

국립국어원 2022-01-21

발간등록번호
11-1371028-000906-01

2022년 맞춤법 교정 말뭉치 연구 분석

연구책임자

남 길 임

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 맞춤법 교정 말뭉치 연구 분석’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2022년 05월 ~ 2022년 10월

2022년 10월 10일

연구책임자: 남길임(경북대학교)

연구 기관: 경북대학교 산학협력단

주식회사 이르테크

연구책임자: 남길임

공동연구원: 곽용진, 안미애, 송현주

안의정, 황은하

보조연구원: 심난희, 현영희, 강신아

백미경, 강윤희, 강민지

황지윤, 안진산, 박시온

고예린, 박아름, 정나현

김수지, 정희연, 성민규

장희선, 조은실, 배수종

안효민, 한도연, 이현영

이담허

<국문 요약>

2022년 맞춤법 교정 말뭉치 연구 분석

이 사업의 목적은 '21년 국립국어원 온라인 대화 말뭉치에서 선별한 100만 발화를 자동 형태소 분석, 기계 번역 등 한국어 처리 도구가 분석할 수 있는 수준으로 교정하고, 온라인 대화 텍스트의 특수성을 살린 교정 병렬 말뭉치의 구축 방안을 연구하고 이를 반영한 온라인 대화 맞춤법 교정 병렬 말뭉치를 구축하는 것이다. 이를 위한 사업의 범위는 다음 세 가지이다.

- 첫째, '21년 국립국어원 온라인 대화 말뭉치에서 맞춤법 교정 대상 대화를 선별한다.
- 둘째, '21년 사업에서 수립된 맞춤법 교정 지침을 정교화하고 개선한다.
- 셋째, 언어학적, 공학적 활용성을 고려한 맞춤법 교정 병렬 말뭉치를 구축한다.

각 하위 사업을 요약하여 제시하면 다음과 같다.

1) '21년 국립국어원 온라인 대화 말뭉치에서 맞춤법 교정 대상 대화를 선별

이 사업은 '21년 국립국어원 온라인 대화 원시 말뭉치 중 무의미 발화를 제외한 100만 개 이상의 발화에 대한 맞춤법 병렬 말뭉치를 구축하는 것이다. 따라서 426만 개 발화 중 125만 개의 발화를 선별하였는데 이에 대한 범위와 기준은 다음과 같다. 첫째, 원시 말뭉치인 '21년 온라인 대화 말뭉치 전체에서 8만 개의 대화 세트(대화별 발화 수 10개 이상, 총 발화 수 100만 개 이상)를 선별한다. 둘째, 대화 세트의 기준은 5개 발화 이상(Egbert et al. 2021: 716~721), 8~10회 발화(21년 한국어 대화 요약 데이터, 한국지능정보사회진흥원) 등 기존 연구에 근거하여 가장 엄격한 기준인 말풍선 단위 수 10회 이상으로 산정하되, 실질적인 대화 내용을 담는 발화만을 선별하기 위해 최초 25개 이상의 말풍선 단위로 구성된 대화 파일을 선별하였다. 여기서 말풍선 단위는 구어 원시 말뭉치의 구어 전사의 기본 단위로서 억양 단위(intonation unit)에 대응되는 단위로, '21년 온라인 대화 원시 말뭉치의 기본 단위이다. 셋째, 온라인 언어 특성을 잘 반영하는 발화를 우선적으로 선별하고, 부절적인 표현(욕설, 혐오 표현 등)이 포함된 발화는 구축 대상에서 제외하였다.

2) 맞춤법 교정을 위한 지침의 수립

이 사업의 주요 교정 대상인 '21년 국립국어원 온라인 대화 원시 말뭉치는 '21년 맞춤

법 교정 말뭉치 연구 분석 사업에서 이루어진 메신저 대화 말뭉치와 크게 다르지 않다. '21년 맞춤법 교정 병렬 말뭉치의 주요 방향은 언어학적, 공학적 연구를 위해 온라인 대화의 특성을 살리되, 구어 전사 말뭉치의 수준으로 정제하는 것이었다. 올해 사업 역시 기존 '21년 맞춤법 교정 말뭉치 연구 분석 사업의 주요 방향을 유지하되, 아래 세 가지 사항을 포함하는 개선 방향으로 지침을 보완하였다. '22년 맞춤법 교정 말뭉치 연구 분석 사업의 지침 개선 방향은 다음과 같다. 첫째, 띄어쓰기, 오타자 등에 대한 교정의 수준은 형태 분석의 효율 등을 고려하여 구어 전사 말뭉치 수준으로 교정하며, 원칙적으로 <우리말샘>을 기준으로 하였다. 둘째, <우리말샘>의 미등재어 및 비표준어에 대해서는 '21년도 지침과 동일하게 'OoV'(Out of Vocabulary)로 분류하되, '의미불명어' 범주를 새롭게 도입하여 관련 분류 항목을 세분화하였다. 셋째, 온라인 환경에서 나타나는 특수 표현을 원문자로 복원하는 방향으로 지침을 개선하고, 이모티콘은 별도의 범주를 부여하여 JSON 구조에 반영하였다. 교정 지침은 교정 유형별 지침으로 구성되며, 표준형과 비표준형의 판별은 <우리말샘>을 주요 기준으로 하되, 유형에 따라 별도의 지침을 수립하여 목록을 관리하였다. 이와 더불어 민간에서 변환 및 호환이 용이한 공공재로서의 말뭉치 활용을 위해 온라인 대화 말뭉치에서 나타나는 개인 정보 및 부적절한 표현을 파악하고, 비식별화하는 기준과 방안을 마련하였다.

3) 맞춤법 교정 병렬 말뭉치의 구축

맞춤법 교정 병렬 말뭉치의 효율적 구축을 위해서는 원시 텍스트에 맞춤법 검사기를 일괄적으로 적용하는 자동 교정과 개별 텍스트 맥락을 고려한 수작업 교정의 두 단계가 모두 필요하다. 따라서 본 연구는 원시 말뭉치에 맞춤법 자동 검사기를 활용하여 전처리 교정을 한 후, 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식으로 이루어졌다. 교정 이력의 체계적 관리를 위해 교정 병렬 말뭉치 구축 도구인 Kronoth를 사용하였으며, 교정 작업의 효율화와 일관성 검수를 위해서 Excel을 동시에 활용하였다. 또 ㈜이르테크의 말뭉치 검증 시스템을 활용해 분석 결과의 정확도를 확보하였다. 맞춤법 교정 말뭉치의 구축은 (1) 맞춤법 교정 대상 대화의 선별, (2) 텍스트 전처리를 통한 맞춤법 교정용 말뭉치 변환, (3) 자동 맞춤법 교정 도구를 이용한 1차 자동 교정, (4) 수작업 전수 교정, (5) 개인 정보와 부적절한 표현의 비식별화, (6) 세 차례의 품질 검수, (7) JSON 구조화, (8) 최종 형식 검수의 과정으로 구축되었다. 한편 이 연구에서는 기계 학습에 효율적인 구조 설계를 위해 현재 말뭉치 단위의 원문-교정 대응쌍의 구조를 원 어절-교정 대응쌍 구조로 개선하는 방안을 연구하였고, 어절 단위 교정쌍의 빈도와 실제 오류 경향성을 도출하였다. 이러한 연구 결과는 실제 온라인 대화 텍스트의 오류 경향성에 대한 최초의 연구로, 컴퓨터 공학에서 연구되어 온 자동 생성 오류 기반 연구를 보완하는 실증적 자료를 제공한다.

주요어: 맞춤법 교정 말뭉치, 온라인 대화 말뭉치, 구어 말뭉치, 교정 병렬 말뭉치, 맞춤법
검사기, 병렬 대응쌍

<Abstract>

Analysis of the 2022 Normalized Spelling Corpus

This project aims to select 1 million utterances from the 'Online Chat Corpus 2021' and build a Normalized Spelling Corpus to be analyzed with Korean language processing tools, such as, automatic translators, morphological analyzer, and so on. In order to achieve this, the project entails the following.

First, selecting 1 million utterances(Instant Message transmission units) for normalization from the 'Online Chat Corpus 2021'.

Second, refining and improving the guidelines of the normalization of the spelling which were established in 2021.

Third, constructing a parallel corpus of normalized spelling to be used as machine learning data for AI.

The above can be summarized as follows.

1) Selecting 1 million utterances(Instant Message transmission units) subject to spelling normalization

The Online Chat Corpus 2021, which is the object of study, includes unique conversation data types, such as, system messages (shared photos, shared videos, etc.) or conversations solely consisting of emoticons. In order to build a valid Normalized Spelling Corpus, more than 1 million utterances containing practical conversations on one or more topics were extracted, which represent well the characteristics of online messenger conversations and exclude meaningless conversations. In fact, more than 1.2 million utterances were selected from which were excluded conversation files that contained hate and discriminatory expressions in the context of de-identification of inappropriate expressions. The number of utterances subject to spelling correction thus extracted amounts to 1.25 million utterances.

