교정 대상의 유형>
9. [유형1] 해시태그(#) 이하의 처리
10. [유형2] 웹 말뭉치 특유의 외래어/ 외국어의 처리
11. [유형3] 웹 말뭉치 특유의 기호, 문장부호의 처리
12. [유형4] 기타 세부 사항
13. [유형5] 웹 말뭉치의 비식별화

마지막으로 개인정보 및 혐오, 부적절한 표현의 비식별화 범위를 지정하기 위한 지침 작업을 시행하였다. 개인정보는 사업의 전체적 유기성을 고려하여 2019년에 추진한 국립국어원 메신저 대화 자료 수집 및 말뭉치 구축 사업에서 제시된 범주에 의거해 1차 비식별화 작업을 진행하였다. 혐오 및 부적절한 표현은 먼저 <우리말샘> 등재어와 용례를 통해 관련 표현을 확보하였다. 이 목록을 기준으로, 100만 어절을 대상으로 한 샘플링 검수 작업을 통해 웹과 메신저 말뭉치에 나타나는 개인정보와 혐오, 부적절한 표현의 목록을 보완하였다. 이 목록은 법학 및 사회 언어학 분야의 전문가의 자문을 받아 비식별화 지침에 반영하였다.

비식별화 지침 수립 시, 혐오 차별 표현의 경우, 작업자와 검토자의 이해를 돕기 위해 유형별로 구체적인 사례를 제시하여 유형별 처리 방안을 수립하여 시행하였으며 작업 중에도 혐오 및 차별 표현의 유형을 수집하여 지침을 수정하였다.

2. 맞춤법 교정 말뭉치 교정 지침

이 장에서는 본 사업에서 메신저와 웹 말뭉치의 맞춤법 교정 말뭉치를 구축하기 위해 적용한 지침을 제시하고자 한다.

I. 사업의 목적과 교정의 수준

1. 사업의 목적과 교정의 수준
2. 기본 원칙

II. 유형별 지침

[지침1] 한글 맞춤법에 따른 띄어쓰기, 오타자 등의 교정

[지침2] <우리말샘> 미등재어(외래어, 신어 등)의 처리

[지침3] 개인정보 및 부적절한 표현(욕설, 혐오 표현 등)의 비식별화 처리

[지침4] 특수 표현의 유형 분류와 유형별 처리

[지침5] 방언형의 처리

[지침6] 문장부호의 처리

Ⅲ. 웹 말뭉치 작업 지침

[지침1] 해시태그(#이하 내용)에 대한 지침

[지침2] 외래어/외국어 관련 지침

[지침3] 기호, 문장부호 관련

[지침4] 기타 세부 사항

[지침5] 비식별화

1. 사업의 목적과 교정의 수준

이 사업의 목적은 자동 형태소 분석, 기계 번역 등 한국어 처리 도구가 메신저 원시 말뭉치를 분석할 수 있는 수준으로 교정하되, 메신저 언어의 특수성을 살린 교정 병렬 말뭉치를 구축하는 데 있다.

본 사업의 목적이 이상적인 교육용 규범 말뭉치를 구축하는 데 있기보다는 일종의 이개어(원 문장과 교정 문장) 병렬 말뭉치, 기계를 위한 학습 데이터로서의 교정 말뭉치를 구축하는 데 있는만큼, 다음과 같은 교정의 수준을 전제로 한다.

- 1) 메신저 표기의 오타자, 비표준형, 띄어쓰기 등을 **구어 전사 말뭉치 수준**으로 교정한다.
- 2) 맞춤법 검사기, 자동 형태소 분석기 등 기계가 처리할 수 있는 수준에 한해서 엄격한 규범을 추구하기보다는 **지침에서 명시된 허용 규정을 적용**한다.(예: 보조용언의 띄어쓰기, 문장부호 지침, 외래어 규정 등은 별도 지침 참조)
- 3) 개인정보를 비롯하여, 욕설, 혐오 및 차별 표현 등 **부적절한 표현을 비식별화한 교정 말뭉치**로 가공한다.

2. 기본 원칙

1) 유형별 지침

문어, 구어와 구분되는 메신저의 특수성을 살린 교정 대응쌍의 구축을 위해, 메신저에서 자주 등장하는 감탄사나 부사, 특수 표현, 자소형 표기, 이모티콘 등의 교정은 별도의 유형별 지침을 따른다.

2) 기본 작업 과정

이 작업은 자동 검사기 처리를 거친 후 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식으로 이루어지므로, 맞춤법 검사기의 오교정과 과교정에 유의하여야 한다.

3) 병렬 구조를 고려한 단위 설정과 교정

맞춤법 교정 병렬 말뭉치의 병렬 구조는 기본적으로 메신저의 말풍선에 해당하는 ‘발화 단위’의 대응을 기본으로 하고, ‘발화 단위’ 즉 하나의 말풍선 내에 두 문장 이상이 포함된 경우는 문장 대응 구조를 상정한다. 이를 위한 문장의 판별, 원 문장 대 교정 문장의 대응 구조 설계를 위한 구획 표지는 [지침 6] 문장부호 지침을 따른다.

Ⅱ. 유형별 지침

[지침1] 한글 맞춤법에 따른 띄어쓰기, 오타자 등의 교정

1. 기본 작업 과정

1) 이 작업은 자동 검사기 처리를 거친 후 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식으로 이루어진다.

❶ 메신저 원시 말뭉치	❷ 자동 검사기 처리	❸ 최종 작업(수작업 교정)
아이구 힘들었겠다 아침에8시에나와서두유만먹어가 지구 끝났웅? 오늘할수있는일은곳 적응잘하려는지거경되죽겠다.	아이고 힘들었겠다. 아침에 8시에 나와서 두유만 먹어서 끝났어? 오늘할수있는일은곳 적응잘하려는지거경되죽겠다.	아이구 힘들었겠다. 아침에 8시에 나와서 두유만 먹어 가지고 끝났어? 오늘 할 수 있는 일은 끝 적응 잘하려는지 걱정돼 죽겠다.

<표 6> 교정 단계와 단계별 교정 내용

2) 표준형과 비표준형의 판단 기준: 표준형과 비표준형의 기준은 <우리말샘>의 등재 여부로 한다.

예) 끝났웅? → 끝났어?

※ 단, 아래의 ‘괘찮’과 같이 용언의 어간형만으로 준말을 삼은 경우는 <우리말샘>이라도 따르지 않는다.

괘찮 「001」 주로 인터넷상에서 별로 나쁘지 않고 보통 이상임을 표현할 때 쓰는 말.

3) 맞춤법 검사기에서 구어 표현을 다른 단어로 대치하는 등 과하게 수정한 경우, 원래

의 단어로 복원한 후 표준형으로 수정한다.

예) <1> 먹어가지구 → <2> 먹어서 → <3> 먹어 가지고(또는 먹어가지고) <1> 아이구 → <2> 아이고 → <3> 아이구
--

2. 띄어쓰기

1) 한글 맞춤법의 띄어쓰기 규정을 따른다.

2) 보조용언의 띄어쓰기

보조용언은 맞춤법의 규정에 따라, 띄어 쓰음 원칙으로 하되, 경우에 따라 붙여 쓰기도 허용한다. 이는 현재의 맞춤법 규정을 따르는 동시에, 현재 형태소 분석기, 맞춤법 검사기 등의 한국어 처리 도구의 정확도가 보조용언 띄어쓰기에 영향을 받지 않는 것을 고려한 것이다.

예1) 가보고 싶어(o), 가 보고 싶어(o)

예2) 가보니(o), 가 보니(o)

※ 단, 어절이 지나치게 길어지는 경우는 띄어 쓴다.

예) 이야기해봐야겠구먼 → 이야기해 봐야겠구먼

※ 단, 맞춤법 규정에 따라 ‘-어지다’, ‘-어하다’의 구문의 경우는 붙여 써야 한다.

예) (낙서가) 지워진다, (아기를) 예뻐한다

3) 구어에서 ‘-고 하-’나 ‘하-’가 생략된 경우 띄어쓰지 않는다.

예) 가야겠어요, 가봐야겠어요, ...

4) 신조어 파생어의 경우 붙여 쓴다. 이들은 <우리말샘>에 등재되지 않은 새로운 의미의 접사가 붙어 만들어진 단어들로 다음과 같은 것들이 있다.

ㄱ. <우리말샘>에 ‘갓(god)’ 명사는 등재되어 있지만 접사에 대한 기술은 없음.

예) 갓-: 갓동원, 갓뚜기, 갓보검, ...

ㄴ. <우리말샘>에 ‘꿀’ 명사(매우 뛰어나거나 좋음.. 목소리가 꿀이다)는 등재되어 있지만 접사에 대한 기술은 없음.

예) 꿀-: 꿀나은, 꿀잼, 꿀목소리, ...

ㄷ. <우리말샘>에 ‘핵(核)’ 명사는 등재되어 있지만 접사에 대한 기술은 없음.

예) 핵-: 핵공감, 핵노잼, 핵인싸, ...

ㄹ. <우리말샘>에 접사 ‘개-’는 부정적 의미로 정도가 심한 뜻만 등재되어 있음.

예) 개-: 개간지, 개꿀잼, 개귀엽다, 개좋아, ...

ㄹ. ‘급’: <우리말샘>에 접사 ‘급-’은 “(일부 명사 앞에 붙어) ‘갑작스러운’의 뜻을 더하는 접두사.”로만 기술되어 있음.

예) 급-: 급생각해서, 급어필했지...

※ 다만, 다음처럼 접사가 아닌 부사처럼 사용된 ‘개-’, ‘급-’은 띄어 쓴다.

-‘(시험을) 개잘봤네’, ‘개폭자버렸다’와 와 같이 접사로 보기 어려운 경우 ‘개’는 부사로 보고 ‘개 잘 봤네’, ‘개 폭 자 버렸다’로 띄어 쓴다.

-‘급’의 경우, <우리말샘>에는 명사 파생 접사로만 등재되어 있으나, 명사 파생 접사로 설명될 수 없는 아래 예들은 부사로 보고 띄어 쓴다.

예) 급 먹고 싶어, 급 흥미 잃었어, 급 채솟값도 오르겠네, 급 둘이 술 먹다가, 급 떠오르는, 급 닭볶음탕 당기네

5) 반복 표현의 띄어쓰기

감탄사나 응답 표현, 정도부사 등은 <우리말샘>에 준하여 붙여 쓰고, <우리말샘>에 등재되지 않은 반복 표현의 경우에도 감탄사, 부사에 한해서는 붙여 쓴다.

예) 빨리빨리, 너무너무, 야금야금, 팡팡팡, 그래그래, 그치그치, 그쵸그쵸, ...

(2) 용언의 반복은 띄어 쓴다. 가운데에 쉼표를 넣는다.

예) 알아알아 → 알아, 알아.

조아조아 → 좋아, 좋아.