2) Refining and improving guidelines 2021 for Normalized Spelling Corpus 2021

The Online Chat Corpus 2021 displays characteristics, such as, emotional expression using variations of symbols or spelling, colloquial expressions, non-normative expressions, typos and spelling mistakes, and ethical issues of hate speech expressions. Not only do such characteristics make it difficult to apply natural language processing tools, such as the morphological analyzer trained for written and spoken languages, but they also raise an ethical issue as for their use as machine learning data. In this project, normalization

guidelines were established by studying and analyzing the linguistic characteristics of typos, non-standard forms, and word spacing in the Online Chat Corpus in order to normalize it to the level of a spoken corpus. The guidelines consist of a set of principles aimed at each error type and are based on the guidelines for the 2021 Normalized Spelling Corpus, which have been refined and supplemented. While Urimalsaem served as the main standard for determining standard and non-standard forms, these were sorted out into lists using separate guidelines for each type. In particular, the two categories of ‘OoV’ (out-of-vocabulary) and ‘unknown sense’ were introduced to deal with the unrecorded words and non-standard forms of Urimalsaem, and further subdivide and classify the relevant items. Moreover, the guidelines were improved so that expressions that are typical of online contexts be preserved as the original text and emoticons be reflected in the Json structure by assigning them a separate category.

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) selection of the utterances subject to spelling normalization; (2) preprocessing of the text to convert the corpus for spelling normalization; (3) first-step automatic correction using an automatic spelling normalization tool; (4) manual correction; (5) de-identification of personal information and inappropriate expressions; (6) three-step check of overall quality; (7) JSON structuring; (8) final verification of the overall format.

Key-words: normalized corpus, online chat corpus 2021, spoken corpus, parallel corpus
normalized, automatic orthography checker, parallel equivalent

차 례

제1장 서론

1. 사업의 목적	1
2. 사업의 범위	2
2.1. 맞춤법 교정 대상 대화의 선별	2
2.2. 맞춤법 교정 지침 개선	3
2.3. 맞춤법 교정 병렬 말뭉치 구축	3

제2장 맞춤법 교정 말뭉치의 구축 및 연구

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	9
2.1. 맞춤법 교정 대상 대화의 선별	9
2.2. 맞춤법 교정용 말뭉치 변환	12
2.3. 자동 교정 및 후처리	13
2.4. 작업 교육	16
2.5. 수작업 전수 교정	19
2.6. 개인 정보와 부적절한 표현의 비식별화	21
2.7. 품질 검수	22
2.8. 최종 결과물 산출	24
3. 맞춤법 교정 병렬 말뭉치의 구조 및 주요 오류 유형 연구	29
3.1. 맞춤법 교정 병렬 말뭉치의 정렬 단위	29
3.2. 맞춤법 관련 주요 오류 유형	32

차 례

제3장 맞춤법 교정 말뭉치의 교정 지침 수립

1. 기본 지침과 지침 연구	43
1.1. 기본 지침	43
1.2. 지침 연구	44
2. 맞춤법 교정 말뭉치의 교정 지침	49

제4장 결론

1. 사업 요약	87
2. 향후 연구 및 정책 제언	88
2.1. 맞춤법 정렬의 단위 관련 제언	88
2.2. 맞춤법 오류 유형 연구 관련 제언	90

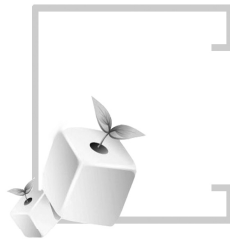
참고 문헌	92
-------------	----

표 차례

<표 1> 교정 대상 말뭉치의 유형과 규모	2
<표 2> 교정 대상 대화 선별 결과	10
<표 3> 교정 대상 대화의 메타 정보 비율	11
<표 4> 일괄 교정 항목	14
<표 5> 온라인 대화 특유의 고빈도 어형 목록	15
<표 6> 교정 작업 관련 유형과 횟수	15
<표 7> 맞춤법 교정 말뭉치의 JSON 형식 기본 구조	26
<표 8-1> 온라인 대화 맞춤법 교정 말뭉치의 JSON 양식	27
<표 8-2> 온라인 대화 맞춤법 교정 말뭉치의 JSON 양식	28
<표 9> 청크 단위 정렬 교정 말뭉치의 JSON 양식	35
<표 10> 교정 대상별 교정 양상 규모	35
<표 11> 문자열의 교정 값 기준 고빈도 목록	37
<표 12> 고빈도 미등재어 목록	38
<표 13> 고빈도 부호형 이모티콘	39

그림 차례

[그림 1] 맞춤법 교정 말뭉치 구축 단계	9
[그림 2] 교정 작업 파일 예시 (1)	12
[그림 3] 교정 작업 파일 예시 (2)	13
[그림 4] 구글 문서로 공유한 교정 지침 예시	16
[그림 5] 작업자 대상 1차 교육 자료 일부	17
[그림 6] 작업자 대상 2차 교육 자료 일부	17
[그림 7] 검토자의 피드백 예시	18
[그림 8] 질의응답용 구글 스프레드 시트 예시	19
[그림 9] Kronoth(v1.0) 교정 작업 예시	20
[그림 10] 엑셀(Excel)을 이용한 작업자 화면 (1)	20
[그림 11] 엑셀(Excel)을 이용한 작업자 화면 (2)	21
[그림 12] Kronoth의 작업자 화면	33
[그림 13] 주의해야 할 띄어쓰기 목록(구글 문서)	46
[그림 14] 수집한 의미불명어 목록 예시	46
[그림 15] 수집한 OoV 목록 예시	47



제 1 장

서 론



1. 사업의 목적

이 사업의 목적은 '21년 국립국어원 온라인 대화 말뭉치에서 선별한 100만 발화를 대상으로 자동 형태소 분석, 기계 번역 등 한국어 처리 도구가 분석할 수 있는 수준으로 교정하고, 온라인 대화의 특수성을 살린 교정 병렬 말뭉치의 구축 방안을 연구하고 온라인 대화 맞춤법 교정 병렬 말뭉치를 구축하는 것이다. 이를 위한 사업의 범위는 다음 세 가지이다.

첫째, '21년 국립국어원 온라인 대화 말뭉치에서 맞춤법 교정 대상 대화를 선별한다.

둘째, '21년 사업에서 수립된 맞춤법 교정 지침을 수정 보완하여 정교화한다.

셋째, 언어학적 및 공학적 활용성을 고려한 맞춤법 교정 병렬 말뭉치를 구축한다.

온라인 대화를 대상으로 한 교정 말뭉치 구축은 국어 자원의 활용도와 가치를 제고하고, 언어 사용 환경의 변화에 맞춘 언어 자료를 제공할 수 있다는 점에서 학계 및 산업계의 요구가 크다. 온라인 대화는 신어와 미등재어 등 다양한 언어 변이형을 포함하고 있으며, 방언을 포함한 다수의 비표준형이 나타난다. 또한 입력의 편의성으로 인한 띄어쓰기 무시, 오탈자, 표음적 표기, 기호나 철자의 변형 등 전통적인 말뭉치에서는 거의 나타나지 않는 다양한 현상이 등장한다. 따라서 현대 표준어를 기반으로 개발된 형태 분석기나 구문 분석기 등의 언어처리 도구로는 온라인 대화를 분석하는 데 큰 어려움을 겪을 수밖에 없다. 따라서 이 사업은 한국어 형태소 분석기의 적용이 가능한 말뭉치를 구축하는 것을 교정의 수준으로 삼는다.

본 연구팀은 이 사업을 통하여 앞서 구축된 온라인 대화 말뭉치를 대상으로 맞춤법 교정 말뭉치를 구축하는 데 필요한 지침을 수립하고 표준화된 말뭉치를 제공함으로써, 한국어 형태소 분석기가 적용 가능한 수준의 데이터를 제공하여 학계와 산업계의 요구에 부응할 수 있을 것이다. 또한 온라인 대화에 대한 맞춤법 교정 말뭉치는 새로운 언어 환경에 적합한 다양한 인공지능 기술 개발의 기반 자료로 활용 가능하므로, 대국민 서비스 강화에 이바지할 것이다.

2. 사업의 범위

이 사업은 다음의 세 가지 범위에서 수행된다. 첫째, 2021년 국립국어원 온라인 대화 말뭉치에서 맞춤법 교정 대상을 선별하는 것이다. 둘째, 2021년 사업을 통해 수집된 메신저 및 웹 대화를 대상으로 한 맞춤법 교정 지침을 개선하는 것이다. 셋째, 온라인 대화 말뭉치를 ‘원문-교정문’ 형식의 병렬 말뭉치로 정제 및 가공하는 것이다.

2.1. 맞춤법 교정 대상 대화의 선별

이 사업의 교정 대상 말뭉치는 국립국어원에서 2021년에 구축한 ‘온라인 대화 말뭉치’이다. ‘온라인 대화 말뭉치’는 메신저를 통해 수집한 대화문으로 구어의 형식을 띤 문어 텍스트 원시 말뭉치이며, 이 사업의 대상은 이 원시 말뭉치에서 별도로 선별한 약 400만 어절의 말뭉치이다.

제안 요청서 상으로 교정 대상 대화를 선별하는 조건은 대화별 발화 수 10개 이상이면서 전체 발화 수가 100만 개 이상인 대화 8만 세트를 구축하며, 대상 데이터에서 온라인 언어 환경의 특성을 잘 반영하는 발화를 선별하되, 부적절한 표현(욕설, 혐오 표현 등)이 포함된 발화는 구축 대상에서 제외하는 것이다. 이에 대해 이 사업을 위해 최종적으로 선별한 교정 대상 대화의 규모는 약 125만 발화로 아래 <표 1>과 같다.

교정 대상 말뭉치 유형	어절 수	발화 수
국립국어원 온라인 대화 말뭉치	약 400만 어절	약 125만 발화

<표 1> 교정 대상 말뭉치의 유형과 규모

여기서의 “발화”는 <2022년 맞춤법 교정 말뭉치 연구 분석> 제안 요청서상의 용어이다. ‘발화’가 온라인 대화 원시 말뭉치 구축의 기본 단위임을 고려할 때, 이는 구어의 발화 단위로 간주되는 구어 말뭉치의 억양 단위와 대응되는 개념이며, 엄밀히 말해 말뭉치 단위를 지칭함을 분명히 해 둘 필요가 있다.¹⁾ 여기서 ‘발화’ 단위 즉 말뭉치 단위는 ‘21년 온

1) 본 연구의 주요 대상인 “2021년에 구축한 ‘온라인 대화 말뭉치’”는 온라인 대화의 말뭉치를 기준으로 구축되었다. 여기서 ‘말뭉치’는 문어의 문장, 구어의 억양 단위 등과 상응하는 개념으로, ‘19년 메신저 대화 말뭉치’, ‘21년 온라인 대화 말뭉치’의 원시 말뭉치 구축 단계에서 기본 단위를 가리킨다. 여기서는 제안요청서상의 용어인 “발화”를 사용하였으나, 구어 말뭉치에서 ‘억양 단위’가 발화의 기본 단위이듯이, 온라인 대화 말뭉치에서 ‘발화 단위’는 ‘말뭉치 단위’에 상응한다. 한편, 온라인 대화의 단위로서 ‘말뭉치’는 만화 등에서는 ‘Speech bubble’로 번역되기도 하나, Baron(2010) 등 온라인 언어 연구자들 사이에서 ‘IM(instant messaging) transmission unit’으로 지칭되기도 한다. 본 연구에서는 제안요청서에 따라 과업 전반을 기술하는 자리에서는 ‘발화’라는 용어를 그대로 사용하되, 별도의 명시가 필요한 경우 ‘말뭉치(IM transmission unit)’를 사용하기로 한다.

라인 대화 말뭉치의 구축 단위로, 본 연구의 주요 과업인 맞춤법 교정 병렬 말뭉치의 대응쌍의 기본 단위이기도 하다.

2.2. 맞춤법 교정 지침 개선

'21년 맞춤법 교정 병렬 말뭉치의 주요 방향은 언어학적 공학적 연구를 위해 온라인 대화의 특성을 살리되, 구어 전사 말뭉치의 수준으로 정제하는 것이었다. 이 사업의 주요 교정 대상인 '21년 국립국어원의 온라인 대화 원시 말뭉치는 2021년 맞춤법 교정 말뭉치 연구 분석 사업에서 이루어진 메신저 및 웹 말뭉치를 대상으로 한 사업과 크게 다르지 않다. 따라서 2022년 사업 역시 기존 '21년 맞춤법 교정 말뭉치 연구 분석' 사업 결과 보고서의 주요 방향을 유지하되, 아래 세 가지 사항을 포함하는 개선 방향을 제시하였다.

첫째, 현행 국어 어문규정에 따른 띄어쓰기, 오타자 등에 대한 교정의 수준은 형태 분석의 효율 등을 고려하여 구어 전사 말뭉치 수준으로 교정하며, 원칙적으로 <우리말샘>을 기준으로 한다.

둘째, <우리말샘>의 외래어, 신어 등과 같은 미등재어 및 비표준어에 대해서는 '21년도 지침과 동일하게 'OoV'(Out of Vocabulary)로 분류하되, 2022년 사업에서는 '의미불명어' 범주를 새로 도입하여 목록을 관리할 수 있게 하였다.

셋째, 온라인 환경에서 나타나는 특수 표현을 원문자로 복원하는 방향으로 지침을 개선하고, 이모티콘은 별도의 범주를 부여하여 JSON 구조에 반영하였다.

이상의 세 가지 개선 방향을 포함한 맞춤법 병렬 말뭉치 구축을 위한 교정 지침은 교정 유형별 지침으로 구성되며, 국립국어원의 <우리말샘>을 우선 기준으로 삼되, <우리말샘>을 따르지 않는 유형은 별도의 지침을 수립하고 목록을 관리하였다. 이와 더불어 학계 및 산업계의 활용 가능성을 고려하여, 온라인 대화 말뭉치에서 나타나는 개인 정보 및 부적절한 표현을 파악하고, 비식별화하는 기준과 방안을 정교화하였다.

2.3. 맞춤법 교정 병렬 말뭉치 구축

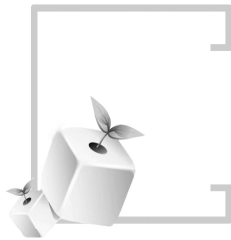
온라인 대화에 대한 맞춤법 교정 병렬 말뭉치는 문장부호를 포함한 맞춤법 교정과 함께 문장 단위 구획, 부적절한 발화의 제의를 통해 구어 전사 말뭉치 수준의 형태 분석은 물론이고 구문분석, 감성분석, 문서 요약 등에 활용할 수 있도록 구축하였다.

'원문-교정문' 형식의 병렬 말뭉치로 정제 및 가공하기 위해서는 원시 텍스트에 맞춤법 검사기를 일괄적으로 적용하는 자동 교정과 개별 텍스트 맥락을 고려한 수작업 교정의 두 단계가 필요하다. 따라서 본 연구는 원시 말뭉치에 맞춤법 자동 검사기를 활용하여 전처리 교정을 한 후, 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식으로 이루어졌다. 교정 작업의 이력의 체계적 관리를 위해 교정 병렬 말뭉치 구축 도구인 Kronoth를 사용하였으

며, 교정 작업의 효율화와 일관성 검수를 위해서는 Excel을 동시에 활용하였다. 또 (주)이르테크의 말뭉치 검증 시스템을 활용해 분석 결과의 정확도를 확보하였다.

맞춤법 교정 병렬 말뭉치는 ‘텍스트 전처리 → 자동 맞춤법 교정 도구를 통한 1차 교정 말뭉치 구축 → 세 차례의 검수 → 개인 정보와 부적절한 표현의 비식별화 → JSON 구조화 → 최종 형식 검수의 과정’의 단계로 구축되었다.

이 연구에서는 기계 학습에 효율적인 구조 설계를 위해 현재 말뭉선 단위의 원문-교정 대응 쌍의 구조를 원 어절 - 교정 어절 대응 쌍 구조로 개선하는 방안을 연구하였고, 어절 단위 교정 쌍의 빈도와 실제 오류 경향성을 도출하였다. 이러한 연구 결과는 실제 온라인 대화의 맞춤법 오류 유형을 파악하고 교정 정보를 확인할 수 있게 한 것으로, 컴퓨터 공학 분야에서 연구되어 온 자동 생성 오류 기반 연구를 보완하는 실증적 자료를 제공한다.



제 2 장

맞춤법 교정 말뭉치의 구축 및 연구



1. 맞춤법 교정 말뭉치의 유형과 특성

1.1. 교정 대상 말뭉치의 유형과 특성

이 사업의 교정 대상 말뭉치는 국립국어원에서 2021년에 구축한 ‘온라인 대화 말뭉치’이다. 이 말뭉치는 온라인 메신저인 카카오톡과 심심이를 통해 수집한 대화문으로, 구어의 형식을 띤 문어 텍스트 원시 말뭉치이며, 이 사업의 대상은 이 원시 말뭉치 대화 파일 전체인 88,949건, 약 400만 어절, 125만 발화이다. 이 말뭉치는 3,000명 이상의 대화 참여자를 사전에 모집하여, 이들을 대상으로 실시간 대화와 기존 대화 수집의 방식으로 수집한 대화문이다. 발화 수 기준으로 전체 온라인 대화 말뭉치의 51.24%가 기존 대화에서 수집된 대화문이며 나머지 말뭉치는 카카오톡 실시간 대화 25.53%, 심심이 채팅 대화 30.44%로 구성되어 있다. 대화의 유형은 2인 대화와 다자 대화(3인, 4인, 5인, 6인)로 구성되며, 다자 대화는 전체 중 0.42% 정도이다. 이 말뭉치는 텍스트로 된 대화문이므로 문어지만 구어의 특징도 가지는 말뭉치이다.

온라인 대화 말뭉치는 사용자의 어문규범이나 표현 규약이 사전에 정돈되지 않고, 예외적인 표기 등이 다수 나타난다는 특징이 있다. 온라인 대화의 특성상 이모티콘 등도 다수 사용된다. 이러한 점은 기존의 형태소 분석기나 맞춤법 교정기 등과 같은 NLP 도구의 자동 처리 정확도를 떨어뜨린다.

그럼에도 불구하고 구어와 문어의 특성을 공유하는 온라인 대화 말뭉치는 자연어 처리 연구에서 핵심적인 역할을 기대할 수 있다. 그러나 이 온라인 대화 말뭉치의 가치와 활용도를 높이기 위해서는 온라인 대화의 특성을 살리되 구어 전사 말뭉치의 수준으로 정제할 필요가 있다.