(3) 다음과 같은 유형의 명사 반복은 감탄과 강조의 의미를 지니므로 붙여 쓴다.

예) 기대기대, 인정인정, ...

(4) 반복된 명사가 파생어를 만든 경우도 역시 붙여 쓴다.

예) 힐링힐링해, 여자여자해, 주저주저주저하다, 두근두근두근거리다...

단, 강조의 의미가 아닌 인용의 의미로 명사를 반복하거나 구 단위와 결합하는 경우, 하나의 용언을 만들었다고 보기 어려우므로 띄어 쓴다.

예) 화자 1. 알겠어, 그래서 오늘 운동할 거냐고.

화자 2. 뭘 운동 운동 거리고 있어.

(5) 대명사 등을 반복한 경우는 띄어 쓴다. 사이에 쉼표를 붙이지 않는다.

예) 어디 어디, 그거 그거, 뭐 뭐, ...

6) 전문용어의 띄어쓰기

전문용어의 띄어쓰기는 허용 규정을 따라 띄어 쓰는 것과 붙여 쓰는 것 모두를 허용한다.(자동교정 반영도 허용함)

예) 영상통화(○), 영상 통화(○) ※<우리말샘>에 표제어로 ‘영상^통화’가 등재되어 있음.

(2) 전문용어의 기준은 <우리말샘>의 전문어 표지 부착 여부에 따른다. 따라서 전문용어가 아닌 구 단위(전문어 표지가 없는 경우)는 <우리말샘>에 따라 띄어 쓴다.

예) 단체 사진, 남자 친구, 코인 노래방, 커피 향, 감상 평 등

7) 기타 띄어쓰기의 허용 지침

(1) <우리말샘>에 미등재된 ‘-하다, -되다, -시키다, -받다’ 등의 결합 복합어는 붙여 쓰되, 띄어 쓰는 것도 허용한다.

예) 공부시키다(○)/공부 시키다(○), 충격받다(○)/충격 받다(○)

(2) ‘명사+명사’ 합성어에서 <우리말샘> 등재어의 구성성분과 동일한 구성성분을 가지고 계열 관계를 이루는 미등재 복합어의 경우, 미등재 어형의 일반적인 지침을 따라 띄어 쓰되, 등재된 어형의 계열관계를 고려하여 붙여 쓰는 것도 허용한다.

예1) 염통꼬치(○)/염통 꼬치(○), 순대꼬치(○)/순대 꼬치(○) cf. 양꼬치(등재), 염통구이(등재)

예2) 양꼬치집(○)/양꼬치 집(○)

10) 기호 관련

비식별화 기호가 들어가 있는 경우라도 다음과 같이 조사, 의존명사와 단어를 구별하여 띄어쓰기를 한다.

예) name2언니	→ name2 언니	(띄어쓰기 필요)
name3선생님이	→ name3 선생님이	(띄어쓰기 필요)
name3 님은		(띄어쓰기 필요, 님: 의존명사)
name2이랑		(띄어쓰기 불필요, 이랑: 조사)

3. 구어형(축약형 포함)과 메신저 비표준형의 교정

1) 구어 표현과 비표준형의 처리

구어 표현이나 메신저에서 많이 나타나는 다음과 같은 비표준형은 괄호와 같이 수정한다.

(1) 음운이 탈락된 경우

예) 마이(많이), 암꺼나(아무거나), 개안나(괜찮아), ...

(2) 음운이 첨가된 경우

예) 고마워염(고마워요), 제법 해욱(제법 해요), 가야징(가야지), 줄일려고(줄이려고), 싫으다(싫다), 꼬옥(꼭), 감좌(감사), 너동(너도), 그럴깁(그렇게)...

(3) 음절을 첨가하여 장음을 표현한 경우: 의미를 강조하기 위해 음절을 첨가하여 장음을 표현한 단어는 교정한다.

예) 고오오급 음식(고급 음식), 최에에고(최고), 너어무(너무), 좋다아아아(좋다), ...

(4) 음운이 교체된 경우

예)쌔돈(생돈), 사주께(사줄게), 찔라야(잘라야), 할꼬얌(할 거야), 체험하구(체험하고), 별루(별로), 지대로(제대로), 이뿌다(이쁘다), 슬푸당(슬프다), 귀웁다(귀엽다), 모야(뭐야), ...

2) 군더더기 표현

다음의 ‘달, 날’과 같이 구어 특유의 군더더기 표현은 수정하지 않는다.

예1) 1월달, 2월달, ...

예2) 1일날, 2일날, 토요일날, ...

예3) 평일날, 생일날, ...

3) 고빈도 구어형 ‘니’, ‘지’, ‘거’, ‘머’의 처리

(1) ‘니’와 ‘지’의 처리

예) 니꺼만 → 니 거만

ㄱ. ‘니’와 ‘거’가 <우리말샘>에 구어 표현으로 등재되어 있으므로 ‘니 거만’으로 수정

※ 현재 검사기에서는 ‘네 것만’으로 수정하므로 주의

ㄴ. <우리말샘>에 ‘꺼’는 다음과 같이 기술되어 있으므로 ‘거’로 수정한다.

※ 꺼 의존명사 001 ‘것’을 구어적으로 이르는 말. → 규범 표기는 ‘거’이다.

ㄷ. 방언형 ‘니’의 경우는 수정이 필요하므로 [지침5]에 기술된 유형에 따라 수정을 한다.(참고: 지침5의 ‘니’ 관련 기술: 예1) 방언형 ‘니’는 → 너는(‘니’를 수정하지 않는 경우는 ‘네’의 의미일 때뿐임 주의.)

ㄹ. ‘지가’와 ‘지 말대로’의 ‘지’는 ‘제’가 규범형이지만, 구어에서 자주 등장하고, ‘제가, 제 말대로’의 교정형 대로는 거의 쓰이지 않으므로 수정하지 않고 그대로 둔다.

(2) ‘거’의 처리(‘거’는 우리말샘 등재어임을 고려)

ㄱ. ‘르’이 덧나는 경우

예) 가는걸로 → 가는 거로

ㄴ. ‘거’의 활용형의 경우: 그대로 둠

예) 거임, 하는 거다.

(3) ‘머’와 ‘모’(대명사)의 처리

‘머’는 사전에 구어적으로 이르는 말로 기술되어 있으므로 수정하지 않는다. 단, ‘모’는 ‘뭘’로 교정 한다.

4) 비표준 종결어미의 처리

(1) 맥락과 화계를 고려한 처리

비표준 종결어미의 경우 맥락과 화계를 고려해 아래에 제시한 예시와 같이 적절한 종결어미로 교정한다(괄호 안과 같이 교정한다.).

예) 거기 가삼.(거기 가.), 내가 했삼(내가 했어.), 배불러서리(배불러서), ~를 찾아싸?(찾아쌍니?)...

※ 단, ‘-지롱’과 같이 특별한 의미(놀림)가 추가된 경우는 위와 같이 교정하지 않는다.(OoV에 추가)

(2) 문말에 사용된 ‘-음’의 처리

예1) 나 놀고 있음. → ‘있어’로 교정하지 않음

예2) 내일 집에 감? → ‘가/가요?’로 교정하지 않음

5) 호칭 변용의 처리

호칭 변용의 경우, 아래에 해당하는 예시는 그대로 둔다(이름만 비식별어로 처리함).

예) 효진쓰 → name1 쓰

비식별어의 번호는 기준에 따라 부여한다.

의미 없이 붙는 ‘-쓰’의 경우 위의 5)의 예와 같은 고유명사에 붙는 경우 이외는 모두 생략한다.(괄호처럼 수정)

예) 위험쓰 → 위험, 꿀잼쓰 → 꿀잼, 다행쓰 → 다행, 빵쓰 → 빵 ...

용언 어간에 붙는 ‘-쓰’의 경우 ‘-쓰’를 삭제하고, 화계를 고려하여 적절한 어미를 추가한다.

예) 괜찮쓰 → 괜찮아/괜찮아요, 먹쓰 → 먹어/먹어요

6) 축약 표현의 처리

축약 표현의 경우, <우리말샘>에 등재된 유형은 교정하지 않는다. 이때 방언도 등재된 유형으로 보아 수정하지 않는다. 그리고 <우리말샘>에 없더라도 고빈도로 쓰이고 자주 쓴다고 판단될 경우, 별도의 목록을 관리하고 수정하지 않는다.

(1) <우리말샘>에 방언으로 되어 있지만, 다음은 실제 통용되는 구어이고 널리 자주 쓰이는 유형은 수정하지 않는다.

예) 강, 그니까, 이케, 냅두다, 함, 여튼, ...

(2) <우리말샘>에 등재되지 않았지만 자주 쓰는 준말도 교정하지 않는다.

예) 어딴다, 여깁다, 왜냐면, ...

(3) 괄호처럼 교정하는 경우

예) 집(x) → 지금(o), 어캄(어떡함), 어카지(어떡하지)

(4) 일반적으로 사용하는 조사, 어미의 축약형 등은 교정하지 않는다.

예1) 난, 날 (o) : 나는, 나를...등으로 교정하지 않고 그대로 둔다.

예2) 하고 싶긴 했는데(o) : ‘기는’으로 교정하지 않고 그대로 둔다.

(5) ‘-으면’의 축약형인 ‘-음’은 교정하지 않는다.(<우리말샘>에 등재)

예) 맞아요, 밖에서 사 먹음 국밥도 7천 원씩 하더라고요.

(6) 3중 모음의 축약은 축약 전의 형태로 교정한다.

예) 바뀌서 → 바뀌어서(o), 사겨서 → 사귀어서(o)...

4. 감탄사와 의성의태어의 교정

1) 감탄사와 의성의태어의 경우 메신저 언어의 감정 표현의 다양성과 특수성을 고려하여 다양한 변이형이 존재하고, <우리말샘>에 소극적으로 등재되어 있으므로, 되도록 수정하지 않는다.

예) 넵, 넹, 아앗, 와아, 크으, 호오, 아우, 오웅, 와아, 우와와, 우앙, 웅, 웅웅, 까짓거, 까아, 까아아아아, 까아아아악, 까아아악, 까아앙, 까악, 까오, ... 네에, 네에네, 네에에, 네, 네에, ...아아니, 아아아, 아아아아, 아아아아아, 아아아아악, 아아앗, 예그마, 예또, 예라이, 예에에, 카하, 크으, 크크, 크흐, 크흑...

2) 단, 아래와 같은 명백한 오류는 수정하고 혐오 표현은 비식별화한다.

예) 아쌈/아씨 → 욕설 및 혐오표현, 오면아 → 어머니, ㄴㅏㅓㅓ → 우와

5. 어미 생략의 처리

다음과 같이 어간만으로 끝나는 경우 "-어/아' 또는 '-어/아요' 둘 중 하나를 화계와 맥락에 따라 선택한다. 단, '-(ㄴ)다'의 형태로 수정하지 않는다(예: 괜찮아(o), 괜찮아요(o), 괜찮다(x))

예) 아래와 같은 문맥에서 밑줄 친 ‘괜찮-’의 처리

문맥 어디 갈지 고민 중
 엥 심해?
 오웅
 오늘 계속 나서
 괜찬?

처리 전	처리 후
괜찬	괜찮아.

괜찬?	괜찮아?
괜찮	괜찮아.
괜찮으?	괜찮아?
괜찬!	괜찮아!
괜춘	괜찮아.
괜춘?	괜찮아?
괜춘쓰	괜찮아.
괜춘해	괜찮아.
괜툐	괜찮아.
괜툐?	괜찮아?

※ ‘대단하다, 훌륭하다’와 같이 어근 ‘대단, 훌륭’으로 끝난 경우는 수정하지 않는다.

6. 외래어/외국어의 교정

1) <우리말샘>의 등재어를 기준으로 규범에 맞게 교정한다.

예) 요거트 → 요구르트

2) 미등재어 또는 등재어 중 규범이 정해지지 않은 외래어나 외국어는 교정하지 않는 것을 원칙으로 하나, 외래어표기법에 따라 교정한 경우도 인정한다.

예) 외래어 고유명사 교정 사례

원문	교정문	교정 근거나 이유
스타듀밸리	스타듀밸리	지침을 따라 그대로 둠
스타듀 밸리	스타듀 밸리	지침을 따라 그대로 둠
스타듀밸리	스타듀 밸리	자동 교정 결과
스타듀 밸리	스타듀 밸리	자동 교정 결과

3) 응답 표현의 처리: 감탄사 처리 원칙에 따라 수정하지 않는다. (아래 예시는 참고용)

예) 노, 늑, 예스, 노우, 노늑, 노우노우, 노노, 예스예스, …

4) 외래어를 순화어로 교정하지 않는다.

예1) 타코 와사비(‘와사비’를 ‘고추냉이’로 교정하지 않음.)

예2) 기스가 찌네(‘기스’를 ‘흠, 흠집’으로 교정하지 않음.)

7. 고유명사(상품명, 상호명 외)의 띄어쓰기, 맞춤법 교정

1) 상호명, 상품명 등의 고유명사는 맞춤법 검사기 결과를 사용하되, 외래어 표기 및 띄어쓰기 전체를 교정하지 않는다. 특히 웹 말뭉치 등에서 외래어 상품명 등이 상당수 반복적으로 등장하는데, 맞춤법 검사기의 결과를 수용하되, 한 문맥 내에서 일관성을 맞추는 방향으로 교정한다.

2) 단, 인접 문맥에 한하여 일관성을 고려하여 수정하고, 연구진 회의에서 논의한다. 문맥의 고려 범위는 다음과 같다.

ㄱ. 한 문맥 안에서 표기나 띄어쓰기가 일치되지 않은 경우는 일관성을 고려하는 선에서 수정한다. 고유명사의 띄어쓰기의 경우는 형태소 분석기나 기계 번역 등에서 큰 문제를 일으키지 않으므로 전체를 일괄 수정하지는 않는다.

ㄴ. 일부 외래어 표기 지침과 상충되는 경우 또는 규범형 표기로 수정하는 것이 합당하다고 판단되는 경우, 맞춤법 검사기에서 일관되게 수정한 경우 등의 사례에 대해서는 연구진 회의에서 논의한 사항을 따른다.

3) 고유명이 줄어든 경우도 수정하지 않음

예) 파바(*파리바게트가 줄어듦), 홈플(*홈플러스가 줄어듦), 필핀(*필리핀이 줄어듦) ...

8. 표현상의 오류 처리

1) 널리 사용되는 틀린 표현, 문법 오류 등은 교정하지 않는다.

예1) 이거랑 저거랑 틀리다. → ‘다르다’로 수정하지 않음.

예2) 어묵 먹자. 그리고 나서 빙수 먹으러 가자. → ‘그리고 나서’로 수정하지 않음

2) 명백한 맞춤법 오류나 과도한 축약은 수정하므로 이에 유의한다.

예1) 정신줄 놔 → ‘놓음’으로 교정

예2) 케이크 만듬 → ‘만들’으로 교정

9. 의미 불명 표현의 처리

다음과 같이 의미를 파악할 수 없어 교정이 불가능한 경우는 교정을 하지 않는다. 다만, OoV로 수집 및 관리할 수 있도록 별도로 처리한다.

예) 만떡챗, 닉병, 력부섬야, 학약 등

10. 기타

화자가 오타를 인지하고 자진 수정한 경우에도 오타는 수정한다.

예1) 아, 역사

역시.

→ 아, 역시

역시.

예2) 잘 키우실 수도 있으니까

까

가족끼리 잘 상의해 봐요~

→ 잘 키우실 수도 있으니까

까

가족들끼리 잘 상의해 봐요~

[지침2] <우리말샘> 미등재어(외래어, 신어 등)의 처리

1. 미등재어의 처리

미등재어는 <우리말샘>에 등재되지 않은 다음과 같은 예를 말한다.

예1) 단어의 경우: 존맛탱, 팬아저

예2) 구의 경우: 코로나 블루

미등재어의 경우 OoV(Out of Vocabulary) 목록을 작성해야 하는데, 이는 교정 작업 완료 후 기계적인 처리를 거쳐 확보한다. 저빈도 미등재어의 경우, 기계 처리에서 누락될 가능성도 있으므로 교정 작업 중에 OoV 칼럼에 기록한다.

2. 미등재어의 범위

미등재어의 범위는 비교적 최근에 새로이 등장한 신어, 사전에 등재되지 않은 외래어/외국어 등을 비롯하여 다음 유형과 같다.

1) 신어의 예

예) 주린이, 캠린이, 코로나 블루, 비대면 강의, 공적 마스크, 낀낀세대, K방역, 배달거지, 먹튀브, 동학개미, 돈쥘, ...

- 신어(줄임말)의 예

예) 돌밥돌밥, 내돈내산, 꾸안꾸, 꾸안꾸룩, ...

- 신어(파생어)의 예

예) 꿀목소리, 핵인싸, 개간지, 개귀엽다 (꿀-, 핵-, 개- 등 생산적 접사를 포함한 어형들)

이상의 신어들은 교정하지 않는다. 특히 줄임말의 경우 풀어서 쓰지 않고 그대로 둔다.

예) 꾸안꾸 스타일 12가지야 → 꾸민 듯 안 꾸민 듯 스타일 12가지야 (x)

2) 외래어/외국어

‘웨이팅, 웨이팅 리스트’처럼 사전에 형태는 있으나 해당 의미가 없는 경우(888)도 추후 OoV 목록에 포함시킨다.

※인명이나 지명과 같은 고유명은 미등재어이지만 OoV 목록에 포함시키지 않는다.

예) 이스너, 케이윌, 임영웅, ...

3) 감탄사와 의성의태어

예) 잇힝힝, 찡챙챙, 헝, 헤헝, 으헤헤헤헝, 아고, 오웅, 호다닥...

[지침3] 개인정보 및 부적절한 표현(욕설, 혐오 표현 등)의 비식별화 처리

1. 비식별화 방법

1) 비식별화란 해당 대상을 엑셀 표에 입력하거나 작업 도구에서 해당 부분을 표시하는 것을 의미하며, 작업자가 해당 내용을 임의의 방식으로 비식별화하지 않는다.

2) 입력은 최종 교정 문장을 기준으로 한다.

3) 부적절한 표현(욕설, 혐오, 차별 등)은 조사를 떼고 입력하되, 엉겨붙은 끝이나 용언의 활용형은 어절 단위로 입력한다.

예) 정말 염병이었다 → 염병(○), 염병이었다(x)

예) 시벌노마 → 시벌노마(○), 시벌놈(x)

※ 그 외 비식별화 대상은 해당 부분만 입력한다.

예1) 양희가 곱창볶음 사줬대. → 양희(○), 양희가(x)

예2) 완이가 오늘 놀자고 해서. → 완(○), 완이(x), 완이가(x)

※ 하나의 열에서 비식별화해야 할 대상이 2개 이상인 경우에는 쉼표로 구분한다. 다만, 한 열의 문장 전체를 비식별화해야 할 경우 어절마다 쉼표를 넣지는 않아도 된다.

2. 개인정보의 비식별화

국어원 메신저 말뭉치 구축 시 다음 범주에 대해 아래와 같이 비식별화를 진행하였다. 본 사업팀에서는 이전 사업과의 유기적 연계성을 위해 본 과제의 대상 자료(메신저 및 웹 말뭉치) 전체에 대해 아래의 내용을 비식별화하기로 한다.

1) 메신저 말뭉치의 비식별화 대상은 다음과 같다.

이름(실명, 별명, 대화명, 필명 등), 온라인(아이디, 이메일 등), 각종 번호(고유 식별 번호, 전화번호, 금융 번호 등), 장소(상세 주소, 건물명 등), 출신 및 소속(학교, 직장, 부대 등) 등에 대해 철저히 비식별화한다.

2) 웹 말뭉치의 비식별화 대상은 다음과 같다.

고유 식별 번호(주민 등록 번호, 사번, 학번 등), 전화번호, 금융 번호, 상세 주소

※ 이름, 웹 주소(URL)는 비식별화하지 않는다.

3) 비식별화하지 않는 대상은 다음과 같으니 주의한다.

일반 애칭 별명, 공인 실명(유재석, 조승우 등), 만화 주인공(신노스케, 짱구, 펑수 등), 동보다 큰 단위의 주소(서초구, 중구, 대전 등), 거주지 역명(신천역, 광화문역 등), 비정기적인 방문 장소(우 소아과, 행복 마트 등), 상호명(굽네치킨, 엽떡 신촌점 등)

4) 이외의 비식별화 지침은 국립국어원(2019:33)의 ‘메신저 대화 자료 수집 및 말뭉치 구축’ 사업의 ‘비식별화 지침2’를 따른다.

3. 부적절한 표현의 비식별화

부적절한 표현은 욕설, 차별, 혐오, 성적인 표현을 이른다. 차별 및 혐오 표현에 대한 정의는 다음을 따른다.

“어떤 사람이나 어떤 집단과 관련하여 그들이 누구인가를 근거로, 달리 말하면 그들의 종교, 종족, 국적, 인종, 피부색, 혈통, 성 또는 기타 정체성 요소(identity factor)를 근거로 하여 이들을 공격하거나 경멸적이거나 차별적인 언어를 이용하는, 말, 문서 또는 행동으로 하는 모든 종류의 소통” (United Nations Strategy and Plan of Action on Hate Speech, 2019)

성별, 장애, 종교, 나이, 출신지역, 인종, 성적 지향 등을 이유로 어떤 개인 · 집단에게 1) 모욕, 비하, 멸시, 위협 또는 2) 차별 · 폭력의 선전과 선동을 함으로써 차별을 정당화 · 조장 · 강화하는 효과를 갖는 표현 (국가인권위원회, 혐오표현 리포트, 2019)

연구 등의 특수한 목적으로 욕설이나 혐오 표현에 대해 연구하고자 하는 경우가 있더라도, 원시 말뭉치에서 원문이 보존되므로, 일반 사용자를 고려하여 아래와 같은 유형에 대해 비식별화한다.