1.2. 맞춤법 교정 말뭉치의 구축 방향

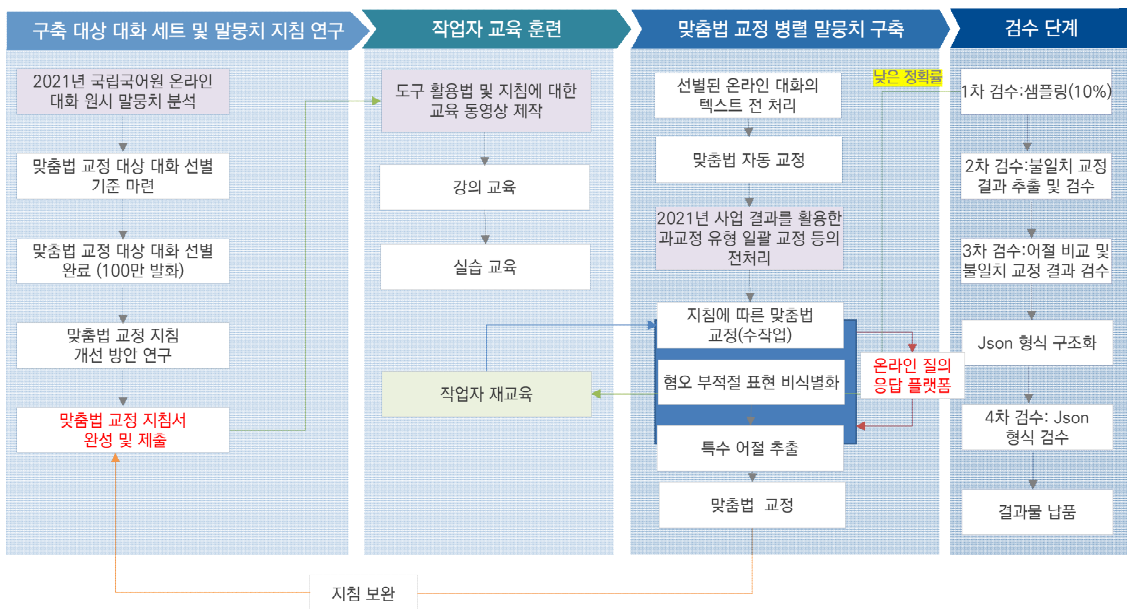
1.1에서 언급한 바와 같이 앞서 구축된 온라인 대화 8만 세트는 문어와 구어의 특징을 동시에 가지며, 장르 특징상의 비규범적 형태가 다수 나타나 자연어 처리 도구로 일정 수준의 교정을 기대하기 어렵다. 또한 이모티콘만으로 이루어진 대화나 자모 연쇄 발화 등의 정제되지 않은 온라인 대화가 그대로 남아 있다. 이에 이 사업은 2021년에 구축된 온라인 대화 말뭉치에서 의미 있는 발화 25개 이상으로 이루어진 대화를 선별하되, 20,000 어절 미만으로만 이루어진 대화 파일을 1차 선별하여 이를 대상으로 총 125만 발화를 선별하는 것을 사업의 첫 번째 목표로 삼는다. 다음으로 2021년 맞춤법 교정 병렬 말뭉치 사업의 지침을 정교화하여 이를 바탕으로 선별한 온라인 대화 원시 말뭉치를 규범적 형태로 전환한다. 최종적으로 이 사업은 기계 학습의 효율성을 고려하여, 원문과 교정문의 형

식을 갖춘 맞춤법 교정 온라인 대화 병렬 말뭉치를 구축하는 데 사업의 목적을 두고 있다.

맞춤법 교정 병렬 말뭉치는 말뭉치의 형태 분석이나 구문 분석 등의 심층적 분석을 시도하기 위한 필수 단계이다. 기존의 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 온라인 대화 말뭉치는 예측 범위 외의 비규범 변이형이 다수 존재하므로 맞춤법 교정기만으로 말뭉치의 품질을 제고하기 어렵다. 이에 이 사업은 2021년 맞춤법 교정 말뭉치 사업의 결과물을 고려하여, ‘표준화를 위한 초기 학습 데이터로서의 맞춤법 교정 병렬 말뭉치’를 ‘맞춤법 교정 온라인 대화 병렬 말뭉치 구축’의 기본 방향으로 삼는다. 여기서 ‘맞춤법 교정’의 수준은 완벽한 수준의 맞춤법 교정이 아니라 한국어 형태소 분석과 언어학 및 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 개인 정보와 부적절한 표현이 비식별화되어 있으므로, 공공재로 적절한 수준이다.

2. 맞춤법 교정 말뭉치의 구축 단계

맞춤법 교정 말뭉치 사업은 첫째, 언어학적 정밀성과 공학적 활용도를 고려한 맞춤법 교정 지침의 정교화, 둘째, 공공재로서의 가치와 활용도를 고려한 개인 정보 및 부적절한 표현 등에 대한 비식별화, 셋째, 실제 맞춤법 교정 병렬 온라인 대화 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 네 단계를 거쳐 구축되었다. 2장에서는 위의 [그림 1]의 구축 단계별 세부 공정 중 맞춤법 교정 대상 말뭉치를 선별²⁾한 후의 구축 단계 및 품질 검수와 최종 결과물의 산출 단계에 관해 설명하고자 한다.

2.1. 맞춤법 교정 대상 대화의 선별

서론에서 기술한 바와 같이, 이 사업은 맞춤법 교정 대상 대화의 선별부터 구축 공정에 포함된다. 제안 요청서상으로는 대화별 발화 수 10개 이상에 해당하는 총 발화 수 100만 개 이상을 선별하되, 대화 세트로는 8만 세트를 구축하는 것이나 실제 사업 수행에서는 실제 제안요청서상 선별 조건을 좀 더 엄격하게 해석하여 적용하였다. 이 연구에서 적용

2) 맞춤법 교정 대상 대화의 선별 단계에 대해서는 서론의 2.1절에서 기술하였다.

한 교정 대상 대화 선별 결과는 다음과 같다.

조건	제안요청서	실제
대화 파일별 최소 발화 수	10개	25개
총 발화 수	100만 개	125만 개
대화 파일별 최대 어절 수	.	20,000어절

<표 2> 교정 대상 대화 선별 결과

위의 <표 2>의 기준을 적용한 대화의 선별의 구체적인 과정은 다음과 같다.

첫째, 발화 수 25개 이상의 대화 파일만을 선별한다. 발화 선별 시 메신저의 ‘시스템 메시지(사진 공유, 동영상 공유 등)’, ‘자모 연쇄 발화(ㅋㅋ, ㅎㅎ 등)’, ‘이모티콘만으로 이루어진 대화’를 제외하였다. 이러한 선별 작업의 목적은 피상적인 대화 내용만을 담고 있는 대화 파일을 배제하고, 온라인 대화의 특성이 잘 드러나며, 하나 이상의 주제에 대해 실제적인 대화 내용을 담고 있는 발화만을 교정하여 맞춤법 교정 말뭉치의 활용도를 높이기 위해서이다.

둘째, 120만 발화 이상을 선별한다. 제안 요청서 상으로 최종 납품 발화 수는 100만 개 이나 이 사업에서는 이보다 20% 정도 많은 125만 개를 최종 납품 발화 수로 선정하였다. 이는 부적절한 표현에 대한 비식별화, 맥락상의 혐오 및 차별 표현 등이 드러난 대화 파일의 삭제 등으로 줄어들 수 있는 발화 수를 고려한 것이다.

셋째, 20,000어절 미만의 대화 파일만을 선별한다. 이는 말뭉치 구성의 다양성을 확보하고, 교정 작업의 일관성을 유지하기 위해서이다. 2021년 온라인 대화 말뭉치에 포함된 대화 파일의 어절 수를 살펴보면, 각각의 대화 파일은 최소 8어절에서 최대 45,965어절로 구성된다. 파일별로 어절 수의 차이가 큰 상황에서 말뭉치 구성의 다양성을 확보하기 위해서는 지나치게 용량이 큰 파일을 배제할 필요가 있다. 또, 개별 파일의 어절 수의 상한선을 둔 것은 한 명의 작업자가 온전히 하나의 대화 파일을 교정하기 위한 것이기도 한데, 20,000어절 이하의 경우 교정 및 주석 작업의 일관성, 비식별화 작업의 일관성, 2차 검수의 일관성 등을 확보하기에 용이하다. 실제로 2021년 맞춤법 교정 말뭉치 구축 사업 수행 결과, 한 명의 작업자가 수행하는 주별 작업 규모는 15,000어절에서 20,000어절 사이가 적절했으며, 20,000어절이라는 기준은 작업자의 작업 부담과 피로도, 사업 수행 일정 등을 종합적으로 고려해 적용한 것이다.

이러한 파일 선별과 배분을 통해 개인 정보의 비식별화 여부 및 맥락상의 혐오 및 차별 표현의 삭제 여부 등을 보다 일관되고 정확하게 판단할 수 있다. 한편 이와 같은 방법으

로 수집된 온라인 대화는 담화 단위(discourse units)의 측면에서 최소 12만 개 이상의 대화 세트로 해석될 수 있는데, 주요 선행 연구에서는 ‘의사소통의 일관성, 자족적 단위, 길이의 요건’ 등을 근거로(Egbert et al. 2021: 716~721), 5개 발화 이상(Egbert et al. 2021), 8~10회 발화(21년 한국어 대화 요약 데이터, 한국지능정보사회진흥원) 등을 기준으로 담화 단위를 설정한 바 있다.³⁾

또한, <표 2>의 선별 조건과 더불어 원시 말뭉치의 발화 메타 정보(사용 환경, 주제, 화자 정보(연령, 직업, 성별, 출생지, 신고 거주지, 실제 거주지, 입력 기기와 키보드 유형 등), 발화 세팅 정보(화자 간 관계, 친밀 정도, 연락 빈도)를 고려할 필요가 있었다. 2021년 온라인 대화 말뭉치는 특정 연령대와 성별이 큰 비중을 차지한다. 주요 메타 정보의 비율은 다음의 <표 3>과 같다. 본 연구팀은 이러한 원시 말뭉치의 한계를 최대한 고려하여 교정 대상 대화를 선별하였다.