1) 욕설의 비식별화

ㄱ. 다음과 같은 욕설(밈줄)은 비식별화한다.

예1) 아 씨, 기억이 안 나. ㅋㅋㅋㅋㅋㅋ

예2) 아쉽다고 전해줘 오즈르

ㄴ. 비속한 표현이기는 하나, 강조의 의미를 갖는 다음과 같은 경우에는 비식별화하지 않는다.

예) 진짜 연출력 미쳤다. / 미친 연기력

※ ‘미치다’+ ‘사람 명사(놈, 년, 새끼, 자식, 부장 등)’와 함께 쓰일 경우 비식별화하고, ‘미친’+ ‘연기력, 날씨, 가창력’ 등과 결합할 경우에는 비식별화하지 않는다.

ㄷ. 비속한 표현이지만 욕설이라고 보기 어려운 표현은 비식별화 대상으로 삼지 않는다.

예) 존맛, 존맛탱, 개존맛, 존예, 존네, 대존맛, 까먹다, 개좋다, 개꿀, 찼다 등

2) 차별 및 혐오 표현의 비식별화

ㄱ. 차별 및 혐오 표현은 맥락을 고려하여 비식별화 여부를 결정하되, ‘성별, 인종, 국적, 종교’ 등에 대한 표현을 포함한다. 단, 차별 및 혐오 표현의 판단 여부는 맥락을 고려할 필요가 있는데, 다음과 같은 경우는 비식별화 대상이 아니다.

예) 사실 제 아내가 중국사람이라.

오! 중국음식을 꽤 많이 먹는 편이에요 ㅋㅋㅋ

ㄴ. 비식별화하는 차별 및 혐오 표현(밈줄)의 예는 다음과 같다.

예)

사우디 위험하다.

종교 경찰 있고

즈르보수적 사회다.

3) 특정 대상에 대한 부정적 평가 표현의 비식별화

공인이나 기관의 경우에는 비식별화 대상이 아니지만, 부정적인 내용이 포함될 경우에는 해당 대상을 비식별화한다. 상품, 상호, 영화명 등은 비식별화하지 않는다.

예) 겨울왕국도 봤어?

안 볼 예정.

보려고 했는데 엘사 보고 안 봄.

끔찍해, 진짜.

아기들 진짜 안 좋은 것만 빨리 흡수해서

벌써 엘사 화장 따라 하고 엘사 머리 따라 하는 거 보고

진짜 유해하다고 생각했어.

→ 겨울왕국과 같은, 영화에 대한 부정적 평가는 비식별화하지 않는다.

4) 성적인 표현에 대한 비식별화

: 성적 표현은 비식별화한다.

4. 대화문의 상당 부분에 대한 비식별화가 필요하다고 판단되는 경우

비식별화해야 할 부분이 광범위하여, 대화문의 상당 부분 또는 해당 대화 전체에 대한 비식별화가 필요한 경우에는 검토자에게 알리고 전체 공동 연구진 회의에서 삭제 여부를 결정한 후 발주 기관과 협의한다.

[지침4] 특수 표현의 유형 분류와 유형별 처리

온라인상에서 사용되는 특수 표현에는 자소형 이모티콘, 감탄사나 응답 표현 대체어, 내용어의 초성 연쇄 등이 있다. 이들은 다음의 2가지로 나누어 처리한다. 우선, 특수 표현 가운데에 원래 형태에 이견이 없이 확실한 내용어의 경우에 원래 형태를 복원한다. 원래 형태를 복원하는 특수 표현은 별도의 목록으로 관리한다. 다음으로 원래 형태를 확정하기 어렵거나 복원한다 하더라도 우리말샘에 등재되지 않은 표현은 복원하지 않는다. 이러한 유형도 목록을 만들어 둘 필요가 있는데, 특수한 어절 구성을 보이므로 자동으로 목록을 만들어 관리할 수 있다.

1. 자소형 이모티콘의 교정

1) 자소형 이모티콘은 자음이나 모음의 연쇄를 이용한 도상을 통해 인간의 표정을 흉내 내어 문장에 동반하는 감정을 드러내는 역할을 하고, 변형이 다양하게 이루어지므로 교정하지 않는 것을 원칙으로 한다.

예) ^^, ^^;;, ^~~, ^^~, >_<, ㅇㅅㅇ, ^.^, ^0^, -_- , ㄴㅈ, ㅋㅋ, ㅠ, ㅈㅈ, ㅈ, ㅈㅈㅈㅈ, ㄴㅈ, ㅈㅈㅈ; , @@@@.

2) 형태를 교정하지 않되, 문자열에 붙어 있는 이모티콘은 한 칸을 띄워서 수정한다.

예) 참 조아^0^ → 참 좋아. ^0^

개꿀이지>_< → 개꿀이지. >_<

ㅈㅈ미안해요 → ㅈㅈ 미안해요. ㅈㅈㅈㅈ

3) 자소형 이모티콘이 문자열과 문장부호 사이에 나오면, 이모티콘 앞뒤로 한 칸 띄워서 수정한다.

예) 눈치는 대리님이^^..... → 눈치는 대리님이... ^^

name1님^^~ 반갑습니다 → name1 님~, ^^ 반갑습니다.

2. 감탄사나 응답표현 기능을 하는 자소형 표현의 교정

1) 자소의 일부만을 사용하여 감탄사나 응답 표현을 대체하는 경우, 내용어 대치가 가능하여 대응쌍을 명백히 줄 수 있는 사례와 그렇지 않은 사례로 구분하여 처리한다.

예1) ㅇㅇ, ㅋㅋㅋㅋㅋㅋ, ㅎㅎㅎ (교정하지 않음.)

이들 뒤에는 별도의 문장부호를 추가하지 않는다.

ㅇㅇ.(x), ㅇㅇ.(x), ㅎㅎㅎ.(x)

예2) ㅇㅋ → 오키

ㅎㅇ → 하이

ㅎㅇㅎㅇ → 하이하이

ㅇㅈ → 인정

ㅁㅈ → 맞아. (용언의 활용형이므로 마침표 부여)

ㅁㅈㅁㅈ → 맞아, 맞아. (용언의 활용형이 중복해서 사용될 경우 쉼표와 마침표 부여)

ㄹㅇ → 레알 (<우리말샘> 등재어 '레알')

2) 문자열에 붙어 있는 자소형 표현은 한 칸을 띄워서 수정하고, 필요한 경우 문장부호를 추가하고 한 칸을 띄어 수정한다.

예1) 역시ㅋ → 역시 ㅋ

ㅎㅎㅎㅎ그래도 어제 → ㅎㅎㅎㅎ 그래도 어제

ㅎㅎㅎㅎ어머니아버지께 → ㅎㅎㅎㅎ 어머니, 아버지께

ㅋㅋㅋㅋ혼나? → ㅋㅋㅋㅋ 혼나?

예2) 아ㅋㅋ큰거면 → 아, ㅋㅋ 큰 거면

아이고ㅎㅎㅎ → 아이고. ㅎㅎㅎ

꽤있어서ㅋㅋ → 꽤 있어서. ㅋㅋ

3. 기타 특수 표현의 처리

1) 숫자가 문자열과 결합한 경우에 우리말샘에 등재된 어휘로 대체한다.

예) 1도 → 하나도

4가지 → 싸가지

2) 자소의 일부만을 사용하여 내용어를 대체하는 경우, 내용어 대치가 가능하여 대응쌍을 명백히 줄 수 있는 사례에는 원래 형태를 복원시킨다.

예) ㄱㅈ은듯 → 괜찮은 듯.

ㅅㄱ하셈 → 수고하세요.

3) 신어를 알파벳으로 표기한 것은 그대로 둔다.

예) JMT(존맛탱 뜻으로 쓴 것)

4. 야민정음의 처리

이들이 두루 쓰이어 언중 사이에 단어로 정착이 되면 사전에 등재될 수 있다. 현재 많

은 목록이 <우리말샘>에 등재가 되어 있는데, <우리말샘>에 등재된 어형은 교정하지 않는다.

예) 땡땡이, 머머리, 땡곡, 땡반, 커엽다

5. 언어 유희의 처리

아래와 같은 언어 유희는 교정하지 않는다.

예) 망실(‘실망’의 의미로), 아깝지 않다람쥐.

[지침5] 방언형의 처리

1. 어문규범과 <우리말샘>의 기준

해당 방언형이 <우리말샘>에 등재된 방언형과 형태 의미 면에서 정확하게 일치할 경우, <우리말샘>에서 제시한 규범 표기에 따라 교정한다.

※ <우리말샘>에 미등재된 어형 또는 방언으로 기재되어 있으나 규범 표기가 제시되어 있지 않은 경우에는 표준어형을 추측해 교정하지 않고, “?의미불명” 처리 후 그냥 둔다.

2. 방언 종결 어미의 처리

대화 상황에서 종결 어미는 대화 맥락이나 어감, 화자·청자의 관계 등에 따라 어울리는 표준형이 달라질 수 있으므로 맥락을 고려하여 교정하여야 한다. 따라서 아래 표에서 볼 수 있듯이, 하나의 방언형 종결 어미가 둘 이상의 교정 결과를 가질 수 있다.

- 동남 방언 의문형 종결어미 ‘-나’ → 최대한 원형을 살리되 맥락에 따라 ‘-어/어요’를 우선 적용해 맥락을 고려해 교정함. 아래 예시 참고.

- 어미의 경우, 방언형이 표준형 하나로 대응되지 않는 경우가 많으므로, 맥락을 고려하여 적절한 표준어형으로 교정함.

발화 ID	화자	원 문장	최종 교정 문장	비고
MDRW1 90000002 2.1.1.3	1	우리 점심때 영화 한 편 보 고 저녁에 고 기 먹을라하 는데 엄마는 어 떻노?	우리 점심때 영화 한 편 보고 저녁에 고기 먹 으려 하는데 엄마는 어 때? 우리 점심때 영화 한 편 보고 저녁에 고기 먹 으려 하는데 엄마는 어 떻나?	‘어떻다’ 형용사, 의문형 종결 어미 ‘-나’는 동사에 결합함 표준어형 하나에 1:1 대응 이 되지 않으므로 문맥에 따라 교정형을 선택하되, ‘- 어/어요’형을 우선으로 함.
MDRW1 90000002	2	근데 영화 볼만한거 있더	1. 근데 영화 볼 만한 거 있더나?	맥락에 따라 교정형 선택

2.1.1.5		나	2. 근데 영화 볼 만한 거 있어?	
MDRW1 90000002 2.1.1.19	2	아들은 저 영화는 별로인 거 같나	1. 아들은 저 영화는 별로인 거 같나? 2. 아들은 저 영화는 별로인 거 같아?	'갈니/갈아?'와 같이 맥락 을 고려해 교정

방언형이 여러 개의 표준어에 대응될 경우, 형태적으로 가장 가까운 표준형을 선택하여 교정하며, 맥락에 따라 동일한 방언형도 둘 이상의 교정 결과를 가질 수 있다.