메타 정보	항목	비율
publisher	카카오톡	37.50%
	심심이	62.50%
contact frequency	처음 연락한다.	50.07%
	(거의) 매일 연락한다.	32.43%
	주 3회 이상	10.62%
	주 1~2회	3.95%
	주 1회 미만	2.62%
	월 1회 미만	0.31%
age	10대	7.70%
	20대	54.85%
	30대	31.63%
	40대 이상	5.82%
sex	남성	11.10%
	여성	88.90%

<표 3> 교정 대상 대화의 메타 정보 비율

3) Egbert et al.(2021: 725-730)에서 논의하고 있듯이, 구어 담화 단위로서의 대화 단위 역시 연구의 초기 단계에 있으며, 구어 전사 말뭉치에서나 온라인 대화 말뭉치에서 주제적 요건을 고려하여 담화 단위를 자동으로 식별하는 것은 현재로서는 불가능한 일이다. Egbert et al.(2021)에서 논의한 담화 단위 구획의 세 가지 요소를 좀 더 상세히 소개하면 다음과 같다.

- ㄱ. 의사소통의 일관성: 의사소통 목적에 따라 하나 이상의 의사소통 기능을 가지는 발화의 연쇄임.
- ㄴ. 자족적 단위: 식별할 수 있는 처음과 끝을 가짐.
- ㄷ. 길이의 요건: 5개 발화나 100단어 이상으로 이루어짐.

2.2. 맞춤법 교정용 말뭉치 변환

'21년 온라인 대화 말뭉치는 메신저 시스템 메시지뿐만이 아니라, 무의미 발화, 자소만으로 이루어진 발화, 이모티콘만으로 이루어진 발화, 부적절한 표현으로 이루어진 발화 등이 포함되어 있다는 특징을 가지고 있다. 맞춤법 교정용 말뭉치 변환 단계에서는 이러한 발화를 배제하고 선별한 125만 개(대화별 최대 어절 수 20,000어절)의 대화를 맞춤법 교정을 위한 말뭉치로 변환하는 과정을 거쳤다.

교정용 말뭉치 변환 과정에서는 대화 파일의 메타 정보(출처, 구축 일시, 화자의 나이·성별·직업·출생지·거주지 등)를 제외하고, 대화 파일 ID, 발화 ID, 화자 ID, 원시 발화, 원시 발화의 어절 수 등 교정 작업에 필수적인 항목들만 포함하였다. 또한, 작업자의 조작 실수로 인한 데이터 손상을 막기 위해 원시 말뭉치의 대화 ID, 발화 ID, 화자 ID, 원시 발화가 포함된 열(column)은 도구 내 '시트 보호' 기능을 활성화하였다.

교정 작업 파일에 포함된 내용은 다음과 같다.

- ① 작업 진척 관련 내용: 마감일, 작업자, 검수자, 작업자 및 검수자 메모
- ② 발화 관련 내용: 대화 파일 ID, 발화 ID, 화자 ID, 참고용 원시 발화, 전처리가 완료된 교정 작업용 발화
- ③ 주석 관련 내용: 부적절한 표현 주석, 개인 정보(이름, 주소, 전화번호, 계좌번호) 주석, OoV(Out of Vocabulary) 주석, 이모티콘 주석, 의미불명어 주석

이상의 변환 과정을 거친 후의 교정 작업 파일 예시는 다음과 같다.

	A	B	C	D	E	F	G	H	I
1	대화 ID	마감일	작업자	검수자	발화 ID	누적 어	화자 ID	원 문장	최종 교정 문장(전처리 Ver 2.0)
2	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	1	2	하이하이	하이 하이
3	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	2	1	반가워욤ㅋㅋ	반가워욤. ㅋㅋ
4	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	5	1	name2님 제 이상형은	name2님 제 이상형은
5	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	12	1	코가 예쁘면 일단 외관 통관데 name2님은 어때여	코가 예쁘면 일단 외관 통관데 name2님은 어때요
6	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	17	2	오 저는 무조건 무쌍 존잘이여	오 저는 무조건 무쌍 존잘이여
7	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	22	2	아니 다들 내 이상형 못생겼다구우기는데	아니 다들 내 이상형 못생겼다고 우기는데
8	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	26	2	무쌍 존잘이 찐 존잘이지..	무쌍 존잘이 찐 존잘이지..
9	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	29	1	ㅋㅋㅋㅋ근데 무쌍 존잘	ㅋㅋㅋㅋ근데 무쌍 존잘
10	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	32	1	현실에 별로 없잖아요...	현실에 별로 없잖아요...
11	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	37	2	요새 강철부대 육준서에 마음이 섹섹섹됩니다.^^	요새 강철 부대 육준서에 마음이 섹섹섹 됩니다.^^
12	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	43	2	아 인정 저는 투디로 행복하러구요 ~	아 인정 저는 투디로 행복하러고요 ~
13	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	44	1	ㅎㅎㅎㅎ앗	ㅎㅎㅎㅎ앗
14	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	47	1	그분은 2.5d잖아요	그분은 2.5d잖아요
15	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	54	2	투디 채고!!! 차피 이상형 못만남..	투디 채고!!! 차피 이상형 못 만남.
16	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	55	2	ㅋㅋㅋㅋㅋㅋ인정,	ㅋㅋㅋㅋㅋㅋ인정,
17	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	58	2	무쌍 채고야 찌릿해	무쌍 채고야 찌릿해
18	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	61	1	힐 방금 찾아봤는데	힐 방금 찾아봤는데
19	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	62	1	개존잘이네요	개존잘이네요
20	MDRW21C	2022-06-05	강민지	현영희	MDRW21C	69	2	그저 그냥 전 이름이면 몸을 보는게 아닐까	그저 그냥 전 이름이면 몸을 보는 게 아닐까?

[그림 2] 교정 작업 파일 예시 (1)

	J	K	L	M	N	O	P	Q	R	S
1	험오 및 차별	전화번호	계좌번호	이름	주소	OoV	의미불명	이모티콘	작업자 매	검수자 매
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										

[그림 3] 교정 작업 파일 예시 (2)

2.3. 자동 교정 및 후처리

2.3.1. 맞춤법 자동 교정

맞춤법 자동 교정 단계에서는 2.1과 2.2절을 통해 선별하고, 변환한 온라인 대화 말뭉치를 대상으로 부산대의 맞춤법 검사기를 활용해 맞춤법 자동 교정 작업을 수행하였다. 자동 교정 대상 온라인 대화 말뭉치의 규모는 전체 대화 파일 수 24,524개(발화 수 1,255,570개, 어절 수 3,997,372개)이다. 자동 교정의 정확률이 높을수록 수작업 교정의 부담을 줄일 수 있으므로 맞춤법 교정 도구를 온라인 대화 텍스트 교정에 최적화하는 작업이 필요하다. 이를 위해 '21년 맞춤법 교정 말뭉치 대화 사업의 결과물⁴⁾과 부산대 맞춤법 검사기를 온라인 대화 말뭉치 교정을 위해 활용하였다.

다른 맞춤법 교정 도구들과 마찬가지로 부산대 맞춤법 교정기 또한 일반 텍스트의 교정을 위해 개발된 것으로, 자동 교정 결과에 대한 샘플링 검수를 통해 주요 오교정과 과교정 양상을 분석하여 일괄 교정이 가능한 오류 유형을 정리하였다. 또한 이렇게 도출된 오류 유형을 바탕으로 일괄 교정을 시행하였다. 이에 추가로 2021년 맞춤법 교정 말뭉치 사업의 결과물을 검토하고 1차 납품 오류 사례를 목록화하여 맞춤법 교정 말뭉치의 정확도를 높이기 위한 일괄 교정 작업을 시행하였다. 과교정 결과물에 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. 일괄 교정 대상 유형과 항목의 예는 다음과

4) '21년 맞춤법 교정 말뭉치 사업의 결과물은 국립국어원의 '모두의 말뭉치'에서 학술적 목적으로 사용 승인을 받은 후 활용하였다.

같다.

오류 유형	항목 수	오류 예시	교정형	비고
띄어쓰기	3,157	을만하다	-을 만하다	
붙여쓰기	438	해외 여행	해외-여행	
규범 표기	370	리트리버	레트리버	
과교정	152	데모용	데코용	‘데코용’을 ‘데모용’으로 과교정
지침 미준수	28	28일날, 부름말 뒤 쉼표	지침에 맞게 일괄 교정	

<표 4> 일괄 교정 항목

2.3.2. 온라인 대화 말뭉치의 특성을 고려한 텍스트 전처리

이 단계에서는 기존의 온라인 대화 말뭉치와 2021년 맞춤법 교정 말뭉치 사업을 통해 확보한 비규범적 유형 결과물을 활용하여, 선별한 온라인 대화 말뭉치에 대한 텍스트 전처리를 하였다. 2021년 맞춤법 교정 말뭉치 사업의 결과로 분석된 온라인 대화의 사례 중, 고빈도의 특징적인 사례들을 중심으로 비규범적, 예외적 특징을 정리하면 다음과 같다.

순위	어형	빈도	순위	어형	빈도
2	ㅋㅋ	13,729	102	네네	1,530
3	ㅋㅋㅋ	12,359	117	헉	1,330
19	ㅎㅎ	5,916	139	..	1,168
28	ㅠㅠ	4,463	171	아님	915
43	ㅎ	3,426	178	그치	884
45	헐	3,392	187	응응	858
51	웅	2,900	187	그런거	853
54	ㅇㅇ	2,791	188	마자	832
58	넴	2,608	208	그니까요	137
66	강	2,467	209	!!!	136
67	?	2,455	218	한거	135

순위	어형	빈도	순위	어형	빈도
82	그니까	1,837	219	ㅏㅑ	135
83	ㅓ	1,820	229	맞나	134
88	영	1,716	234	ㅎㅇ	134
96	ㅎㅎㅎ	1,600	237	name8이	133

<표 5> 온라인 대화 특유의 고빈도 어형 목록

위의 표와 같이 기존 형태 분석, 어휘 의미 분석으로 해결할 수 없는 고빈도 유형의 어절을 분류하고 분석한 결과, 특유의 한글 자모 연쇄, 축약어, 다양한 의성의태어, 자소 단위 의성의태어, 문장부호, 기호 이모티콘이 상위 빈도를 차지하는 것으로 나타나 이와 관련된 어형들을 텍스트 전처리 작업으로 정제하는 작업을 시행하였다.

또한 '21년 맞춤법 교정 말뭉치 사업의 결과물 검토 결과와 '22년 맞춤법 교정 말뭉치 사업 과정 중 1차 납품의 오류 유형을 검토한 결과를 바탕으로 고빈도 오류 유형을 목록화하여 텍스트 전처리 작업에 활용하였다. 분석 결과, 비규범형의 경우, 띄어쓰기 오류 교정 항목 수보다 붙여쓰기 오류 교정 항목 수가 더 많은 것으로 나타났으며, 비표준형 오류 교정 작업 수는 '바뀔' 작업이 총 131,596회, 8100여 종이 수행된 것으로 전체 작업 유형 중 높은 비율을 차지한 것으로 나타났다. 이중 다수를 차지한 것은 아래 예 1)과 같이 '-잡아', '돼(되어)'와 같은 어형과 관련된 오류와, '맞아'의 오류 형태를 바꾸는 경우로 전체 교정 작업의 4.3%를 차지했다.

예1) '바뀔' 관련 작업 상위 수행 어형 항목

- ① '-잡아' 관련: 1956회(1.48%) / 버릇자나, 찻자나, 차잔아, 대자낭, ... → -잡아
- ② '돼(되어)' 관련: 1921회(1.45%) / 더해야대, 있어야대, 집가야대, ... → 돼
- ③ '맞아' 관련: 1808회(1.37%) / 마져, 마자, 마적ㅋㅋ, 맞앙, ... → 맞아

교정 작업 전과 후, 다수의 교정 유형을 차지한 붙여쓰기, 띄어쓰기, 비표준형 수정 관련 항목의 횟수를 정리하면 다음과 같다.

항목	교정 횟수
붙여쓰기 오류 교정 항목 수	10,767회
띄어쓰기 오류 교정 항목 수	362,740회
비표준형 오류 교정 항목 수	131,596회

<표 6> 교정 작업 관련 유형과 횟수

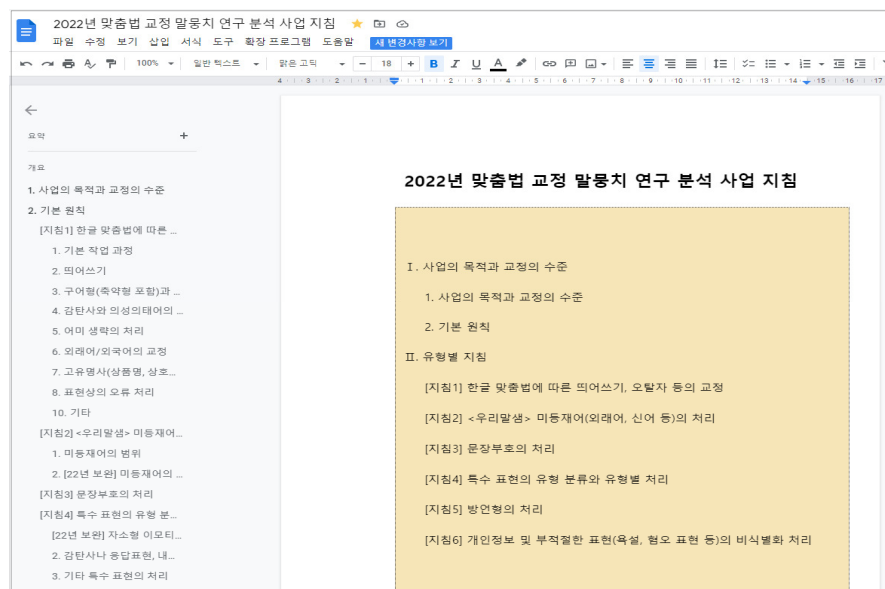
<표 6>에서 비표준형 오류 교정 작업은 주로 ‘바뀔(change)’을 통해 131,596회 수행되었으며, 이는 전체 교정 작업의 4.3%를 차지한다. 비표준형의 오류 유형은 8,100여 종으로 나타나 유형별로 평균 16회 이상 교정된 셈이다. 그밖에, 붙여쓰기와 띄어쓰기 오류는 모두 띄어쓰기 규정을 위반한 오류로, 띄어 써야 할 것을 붙여서 쓴 오류가 훨씬 더 많다 보니, 교정 전의 1,403,275어절에서 교정 후 1,785,791어절로 늘어나 어절 규모가 21.4% 증가하였다.

2.4. 작업 교육

작업자 교육은 크게 1) 사전 지침 교육, 2) 상위 검수자에 의한 샘플링 검수 및 피드백을 통한 재교육, 3) 구글 스프레드 시트를 이용한, 검수자와 작업자 간 실시간 질의응답의 결과를 공유하여 피드백하는 방식으로 진행되었다.

2.4.1. 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.

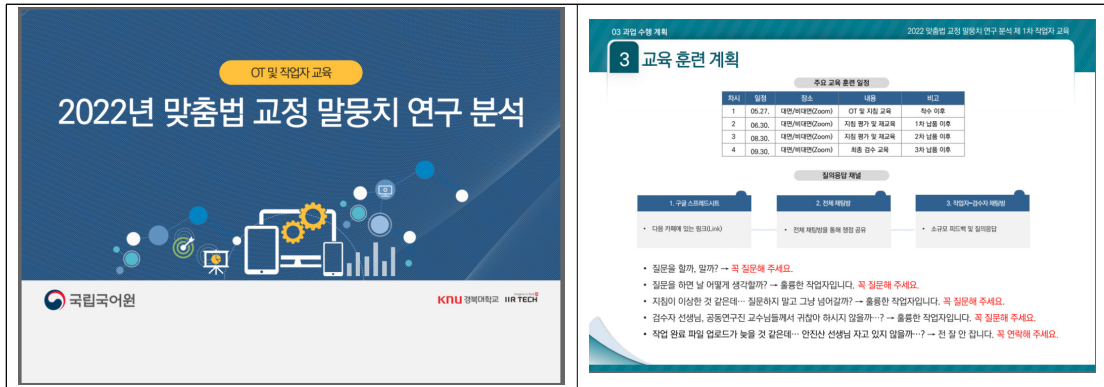


[그림 4] 구글 문서로 공유한 교정 지침 예시

또한 작업자의 지침 및 작업 도구 사용 방법을 숙지를 위해 작업자를 대상으로 한 작업

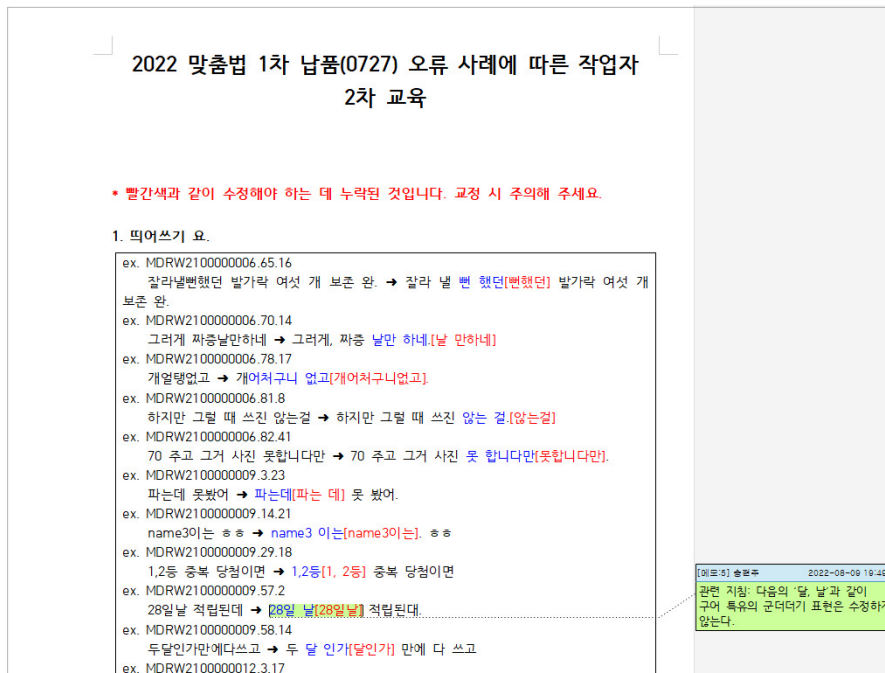
교육이 2회 시행되었으며 메신저와 구글 스프레드 시트를 통해 수시로 작업과 관련된 변경 사항을 안내하였다.