올바른 교정	잘못된 교정
지금 머라카노? → 뭐라고 하니?	지금 머라카노? → 뭐라는 거니?
왔더나→왔어?	왔더나→왔던?

3. 방언형으로, 음운 변용의 처리

음운이 축약, 교체, 생략, 탈락된 경우, <우리말샘>을 기준으로 표준어형으로 교정한다.

원 표현	교정형
너무 많이 지난 것 같애.	너무 많이 지난 것 같아.
그런 생각은 아예 하지를 말어.	그런 생각은 아예 하지를 말아.
아직도 안 돌아갔다구?	아직도 안 돌아갔다고?
전부 수매르 하꺼인대	전부 수매를 할 것인데.
커능기	하는 것이

4. 방언형에서 띄어쓰기의 처리

방언형에서는 띄어쓰기가 무시되었더라도 표준어형으로 교정 시 어문규범에 준하여 띄어 쓴다.

원 표현	교정형
뭐라카노?	뭐라고 하니?

5. 의미 불명 방언의 처리

방언의 의미를 정확하게 해석하지 못해 정확한 교정형을 판단하기 어려운 경우에는 따

로 표시한 후, 검토자와 논의하여 ‘?의미불명’으로 처리한 후, 그냥 둔다. 이렇게 교정한 형태는 방언 교정 지침의 참고 목록으로 제공한다.

예) 누구한테고 한약은 너가 산걸로 다른애들이 서운해 할까봐 돈들도 조금씩 주었는데

→ 누구한테든 한약은 네가 산 거로. 다른 애들이 서운해 할까 봐 돈들도 조금씩 주었는데.

6. 미등재어 방언의 처리

1) 교정하지 않는 경우

<우리말샘>에서 방언임을 확인할 수 없는 심증적으로 방언인 경우, 해당 표현은 미등재어로 두고 교정하지 않는다.

예) 오늘 같은 날엔 패딩...에 찌부대서

→ ‘찌부되다’는 방언으로 사용되나 <우리말샘>의 미등재어로 보고 교정하지 않음.

2) 교정하는 경우

ㄱ. -ㅁ서 → -면서 (‘-면서’의 방언(경남))

예) 맨날 얻어드심서 → 맨날 얻어드시면서

ㄴ. 디지다 → 죽겠다 (‘뉘지다’의 방언(경기, 경남))

예) 디지겠네 → 죽겠네.

ㄷ. 뜨시다 → 따뜻하다 (‘따뜻하다’의 방언(강원, 경상).)

예) 뜨신 물에 → 따뜻한 물에

ㄹ. 아싸리 → 차라리 (‘차라리’의 방언(경상))

예) 아싸리 다 줘 버려. → 차라리 다 줘 버려.

8. 고빈도 구어체 방언의 처리

방언 교정의 기본 지침은 우리말샘의 방언형과 정확하게 일치되는 형태와 의미일 경우, 제시된 규범 표기를 따른다는 것이다. 그렇지 않고 광범위하게 나타나는, 고빈도 방언형으로 판단될 경우, 아래 기준에 따라 비교정 구어체 통용 방언형으로 판단하고, 해당 방언형을 목록에 등재한 후 교정하지 않는다.

1) <우리말샘>에 방언으로 등재되어 있으나, 지역과 무관하게 폭넓게 사용되는 방언형을 ‘구어체 말뭉치 통용 방언’으로 지칭하고, 아래와 같은 기준에 따라 교정 여부를 판단한다.

- 비교정 ‘구어체 말뭉치 통용 방언’의 판단 기준

(1) 기존의 비교정 구어체 말뭉치 통용 방언형인 경우(아래 예시 참고)

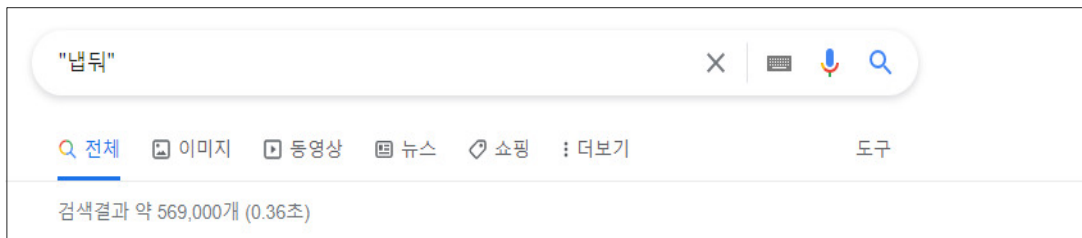
예) 이케, 여튼, 겁나, 겁내, 글고, 달달하다, 맛탱이, 일케, 글케(‘그렇게’의 줄임말로 쓰인 경우), 냅두다(냅둬…), 저번주, 저저번주, 너네

(2) 기존 비교정 방언형과 유사한 결합형인 경우

예) 저번주 → 저저번주

(3) 구글에 해당 어형을 “ ” 안에 넣어 검색해, 검색 결과 같은 의미로 해당 어형이 만 개 이상 검색되는 경우(문맥을 고려함)

예) “이케”의 경우, 글케 등과 인명, 지명이 함께 검색되어 나옴. 이 경우에는 곡용이나 활용형을 고려해 검색함. 예를 들어 “이케”는 “이케해” 7,050개, “이케할” 1,520개, “이케하고” 5,670개와 같이 활용형으로 검색되는 어형 수가 모두 만 개 이상이므로 ‘구어체 말뭉치 통용 방언’으로 판단한다.



<그림 14> 고빈도 통용 방언 검색 예시

구어체 말뭉치 통용 방언으로 판단하고 해당 방언형을 교정하지 않는 경우, 구글 시트를 이용해 비교정 대상에 해당하는 방언형과 근거를 작성한다.

※ 아래 예시와 같이 비교정 대상 구어체 방언형을 확보한다.



얼탱 없다 → 어처구니 없다.

예5) 제주 방언(마지막 페이지의 참고도 확인해주시오.)

-멘: 화계와 시제에 따라 ~지? ~해? 체로 교정

나는 왜 그거 안주멘?: 나는 왜 그거 안 주지?

-클: -(으)르게/해, 해요 등에 대응됩니다. 시제와 화계에 맞게 교정합니다.

예6) 그림 좀 빠칠클: 그림 좀 빠치는데?

비싸클: 비싸.

-연: -여서에 대응

예7) 듀랑고가 최고연: 듀랑고가 최고여서

-헨: -했어

예8) 예약헨: 예약했어.

-겐: -졌어

예9) 예약해야겐: 예약해야졌어.

예10) 따시다: 따뜻하다로 교정, 우째: 어찌로 교정

예11) 땡기다: '당기다'로 교정

예12) 썰다: '세다'로 교정

예13) 쫌: '좀'으로 교정

10. 방언처럼 보이지만 방언이 아닌 경우,

아/어 쌍다: '해쌍다'와 같이 경상 방언으로 표제어가 검색되지만, 앞말이 뜻하는 행동을 반복하거나 그 행동의 정도가 심함을 나타낼 때는 표준어 보조 동사이므로 교정하지 않는다.

[지침6] 문장부호의 처리

1. 기본 지침

한글 맞춤법 문장부호 사용 규정에 따른다. 규정에 어긋나는 위치에 나타나는 문장부호는 삭제할 수 있으며, 복수의 문장부호가 나타나면 하나로 줄인다.

예) ??????????나도 케이크!!!!!!! → 나도 케이크!

다른거 쓰기 뭐해서?? → 다른 거 쓰기 뭐해서?

술도 먹고!!!! → 술도 먹고!

그럴 수도?/ → 그릴 수도?

2. 문장 종결 부호의 부여

발화 단위의 마지막에 나타날 수 있는 문장부호는 마침표, 물음표, 느낌표로 한정한다. 발화 단위가 종결어미나 연결어미로 끝날 때 문장부호를 붙이고, 발화 단위 내에 두 문장이 연속될 때 문장의 경계를 문장부호로 구분한다.

1) 문장부호가 누락된 발화 단위에는 문장부호를 추가한다. 맥락에 따라 문장부호를 추가하되 우선적으로 마침표를 추가한다. 원 발화에서 문장부호가 포함된 것은 바꾸지 않고, 문장부호의 유형은 아래와 같이 맥락에 따라 추가한다.

예1) 땡기네요 → 당기네요.

마자여 → 맞아요.

안녕하세요 → 안녕하세요?

안녕 → 안녕?

안녕하세요~! → 안녕하세요~!

예2) 안녕 뭐해?? → 안녕, 뭐 해?

2) 문장 중간에 종결어미가 도치되어 나타난 경우, 종결어미 뒤에 쉼표를 넣고 문장의 맨 마지막에 마침표를 넣는다.

예) 오래 됐어 바뀐지 → 오래 됐어, 바뀐지.

근데 원작파괴야 드라마. → 아, 근데 원작 파괴야, 드라마.

ㅋ오래가네요 생각보다 → ㅋ 오래 가네요, 생각보다.

3) 한 명의 발화자의 발화가 이어지고 있는 중간에 다른 발화자가 끼어들어 말차례가 바뀐 채로 문장이 이어진 경우에 맥락을 고려하여 발화 단위 마지막이라도 문장부호를 생

략할 수 있다.

예) 발화자1: 나는

발화자2: 맞아.

발화자1: 그랬어.

→ 발화자1의 말이 발화자2의 말로 말 차례가 구분된 경우이므로, 발화자1의 나는 뒤에는 문장부호를 부여하지 않고, 종결 발화 단위인 ‘그랬어’ 뒤에만 마침표를 찍는다.

4) 종결어미로 문장이 끝나면 마침표를 추가하고, 연결어미 이후에 문장이 이어지면 맥락에 따라 쉼표를 추가할 수 있다. 단 자소형 이모티콘 등 특수 표현이 단독으로 나타나면 문장부호를 추가하지 않는다.

예) 샀지 → 샀지.

같은데 → 같은데.

맞아 → 맞아.

ㅌㅌㅌ → ㅌㅌㅌ

ㅋㅋㅋㅋㅋㅋ → ㅋㅋㅋㅋㅋㅋ

3. 어형 및 맥락에 따른 문장부호의 부여

1) [특수 표현] 발화 단위의 마지막에 문장부호 대신에 특수 표현이 나타나면 발화 단위 마지막에 문장부호를 첨가하고 한 칸을 뒀 후에 특수 표현을 남긴다.