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



[그림 5] 작업자 대상 1차 교육 자료 일부

- 2차 교육: 작업 현황, 교정 지침 관련 안내, 작업 안내5)



[그림 6] 작업자 대상 2차 교육 자료 일부

5) [그림 6]은 국립국어원의 1차 납품 검토 결과를 본 사업팀에서 수정하여 사용한 것이다.

수작업으로 진행되는 맞춤법 교정 및 비식별화 작업의 상하위 검토 단계를 마련해 시행하였다. 검수는 (1) 작업자의 1차 교정 작업 결과물에 대한 검토자의 샘플링 검수 및 피드백, (2) 작업자의 피드백 반영, (3) 검토자의 피드백 반영 여부 검토의 단계로 진행되었다. 아래 그림은 엑셀로 진행한 검토자의 검수 및 피드백 예시이다. 이상의 검토 단계는 맞춤법 교정 말뭉치의 정확도와 완성도를 높이는 데 기여하였다.

[그림 7] 검토자의 피드백 예시

온라인 대화 말뭉치의 특성상 교정 과정에서 나타나는 변이형에 대한 고려가 필수적이다. 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 장점이 있다. 아래는 질의응답을 위해 공유한 본 사업팀의 구글 스프레드 시트 예시이다.

[illegible]

[그림 8] 질의응답용 구글 스프레드 시트 예시

2.5. 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.

2.5.1. Kronoth를 이용한 어휘 의미 분석 말뭉치 교정

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

이번 사업에서는 온라인 대화 말뭉치 중 1%에 해당하는 3만 어절 정도에 대해 맞춤법 교정 대응쌍 교정 유형 정보 주석 정보를 구축하고 질적 요소를 관리하기 위해 Kronoth를 활용하였다. Kronoth 도구는 아래의 그림의 교정 예시와 같이 교정 대응쌍 분석과 교정 유형 정보(유형, 빈도 등)를 도출하기에 적합한 도구이다.

2.5.2. 엑셀을 이용한 원시 말뭉치 교정

위의 2.5.1에서 제시한 Kronoth 도구의 강력한 교정 및 작업 관리 기능에도 불구하고, 전체 말뭉치에서 특정 오류 어형을 검색하여 일괄 바꾸기가 불가능하다는 점 때문에 이번 사업에서는 엑셀을 주 작업 도구로 활용하였다. 짧은 사업 기간을 고려할 때, 일괄 교정 기능은 작업 속도 향상과 동일 오류 형태에 대한 일관된 교정에 매우 필요한 기능이다.



[그림 9] Kronoth(v1.0) 교정 작업 예시

엑셀을 이용한 교정 작업은 도구에 대한 작업자의 접근성이 높고 작업이 편리하며 일관된 오류 어형을 찾아서 일괄 처리가 가능한 장점이 있다. 맞춤법 오류는 중복도가 높은 유형이 많으므로, 이를 일일이 읽으면서 교정하는 데는 많은 시간이 필요하며 누락의 위험도 따른다. 이에 따라 본 연구는 온라인 대화 말뭉치 100만 어절에 대해서는 엑셀을 이용해 전수 교정 작업을 진행하였다. 다음은 엑셀을 이용한 작업자 화면인데, 칼럼이 많은 관계로 두 화면으로 나누어서 제시한다.

1	A	B	C	D	E	F	G	H	I	J
1	대화 II	마감일	작업자	검수자	발화 II	누적 어	회자 II	원문장	최종 전처리 문장(Ver. 3.0)	원오 및 전오
13662	MDRW21	2022-07-31	장희선	백미경	MDRW21	8	1	안녕하세요! 혹시 요즘 넷플릭스나 유튜브 재밌게 보시는거?	안녕하세요! 혹시 요즘 넷플릭스나 유튜브 재밌게 보시는 거 있으신가요?	
13663	MDRW21	2022-07-31	장희선	백미경	MDRW21	14	2	저는 요새 대탈출 재밌게 봐서 지난시간 보고있어요	저는 요새 대탈출 재밌게 봐서 지난 시간 보고 있어요.	
13664	MDRW21	2022-07-31	장희선	백미경	MDRW21	20	2	유튜브는 주로 좋아하는 가수 영상 찾아봐요!	유튜브는 주로 좋아하는 가수 영상 찾아봐요!	
13665	MDRW21	2022-07-31	장희선	백미경	MDRW21	22	2	어떤거 즐겨보세요?	어떤 거 즐겨보세요?	
13666	MDRW21	2022-07-31	장희선	백미경	MDRW21	28	1	넷플릭스는 영화나 애니메이션을 주로 보는 편이에요!	넷플릭스는 영화나 애니메이션을 주로 보는 편이에요!	
13667	MDRW21	2022-07-31	장희선	백미경	MDRW21	38	1	유튜브로는 주로 좋아하는 밴드 유비나 예능 모음집 같은걸	유튜브로는 주로 좋아하는 밴드 유비나 예능 모음집 같은 걸 보고 있어요!	
13668	MDRW21	2022-07-31	장희선	백미경	MDRW21	51	1	전체적으로 보는것도 재밌는데 모음집은 재밌는 부분을 모아전체적으로 보는 것도 재밌는데	모음집은 재밌는 부분을 모아놔서 그런지	
13669	MDRW21	2022-07-31	장희선	백미경	MDRW21	59	2	마자요 모음집이 딱 재밌는것만 있어서 지루하지도 않고 좋3맞아요, 모음집이 딱 재밌는 것만 있어서 지루하지도 않고 좋조.		
13670	MDRW21	2022-07-31	장희선	백미경	MDRW21	62	2	예능은 어떤거 보세요?	예능은 어떤 거 보세요?	
13671	MDRW21	2022-07-31	장희선	백미경	MDRW21	73	1	요즘은 신서유기랑 강식당 보고 있어요!!! 전에는 안봤는데 요즘은 신서유기랑 강식당 보고 있어요!	전에는 안 봤는데 요즘 보니까 너무	
13672	MDRW21	2022-07-31	장희선	백미경	MDRW21	82	1	이렇게 재밌는 줄 알았더라면 전작 봤을텐데 왜 안봤나 모르 이렇게 재밌는 줄 알았더라면 전작 봤을 텐데 왜 안 봤나 모르겠어요.		

[그림 10] 엑셀(Excel)을 이용한 작업자 화면 (1)

H	I	J	K	L	M	N	O	P	Q	R	S	T
원 문장	최종 전처리 문장(Ver. 3.0)	혐오 및	전화번호	계좌번호	비밀번호	주소	소속	OoV	의미불	어모터	작업자	검토자
9하세요! 혹시 요즘 넷플릭스나 유튜브 재밌게 보시는 거 있안녕하세요! 혹시 요즘 넷플릭스나 유튜브 재밌게 보시는 거 있으신가요?												
는 요새 대탈을 재밌게봐서 지난시즌 보고있어요	저는 요새 대탈을 재밌게 봐서 지난 시즌 보고 있어요.											
투보는 주로 좋아하는 가수 영상 찾아봐요!	유튜브는 주로 좋아하는 가수 영상 찾아봐요!											
만거 옮겨보세요?	어떤 거 옮겨보세요?											
클로는 영화나 애니메이션을 주로 보는 편이에요!	넷플릭스는 영화나 애니메이션을 주로 보는 편이에요!											
루로하는 주로 좋아하는 밴드 유비나 예능 모음집 같은걸 보유주브로는 주로 좋아하는 밴드 유비나 예능 모음집 같은 걸 보고 있어요!												
레적으로 보는것도 재밌는데 모음집은 재밌는 부분을 모아놔진채적으로 보는 것도 재밌는데 모음집은 재밌는 부분을 모아놔서 그런지 시간 가는 줄 모르고 보네요. ㅋㅋㅋㅋ												
이요 모음집이 딱 재밌는것만 있어서 지루하지도 않고 좋조 맞아요. 모음집이 딱 재밌는 것만 있어서 지루하지도 않고 좋조.												
5은 어떤거 보세요?	예능은 어떤 거 보세요?											
장은 신서유기랑 강서당 보고 있어요!! 전에는 안봤는데 요! 요즘은 신서유기랑 강서당 보고 있어요! 전에는 안 봤는데 요즘 보니까 너무 재밌네요. ㅋㅋㅋㅋ												
장게 재밌는 줄 알았더라면 전작 봤을텐데 왜 안봤나 모르겠어!항게 재밌는 줄 알았더라면 전작 봤을 텐데 왜 안 봤나 모르겠어요.												
ㅋㅋㅋㅋ저요 둘다 재밌조ㅋㅋㅋㅋ	아, ㅋㅋ 맞아요. 둘 다 재밌조. ㅋㅋ											
2봐도 너무 웃겨음ㅋㅋㅋㅋ	봐도 봐도 너무 웃겨요. ㅋㅋㅋㅋ											

[그림 11] 엑셀(Excel)을 이용한 작업자 화면 (2)

[그림 10]을 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 10]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 구분할 수 있도록 하였다. ‘원 문장’ 칼럼에는 교정 작업 전 발화문이 제시되어 있고, ‘최종 작업 문장’ 칼럼에는 자동 교정 및 텍스트 후 처리된 발화문이 들어 있으며, 해당 칼럼에서 작업자가 교정 작업을 진행한다.