예) 응ㅋㅋ → 응, ㅋㅋ

맞아요ㅍㅍ → 맞아요. ㅍㅍ

여긴 안왔어ㅍ → 여긴 안 왔어. ㅍ

2) [감탄사 상당 표현] “어문 규범의 쉼표 규정 (7) 부르거나 대답하는 말 뒤에 쓴다”에 따라 ‘응, 네, 아니...’ 등의 뒤에는 쉼표를 찍고, 인사말 등 감탄사 상당 표현에 대해 쉼표나 물음표 등을 부여한다.

예1) 응 내가 봤어. → 응, 내가 봤어.

웅 난 폰게임안해! → 웅, 난 폰게임 안 해!

웅웅 매일 챙겨와 ㅋㅋ → 웅웅, 매일 챙겨와. ㅋㅋ

우웅 나 집에서 자주해먹어 → 우웅, 나 집에서 자주 해 먹어.

예2) 오호라 그런 거였군 → 오호라, 그런 거였군.

아이고 오타가났네 → 아이고, 오타가 났네.

오호 물까지 챙겨주는 거야? → 오호, 물까지 챙겨 주는 거야?

안녕 뭐해

→ 안녕? 뭐 해?

※ ‘맞아, 좋아, 알았어’ 등과 같이 용언의 활용형이 응답 표현과 같은 기능을 가지는 다 음과 같은 경우, 마침표나 쉼표 등으로 처리한다.

예1) 알았어. 내일 봐.(○)/ 알았어, 내일 봐.(○)

예2) 맞아. 나도 그리 생각.(○)/맞아, 나도 그리 생각.(○)

3) [인용] 인용임이 분명하고 문장부호가 있는 경우에 한해 큰따옴표를 넣어 인용임을 분명히 할 수 있다..

예) 그래서 너도 가입해! 했는데 → “그래서 너도 가입해!” 했는데

공짜 도시락이면 → 공짜 도시락이면

감사합니다. 하고 먹어야지. → “감사합니다.” 하고 먹어야지.

아가가 엄마가 없겠지? 하겠다. → 아가가 “엄마가 없겠지?” 하겠다.

4) [자소형 발화 단위] 발화 단위가 자소형 등으로 이루어진 경우 마침표를 추가하지 않 으며, 특수 표현이 자소형이 내용어로 인정이 되어 복원되는 경우에 마침표를 추가하는 것은 가능하다.

예) ㅇㄷㄷ → ㅇㄷㄷ

ㅇㅇ → ㅇㅇ

ㅇㅈ → 인정./인정

ㅁㅈ → 맞아./맞아

※ 단, 발화 중간에 특수 표현이 있고, 단어의 복원이 가능한 경우, 특수 표현은 소속 문자열에서 한 칸 띄운 자리로 이동시킨다.

예) 아, 맞ㄴ네. → 아, 맞네. ㄴ

복원하지 않는 자소형 감탄사가 문두나 문중에 나올 경우에는 따로 문장부호를 부여하 지 않고 띄어쓰기만 교정한다(아래 6의 지침과 다르니 주의.).

예) ㅇㅇ그렇구나 → ㅇㅇ 그렇구나.

아 ㄴㄴ 싫어. → 아, ㄴㄴ 싫어.

5) [의미 불명어] 의미 불명어의 경우에는 문장부호를 부여하지 않는다.

만떡썰, 닥병, 력부섬야, 학약 등과 같은 표현은 의미를 추정할 수 없으므로 따로 교정 하지 않고, 문장부호도 부여하지 않는다. 의미 불명어는 OoV 칼럼에 입력하여 기록한다.

예) 만떡챵, 닉병, 력부섬야, 학약 등

4. 문장부호의 삭제와 교정

1) 아래 예시와 같이 물결표만 나타난 경우는 발화 단위 마지막에 나타나야 할 문장부호를 추가하며 말줄임표의 역할을 하는 복수의 마침표는 규정에 맞게 세 개의 마침표로 교정한다.

예) 시퍼요~~	→	싫어요~~.
안녕하세요~~~	→	안녕하세요~~~?
거예요~	→	거예요~?
아....	→	맞아...
그래,,, 좋아.	→	그래, 좋아.
헤랑?이	→	헤랑(?)이

2) 발화 단위의 마지막에 종류가 다른 문장부호가 나타날 경우, 맥락에 적합한 하나의 문장부호만 남긴다.

예) 내가해야해?! (의문문이 확실한 경우)	→	내가 해야 해?
20분짜리니까...?	→	20분짜리니까?

3) 어절 단위 내에 나타나는 말줄임표는 삭제한다.

예) 내...가	→	내가
사랑...해	→	사랑해

5. 물결표의 사용

물결표는 문장부호 규정과 상관 없이 유지시킨다.

1) 단어 중간에 나오거나 복수의 물결표가 나와도 물결표는 유지시킨다.

예) 날씬~하고	→	날씬~하고
시퍼요~~	→	싫어요~~.
청국장 된장찌개~~~못먹겠당~~	→	청국장, 된장찌개~~~ 못 먹겠다~~.
안녕하세요~~~	→	안녕하세요~~~?

2) 물결표는 비분절 요소로 이해하여 문장부호보다 선행하여 표시한다.

예) 맞이하면 되겠네!!!!~ → 맞이하면 되겠네~!

6. 기타 부호

1) #이 포함된 경우

예) 물어보는 거지~# → 물어보는 거지~. #

→ 물결표 다음에 마침표 넣고 # 앞에 한칸 띄운다.

2) ^~~

예) 섭섭하네^~~ → 섭섭하네. ^~~

→ 섭섭하네 다음에 마침표 넣고, ^~~를 하나의 이모티콘으로 본다.

3) ~.~

예) 그랬어 ~.~ → 그랬어. ~.~

Ⅲ. 웹 말뭉치 작업 지침

[지침1] 해시태그(#이하 내용)에 대한 지침

(1) 해시태그, 즉 ‘#’부터 첫 번째 빈칸 이전까지의 내용(회색 부분)은 교정 대상이 아니다.

- 비록 오타가 있어도 교정하지 않는다.

예) #맛이찌

- 단어 이상의 단위를 포함하는 경우에도 교정하지 않는다.

예) #그놈이온다

- 문장 종결형으로 끝이 나는 경우에도 수정하지 않고 그냥 둔다. (마침표 붙이지 않음)

예) #뉴스킨제품 ♡

#감사합니다 ♡

원 문장	교정 문장
#온습도계 #모드 #옥아텐 #스우트 #온습도 #우주선램프s3 #스면트 #온도계 #습도계 #육아장비 #출산용품 #출산선물 #육아용품 #막이찌 #육아 #육아만 #육아소통 #육아스타그램	교정하지 않음
#아코이 #우주선램프 하나로.. 스마트폰과 연동해서 수유등을 .. 밝히고 끄고 조절하고 알람설정도 있어.. 약이나 수유시간도 체크해둘 수 있어요.. 폰으로 #온습도 불쾌지수까지 보고.. 이렇게 깨알 육아 장비빨이라는 거 인정..	#아코이 #우주선램프 하나로.. 스마트폰과 연동해서 수유 등을.. 밝히고 끄고 조절하고 알람 설정도 있어.. 약이나 수유 시간도 체크해둘 수 있어요.. 폰으로 #온습도 불쾌지수까지 보고.. 이런 게 깨알 육아 장비빨이라는 거 인정..
#뷰티스타그램 요즘 제일 신경쓰이는것이 바로 주름! 그래서 얼굴주름은 #보나메두사 #페이스웨이브딜리트 로 관리해요~~ 바르는 즉시 리프팅되어 주름이 채워지고 금 오팔 캐비어 등의 고급성분들이 주름개선은 물론 자연스러운 볼륨효과가 그웬잇! 나이가 가장 먼저보이는 목은 넥라인전용 보나메두사 #넥웨이브딜리트 로! 피쉬콜라겐 씨실트 캐비어추출물이 목피부를 탄력있고 밀도높게 관리해줘요~ 말대꾸 참 잘하는 스타일 ☺ ☺ ☺	#뷰티스타그램 요즘 제일 신경 쓰이는 것이 바로 주름! 그래서 얼굴 주름은 #보나메두사 #페이스웨이브딜리트 로 관리해요~~ 바르는 즉시 리프팅되어 주름이 채워지고 금 오팔 캐비어 등의 고급 성분들이 주름 개선은 물론 자연스러운 볼륨 효과가 그웬잇! 나이가 가장 먼저 보이는 목은 넥라인 전용 보나메두사 #넥웨이브딜리트로! 피시 콜라겐 씨실트 캐비어 추출물이 목 피부를 탄력 있고 밀도 높게 관리해줘요~ 말대꾸 참 잘하는 스타일 ☺ ☺ ☺

<그림 17> 웹 말뭉치 원 문장:교정 문장 예시

- 다음과 같은 예에서 작업 구간이 특히 헛갈리므로 주의한다.(밑줄 부분만 작업 대상이므로 주의)

예1) 비오는 날 #흐린날 #비온다 #그놈이온다 #왔다 역광과 흐린날의 콜라보 #현준 #용규세끼 출근전 라이딩 #맛스타그램 #먹스타그램 #남남□ #자전거 #산악라이딩 #라이딩 #마라톤 #싸이클 #프리라이딩 물 웅덩이 호우 #전속사진사 용규세끼 이제서야보는 #곤지암

예2) #토끼정 당신과 함께 머무는 곳 #토끼정 #함밤스테끼 #함박 #밤 #양파튀김 아쉬웠던 #크림카레우동 치킨반반무많이? 아니죠 #숯불구이반반 #ㅋ #할리스커피 강하며 섬세한 보컬리스트 #쥬아나 #juana #너를그리다 #팬미팅 #ㅋ #□□□□□ 싸인해주세요... #용규세끼 #맛스타그램 #먹스타그램 #남남□

(2) #명사(명사구) 형식으로 된 해시태그 다음에 있는 조사나 접사는 교정하지 않는다.
(붙여 쓰지 않음)

예) #아아 가 진리죠~

#아이스커피 가 완성돼요~

진한 #우유아이스크림 과

신메뉴 #꿀호떡썩절편 과

#피치미업 #톤업크림 으로

#아이허브#직구 할 때 산

#무드등 이

[지침2] 외래어/외국어 관련 지침

1) 외래어는 <우리말샘> 등재어에 한해 교정한다.

규범 표기가 미확정인 경우라도 일관성을 위해 현재 등재된 형태를 따른다.

원 문장	교정 문장
바벨 <u>스쿼트</u> 20K	바벨 <u>스쿼트</u> 20K
<u>랫풀다운</u> /바벨 <u>데드리프트</u>	<u>랫 풀다운</u> /바벨 <u>데드 리프트</u>
<u>프로틴</u> 등의 단백질 <u>쉐이크</u>	<u>프로틴</u> 등의 단백질 <u>셰이크</u>
<u>주스 디톡스</u> 니 하는 것들만 먹어서는	<u>주스 디톡스</u> 니 하는 것들만 먹어서는
<u>얼리버드 티켓오픈</u>	<u>얼리 버드 티켓 오픈</u>
<u>에그베네딕트</u>	<u>에그 베네딕트</u>
<u>립플럼퍼</u>	<u>립 플럼퍼</u>
시작된 어색한 <u>포토타임</u> ☹	시작된 어색한 <u>포토 타임</u> ☹
일반히알루론산+ <u>슈퍼히알루론산</u> +초미세히알루론산! 3가지 <u>히알루론산의황금율</u> 로보습력도 UP! 흡수력도UP!	일반 <u>히알루론산</u> + <u>슈퍼</u> 히알루론산 + 초미세 히알루론산! 3가지 <u>히알루론산의 황금률</u> 로 보습력도 UP! 흡수력도 UP!

<그림 18> 외래어/외국어 관련 원 문장:교정 문장 예시

2) 메신저와 마찬가지로 고유명은 수정하지 않는다.

예) 너무 보챌 때는 글라미엘 하고

3) 널리 사용되는 비규범 표기 목록은 다음과 같으니, 이를 유의하여 교정한다.

비규범형	규범형
멘탈	멘탈
커리	카레
바디 케어	보디 케어
프레쉬	프레시
잉글리쉬	잉글리시
케이	케이크
슈퍼	슈퍼
슈퍼마켓	슈퍼마켓
샵	숍
마사지샵	마사지숍
초코렛	초콜릿
오바, 오버하다	오버, 오버하다
개오바	개오버

<표 7> 널리 통용되는 비규범 표기에 대한 규범형 목록

4) 미등재, 비규범형이지만 널리 사용되는 것(구글 검색 10만 이상)은 관용을 인정하여 교정하지 않으며, 그 목록은 다음과 같다.

■ 관용적인 표기를 인정하는 대상

그웨잇(메신저 60만에서 9회 출현) … 그웨잇(50만..구글)

스튜핏... 스튜핏(23만..구글)

맙큐

프로틴

[지침3] 기호, 문장부호 관련

1) 이모티콘과 문자 사이는 앞뒤를 한 칸씩 띄어서 쓴다.

단, 조사가 개입되는 경우는 띄지 않는다.

예) 딸기□랑 바나나□ 사서 집으로...

원 문장	교정 문장
★ 기존에 봐왔던	★ 기존에 봐왔던
☹️ 뒷글은 브런치에서	☹️ 뒷글은 브런치에서
#퀵마마마켓 리뉴얼되어 향 가득한 곳. ♣️	#퀵마마마켓 리뉴얼되어 향 가득한 곳. ♣️
눈 돌아가, 코 킁킁거리, 좋아. 좋아. 🌸	눈 돌아가, 코 킁킁거리, 좋아, 좋아. 🌸
나두 추석 선물 받아따~~예이~~ 怡舟... 너 근데 글씨 못 쓴다... 못알아보게써	나두 추석 선물 받았따~~. 예이~~. 怡舟... 너 근데 글씨 못 쓴다... 못 알아 보
😊 마음이 예쁘니까 괜찮아♡송편대신 월병 먹구 저건 무슨 맥주과자라는데 꿀맛... 🍷 하루에 하나씩만 먹자	😊 마음이 예쁘니까 괜찮아. ♡ 송편 대신 월병 먹고 저건 무슨 맥주 과자라는데 꿀맛... 🍷 하루에 하나씩만 먹자.

<그림 19> 기호, 문장부호 관련 원 문장:교정 문장 예시

2) ‘+’와 ‘=’는 앞뒤를 한 칸 띄어서 쓴다.

예)

원 문장	교정 문장
최대자외선차단지수(SPF50+,PA+++)+ 보습효과+미백.주름개선+피부진정+메이크업 베이스 이모든게 한번에 가능하다고□	최대 자외선 차단 지수(SPF50+, PA+++)+ 보습효과 + 미백, 주름 개선 + 피부 진정 + 메이크업 베이스. 이 모든 게 한번에 가능하다고 □

3) 괄호는 수정하지 않고 그대로 둔다.

다만, 다음 예에서 말줄임표 다음에는 한 칸 띄어 쓴다.

예) □ 계란은 요즘(...아니 이미 예전인가 □) → □ 계란은 요즘(... 아니 이미 예전인가? □)

4) 부호의 수정

부호가 일반적인 문장에서와 달리 사용된 경우나 누락된 경우는 수정한다.

1) 위 (2)의 예 중 다음은 마침표가 잘못 사용된 것이므로 쉼표로 수정한다.

예) 미백.주름개선 → 미백, 주름 개선

2) 위 (2)의 예 중 다음은 문장이 종결되는 것으로 보여 마침표를 넣는다.

예) 메이크업베이스 이모든게 → 메이크업 베이스. 이 모든 게

3) 빗금은 두 개 이상의 어구를 나열할 때 사용할 수 있으므로 수정하지 않는다.

예) 둘째 임신/출산으로 인해,(○)

5) 말차레가 없는 경우에는, 종결이 분명한 경우(㉞, ㉟)에 한해 마침표를 추가한다.

예) ㉠ 처음 사용해 본 #데싱디바

㉡ 생각했던 것보다 편하고 이쁘네. □

㉢ 붙이는 것도 간편하고 길이 조절도 간편하고

㉣ 손톱이 작은 편이라 고민했었는데

㉤ 사이즈도 다양하게 있어서

㉦ 골라쓰니 문제없군.

6) 단어 사이에 붙임표가 들어간 경우 삭제한다.

예) 크림치이-즈 → 크림치즈, 아-무리 → 아무리, 모-오닝 → 모닝, 사알-짝 → 사알짝, ...

[지침4] 기타 세부 사항

1) 다음과 같은 축약형은 수정하지 않는다.

예) 짜치계=짜장 떡볶이+치즈+계란 → 짜치계 = 짜장 떡볶이 + 치즈 + 계란

2) 어미 뒤에 붙은 ‘잉’의 처리

예) 맛있어잉 → 맛있어

3) 다음은 수정하지 않고 OoV에 추가한다.

예) 초크초크해(‘촉촉해’의 뜻): 비표준형이나 용례가 다수 확인됨

4) 신조어 ‘각(角)’의 띄어쓰기

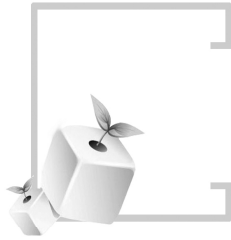
‘xx 각’의 경우 <우리말샘>의 기술에 따라 띄어 쓴다.

각(角) 「의존 명사」 「030」((명사 뒤에 쓰이거나 어미 ‘-을’ 뒤에 쓰여)) 어떤 일이 일어날 조짐이나 분위기.

예1) 마침 배도 출출한데, 야식 각이다.

예2) 시험도 끝났겠다, 놀러 갈 각이다.

예3) 꿀잠 각, 썸 각, 공주님 각, 또 각 나왔다, 트윈 룩 각, ...



제 4 장

결 론



결론

이 사업은 기구축된 국립국어원의 메신저 및 웹 말뭉치 300만 어절에 대한 맞춤법 교정 말뭉치 연구 분석 사업이다. 이 사업을 통해 구축된 맞춤법 교정 말뭉치는 공공재로서의 말뭉치, 초기 인공지능 학습 데이터로서의 말뭉치를 지향한다. 따라서 완벽한 수준의 맞춤법 교정이 아니라 형태소 분석기 등의 NLP 도구 적용이 가능한 수준의 말뭉치를 구축하고, 이를 구축하기 위한 맞춤법 교정 및 개인정보 및 혐오 표현 처리 지침과 처리 방안을 마련하는 것이 이 사업의 목적이다. 이러한 목적하에 이 사업은 아래와 같은 단계로 사업을 수행하였다.

(1) 맞춤법 교정을 위한 지침의 수립

이 사업의 대상인 웹과 메신저 말뭉치는 사용자 생성 콘텐츠(User Generated Content, UGC)로 기호나 철자의 변형을 활용한 감정 표현, 구어체, 비규범적 표현, 오자와 탈자, 혐오 차별적 표현의 윤리적 문제 등의 특성을 가지고 있다. 이 특성들은 문어·구어를 중심으로 학습된 기존의 형태소 분석기 등 NLP 도구의 적용을 어렵게 하며, 인공지능 데이터의 윤리적 활용에서도 문제를 제기한다. 본 사업에서는 메신저와 웹 말뭉치의 표기 오탈자, 비표준형, 띄어쓰기 등을 구어 전사 말뭉치 수준으로 교정한다는 목적에 따라, 이들의 언어적 특성을 연구, 분석함으로써, 맞춤법 교정 지침을 수립하였다. 교정 지침은 교정 유형별 지침으로 구성되며, 표준형과 비표준형의 판별은 <우리말샘>을 주요 기준으로 하되, 유형에 따라 별도의 지침을 수립하여 목록을 관리하였다.

(2) 개인정보 및 부적절한 표현 등에 대한 처리 방안 마련

이 사업에서는 민간에서 변환 및 호환이 용이한 공공재로서의 말뭉치를 구축을 목표로 한다. 이를 위해 웹과 메신저 말뭉치에서 나타나는 개인정보 및 부적절한 표현을 파악하고, 비식별화하는 기준과 방안을 마련하였다. 개인정보의 경우, 기존 사업과의 유기성을 고려해 국어원 메신저 말뭉치(버전 1.0) 구축 시 제시한 범주에 의거해 비식별화 방안을 마련하여 적용하였다. 또 혐오, 차별 표현 및 부적절한 표현의 범주는 메신저와 웹의 특수성을 고려하여 혐오와 차별, 욕설, 성적 표현 등의 별도 범주를 분류함으로써 비식별화 지침을 수립하고 실제 말뭉치 구축에 적용하였다.

(3) 맞춤법 교정 병렬 말뭉치의 구축

이 작업은 자동 검사기 처리를 거친 후 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식

으로 이루어진다. 또한 교정 작업의 효율화를 위해 교정 병렬 말뭉치 구축 도구인 Kronoth를 사용하였으며, (주)이르테크의 말뭉치 검증 시스템을 활용해 분석 결과의 정확도를 확보하였다. 맞춤법 교정 말뭉치의 구축은 (1) 텍스트 전처리를 통한 맞춤법 교정용 말뭉치 변환 (2) 자동 맞춤법 교정 도구를 이용한 1차 자동 교정 (3) 수작업 전수 교정 (4) 개인정보와 부적절한 표현의 비식별화 (5) 세 차례의 품질 검수 (6) JSON 구조화 (7) 최종 형식 검수의 과정으로 구축되었다.

이상의 과정을 통해 이 사업은 300만 어절의 맞춤법 교정 말뭉치를 구축하였다. 이 사업으로 구축된 말뭉치는 자연어 처리의 관점에서 인공지능 학습용 데이터로서의 수준을 만족하는 수준으로, NLP 도구의 적용이 가능한 수준이다. 특히, 메신저와 웹 언어가 가지는 특수한 성격을 고려하여 수립된 지침에 따라 구축되었다는 점에서 이 말뭉치는 향후 맞춤법 검사기의 정밀도와 정확도 향상과 기타 자연어 처리 응용 기술에의 활용이 기대되는 말뭉치이다.

참고문헌

- 김진웅(2021), 자연언어처리에서 윤리적 문제와 해결 방안, 연구방법논총 6(1), 157-180.
- 남길임(2016), 상품평 텍스트에 나타난 감성표현 연구: 감성분석과 국어학 연구의 접점, 언어과학연구, 78, 101-123.
- 남길임(2018), 웹 말뭉치를 활용한 언어 연구의 현황과 쟁점, 한국어 의미학 60, 23-49.
- 남길임, 강현아(2019), 말뭉치언어학적 관점에서 본 감성표현 추출의 쟁점 - 사용자 리뷰 말뭉치를 중심으로-, 어문론총 82, 207-236.
- 남길임, 안진산, 황은하(2020), UGC 표준형 말뭉치 구축을 위한 말뭉치언어학적 연구-유투브 댓글을 중심으로, 한말연구 57, 63-96.
- 박일섭 외(2019), 『메신저 대화 자료 수집 및 말뭉치 구축』, 국립국어원.
- 송현주(2020), 차별과 혐오 표현에 대한 국어교육 내용 연구, 제52회 2020년 국어교육학회(since1969) 전국학술발표대회 발표자료집, 166-188.
- 안의정(2018), 구어 전사 말뭉치 구축에 관한 현황과 쟁점, 언어와 문화 14, 81-101.
- 안의정(2019), 형태 분석 말뭉치 구축을 위한 한국어 구어 분석, 언어사실과 관점 47, 5-24.
- 안의정(2020), 구어 전사 말뭉치의 언어학적 주석과 활용, 동서인문학 58, 7-28.
- 안의정, 송현주, 김진웅(2020), 형태 분석을 위한 메신저 텍스트 처리 방안. 텍스트언어학 49, 27-52.
- 유현경, 황은하(2010), 병렬말뭉치 구축과 응용, 언어정보와 사전편찬 25, 5-40.
- 윤은정, 김진호, 남길임, 송현주, 옥철영, 최준, 박윤배(2018), 교육용 과학언어 연구를 위한 범용 자료로서 과학교과서 말뭉치 K-STeC(Korean Science Textbook Corpus) 구축, 한국과학교육학회지 38(4), 575-585.
- 이승현, 이준일, 정강자, 조혜인, 한상희, 홍성수(2019), 『혐오 표현(Hate Speech) 리포트』, 국가인권위원회.
- 이영희 외(2019), 『웹 말뭉치 구축 최종 보고서』, 국립국어원.
- 황은하 외(2002), Korean-Chinese Machine Translation Based on Verb Patterns, in AMTA '02 Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, 94-103.
- 황은하(2016), 말뭉치 기반 한외(韓外) 대조언어학 연구에 대한 일고찰, 어문론총 69, 39-72.

- 황은하(2016), 말뭉치에 기반한 한중 한자어의 대조분석 연구: 공기 경향성에 대한 관찰을 중심으로, *이중언어학* 64, 327~352.
- 황은하(2017), 언어간 연구를 위한 대응어 주석 말뭉치의 구축과 활용, *언어와 정보* 21(2), 137-157.
- 황은하(2021), 대조분석을 위한 말뭉치의 타당성 연구 -한중 대조분석을 중심으로-, *이중언어학* 82, 259-286.
- Al-Sa'Di, R. A., & Hamdan, J. M. (2005). "Synchronous online chat" English: Computer-mediated communication. *World Englishes*, 24(4), 409-424.
- Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C.R., Hriba L., Longhi J. & Seddah D. (2014). The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres, in Building and Annotating Corpora of Computer-Mediated Discourse, *Journal of Language Technology and Computational Linguistics*, 29(2) : 1-30.
- Chen, T. & Kan, M.(2013) Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources & Evaluation* 47, 299-335.
- Collins, L. (2019). *Corpus Linguistics for Online Communication: A Guide for Research*. Routledge.
- Crystal, D. (2006). *Language and the Internet*. Second Edition. Cambridge: Cambridge University Press.
- Egbert, J.(2017) Meaningful levels of analysis in (corpus) linguistics. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_183133.pdf
- Flint E., Ford E., Thomas O., Caines A., & Buttery P. (2017). A text normalisation system for Non-Standard english words. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 107-115.
- Guterres, A. (2019). United Nations Strategy and Plan of Action on Hate Speech. no. May, 1-5.
- Herring, S. (2014), Research: Computer-mediated communication, *Bulletin of the American Society for Information Science and Technology*. 41-44.
- Ljubešić, Nikola; Erjavec, Tomaž; Fišer, Darja(2014), Standardizing Tweets with Character-level Machine Translation, *Computational Linguistics and Intelligent Text Processing*. 164-175.

- Saito, I., Suzuki, J., Nishida, K., Sadamitsu, K., Kobashikawa, S., Masumura, R., ... & Tomita, J. (2017). Improving neural text normalization with data augmentation at character-and morphological levels. *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. 257-262.
- Eryiğit, G., & Toruno, D. (2017). Social media text normalization for Turkish. *Natural Language Engineering*, 23(6), 835.
- Schulz, S., Pauw, G. D., Clercq, O. D., Desmet, B., Hoste, V., Daelemans, W., & Macken, L. (2016). Multimodular text normalization of Dutch user-generated content. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4), 1-22.
- Sindoni, M. G. (2014). *Spoken and written discourse in online interactions: A multimodal approach*. Routledge.
- Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., & Richards, C. (2001). Normalization of non-standard words. *Computer speech & language*, 15(3), 287-333.
- Yvon, F. (2010). Rewriting the orthography of SMS messages. *Natural Language Engineering*, 16(2), 133.

<Abstract>

Analysis of the 2021 Normalized Spelling Corpus

This project aims to build a ‘Web and Instant Messenger Corpus’ of 3 million *ece/* which is morphologically analyzed, review and normalize the spelling of the raw corpus so it can be analyzed with Korean language processing tools, such as automatic translators, and build and examine a parallel corpus of normalized spelling that has kept the essence of instant messenger language. In order to achieve this, the project entails the following.

First, guidelines for the normalization of the spelling need to be established.

Second, directions for dealing with personal information and inappropriate expressions need to be prepared.

Third, a parallel corpus of normalized spelling needs to be built such that it is optimized as learning data for AI.

The above can be summarized as follows.

(1) Guidelines for the normalization of spelling

As user-generated content (UGC), the Web and Instant Messenger Corpus is characterized by colloquialism, non-normative expressions, typographical errors and omissions, sentiment expressions that are conveyed by means of symbols or spelling variations, and ethical issues regarding hate and discrimination language. These characteristics may hinder the application of NLP tools, such as the morpheme analyzer which is based on written and spoken language, and raise issues in the ethical use of AI data. In that regard, spelling guidelines were established in accordance with the objective of the project, that is, correcting typographical errors, non-standard forms, and word-spacing to the level of a speech transcription corpus after examining and

analyzing their linguistic characteristics. The correction guidelines consisted of a set of instructions for each type of corrections. In addition, while the distinction between standard and non-standard forms was made based on *Urimalsaem*, the lists were treated based on separate guidelines established for each type.

(2) The treatment of personal information and inappropriate expressions

As the project sought to build a corpus of private data that could be converted to public data to be compatible with public use, a system was established to deal with personal information and inappropriate expressions which appear in the Web and Instant Messenger Corpus. In the case of personal information, de-identification measures were designed and implemented based on the standard categories provided by the National Institute of Korean Language for the building of the Instant Messenger Corpus (version 1.0) to remain consistent with previous projects of the same nature. In the case of inappropriate language, it was analyzed based on the characteristics of instant messenger and Web language and inappropriate expressions were classified into the categories of hate expressions, discrimination expressions, profane expressions, and expressions of a sexual nature. De-identification guidelines for such expressions were established and applied to the construction of the corpus.

(3) Construction of the Normalized Spelling Parallel Corpus

The construction of the Normalized Spelling Parallel Corpus is a task that entails running an automatic orthography checker followed by manual correction of spellings and word-spacing. In addition, we used the normalization parallel corpus building tool Kronoth to increase the efficiency of the correction task and utilized the corpus checker system developed by Irtech Co. Ltd. to ensure the accuracy of the analysis results. The building steps of the Normalized Spelling Corpus are as follows: (1) Preprocessing of the text; (2) Construction of the first

version of the Normalized Spelling Corpus using automatic spell checkers; (3) Three-step verification; (4) De-identification of personal information and inappropriate expressions; (5) JSON structuring; and (6) Final verification.

Keywords: normalized corpus, instant message corpus, web corpus,
parallel corpus, automatic orthography checker

Project Director: Kilim Nam(Kyungpook National University)

<기획·연구>

국립국어원 이승재 언어정보과장

국립국어원 유희정 학예연구사

국립국어원 한송이 연구원

<사업 참여자>

사업 책임자 남길임(경북대학교 국어국문학과 교수)

사업 참여자 곽용진((주)이르테크)

안미애(경북대학교 국어국문학과 교수)

김진웅(경북대학교 국어국문학과 교수)

송현주(경북대학교 국어교육과 교수)

안의정(연세대학교 문과대학 강사)

황은하(배재대학교 국어국문·한국어교육학과 교수)

심난희(배재대학교 주시경교양대학 교수)

이후영((주)이르테크)

최지선((주)이르테크)

강신아(연세대 국어국문학과 박사 수료)

강윤희(경북대 국어교육과 박사 과정)

이갑진(경북대학교 국제교류처 강사)

백미경(경북대학교 국제교류처 강사)

강현아(경북대학교 국어국문학과 강사)

안진산(경북대학교 국어국문학과 석사 과정)

황지윤(경북대학교 국어국문학과 석사 과정)

고예린(경북대학교 국어국문학과 석사 과정)

성민규(경북대학교 국어국문학과 학부 과정)

장희선(경북대학교 국어국문학과 학부 과정)

이지혜(경북대학교 국어국문학과 학부 과정)

김수지(배재대학교 한국어교육학과 석사 과정)

정나현(배재대학교 한국어교육학과 석사 과정)

전현진(배재대학교 한국어교육학과 석사 과정)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2021년 12월 21일

발행일: 2021년 12월 21일

인 쇄: 경대디지털

※ “이 책은 국립국어원의 용역비로 수행한 ‘2021년 맞춤법 교정 말뭉치 연구 분석’ 사업의 결과물을 발간한 것입니다.”



NATIONAL INSTITUTE OF KOREAN LANGUAGE



국립국어원