[그림 11]의 발화문 이후 칼럼은 주로 OoV와 함께 의미가 불분명한 어형을 표시하고 비식별화하기 위한 주석 영역이다. 이에 덧붙여 작업자와 검토자의 소통을 위한 칼럼도 마련되어 있다.

2.6. 개인 정보와 부적절한 표현의 비식별화

작업자가 특정 내용을 임의로 비식별화하지 않도록 본 사업단은 개인 정보 및 부적절한 표현에 대한 비식별화 방안을 마련해 작업에 반영하였다. 먼저 개인 정보는 사업의 전체적인 유기성을 고려해 2021년에 추진한 국립국어원의 맞춤법 교정 말뭉치 사업에서 제시한 지침에 의거하여 비식별화하였다.

2.6.1. 개인 정보의 비식별화

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) 등을 철저하게 비식별화하였다.

2.6.2. 부적절한 표현의 비식별화

부적절한 표현은 혐오 및 차별 표현, 욕설, 성적 표현 등을 포함하며, 비식별화 대상이다. 혐오 표현은 국가인권위원회의 혐오 표현 리포트(2019)에 따르면 “성별, 장애, 종교,

나이, 출신 지역, 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 2021년 맞춤법 교정 말뭉치 사업의 지침과 동일하게, ‘형태’가 아닌 ‘맥락’을 기반으로 판단하였으며, 판단한 기준과 결과에 대해 국내 혐오 표현에 대한 전문가 자문으로 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.

2.7. 품질 검수

품질 검수 단계에서는 맞춤법 수작업 전수 교정 결과에 대해 4차례에 걸친 품질 검수를 실시하여 말뭉치의 교정 품질을 최대화하였다.

2.7.1. 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 진행되었다.

1차 샘플링 검수의 결과는 다음과 같은 세 가지로 반영된다. 우선, 작업자에게 주요 오류 유형을 알림으로써, 나머지 90%의 교정에 반영하도록 한다. 다음으로, 주요 오류 유형을 수집하여 목록화함으로써 2차 검수에 대비한다. 끝으로, 정확률이 낮은 작업자에 대해서는 지침 및 작업에 관한 재교육을 제공하는 방식으로 작업의 품질을 높인다.

2.7.2. 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 교정 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 오

류 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 해당하는 목록은 다음과 같다.

- 예2) 1) 컴터 → 컴퓨터, 쌤 → 쌤, 라떼 → 라테, 큐알 코드 → 큐아르 코드...
2) 아냐 아냐 → 아냐, 아냐, 가자 가자 → 가자, 가자...
3) 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 해당한다.

- 예3) 1) 못 하다/못하다, 잘 하다/잘하다, 한번/한 번...
2) -(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.

2.7.3. 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 온라인 대화 말뭉치의 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 확인한다. 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.

2.7.4. 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 교정하면 최종 검수가 완료된다.

2.8. 최종 결과물 산출

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

2.8.1. 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 단위로 원문과 병렬하여 제시한다. 한편, 2021년도 사업 대비 2022년도 사업의 JSON 구조에는 ‘이모티콘’과 ‘의미불명어’에 대한 속성값이 추가되었다. <표 7>은 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 연도
	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

		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_residence	str	구어(준구어 제외), 메신저	전체	주 성장지
			current_residence	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	전체 (단, 원시 제외)	발화
		id		str	맞춤법 교정	전체 (단, 원시 제외)	발화 ID
		original_form		str	맞춤법 교정	전체 (단, 원시 제외)	원문 형태
		form		str	맞춤법 교정	전체 (단, 원시 제외)	정제된 형태
		corrected_from		str	맞춤법 교정 또는 비식별화	전체 (단, 원시 제외)	맞춤법 교정 형태, 또는 부적절한 표현의 경우 "hate-speech"로 비식별화함
		speaker_id		str	맞춤법 교정	전체 (단, 원시 제외)	화자 ID

		emoticon		arr (obj)	맞춤법 교정	전체 (단, 원시 제외)	이모티콘
			begin	str	맞춤법 교정	전체 (단, 원시 제외)	시작 위치 값
			end	str	맞춤법 교정	전체 (단, 원시 제외)	끝 위치 값
			value	str	맞춤법 교정	전체 (단, 원시 제외)	이모티콘 형태
		meaningless_ word		arr (obj)	맞춤법 교정	전체 (단, 원시 제외)	의미불명어
			begin	str	맞춤법 교정	전체 (단, 원시 제외)	시작 위치 값
			end	str	맞춤법 교정	전체 (단, 원시 제외)	끝 위치 값
			value	str	맞춤법 교정	전체 (단, 원시 제외)	의미불명어 형태

<표 7> 맞춤법 교정 말뭉치의 JSON 형식 기본 구조

2.8.2. JSON 양식

JSON으로 변환한 최종 결과물의 JSON 양식을 온라인 대화 맞춤법 교정 말뭉치의 예로 들어 보이면 다음의 <표 8>과 같다.

```
{
  "id": "MXSC2202210100",
  "metadata": {
    "title": "국립국어원 온라인 대화 말뭉치 추출 MXSC2202210100",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2022",
    "category": [
      "온라인 대화 > 2인 대화",
      "온라인 대화 > 다자 대화"
    ],
    "annotation_level": "맞춤법 교정",
    "sampling": "부분 추출 - 임의 추출"
  },
  "document": [
    {
      "id": "MDRW2100000002.1",
      "metadata": {
        "title": "온라인 대화",
        "author": "개인 대화 참여자",
        "publisher": "카카오톡",
        "date": "20210518",
        "topic": "연애와 결혼",
        "speaker": [
          {
            "id": "1",
            "age": "20대",
            "occupation": "사무 종사자",
            "sex": "여성",
            "birthplace": "서울",

```

6) 발화 내용이 부적절한 표현 등의 이유로 비식별화 대상인 경우에 속성 “corrected-form”의 값에는

```

        "pricipal_residence": "서울",
        "current_residence": "서울",
        "device": "스마트폰",
        "keyboard": "2벌식(쿼티)"
    },
    {
        "id": "2",
        "age": "20대",
        "occupation": "전문가 및 관련 종사자",
        "sex": "여성",
        "birthplace": "서울",
        "pricipal_residence": "해외/기타",
        "current_residence": "서울",
        "device": "스마트폰",
        "keyboard": "2벌식(쿼티)"
    }
},
"setting": {
    "relation": "직장>선후배/상사-부하"
},
},
"utterance": [
    {
        "id": "MDRW2100000002.1.1",
        "original_form": "하이하이",
        "form": "하이하이",
        "corrected_form": "하이하이.",
        "speaker_id": "2",
        "emoticon": [],
        "meaningless_words": []
    },
    {
        "id": "MDRW2100000002.1.2",
        "original_form": "반가워욥트ㅋㅋ",
        "form": "반가워욥트ㅋㅋ",
        "corrected_form": "반가워요. ㅋㅋㅋㅋ",
        "speaker_id": "1",
        "emoticon": [],
        "meaningless_words": []
    },
    ...
    {
        "id": "MDRW2100000002.1.10",
        "original_form": "요새 강철부대 육준서에 마음이 선택선택됩니다{emoji:…}{emoji:…}.^^",
        "form": "요새 강철부대 육준서에 마음이 선택선택됩니다.^^",
        "corrected_form": "요새 강철부대 육준서에 마음이 선택선택됩니다. ^^",
        "speaker_id": "2",
        "emoticon": [
            {
                "begin": 25,
                "end": 27,
                "value": "^^"
            }
        ],
        "meaningless_words": []
    },
    ...
    {
        "id": "MDRW2100000005.40.11",
        "original_form": "웨친",
        "form": "웨친",
        "corrected_form": "hate-speech6)",
        "speaker_id": "1",
        "emoticon": [],
        "meaningless_words": []
    },
    ...
]

```

<표 8-1> 온라인 대화 맞춤법 교정 말뭉치의 JSON 양식

“hate-speech”를 입력한다. 예시에서 “웨친”은 비식별화 대상이므로 “corrected-form”의 속성값에 “hate-speech”라고 제시되어 있다.

본 사업에서 제안되지는 않았지만, 말뭉치 활용의 목적 및 용도에 따라 혐오·차별 표현 등 부적절한 표현에 대한 정보를 더 명시적이고 적극적으로 추가하고자 하는 경우, 상기의 JSON 구조에 부적절한 표현에 대한 속성-값을 추가하는 것도 고려해 볼 만하다. 이때, 부적절한 표현에 대한 정보를 획득하고 부적절한 표현이 포함된 발화 전체가 아닌 부적절한 표현을 정확히 구분하기 위해 hate_exp_infos 정보가 포함된다. 이를 위한, 가능한 JSON 구조의 예시를 보이면 다음과 같다.

```
...      {
          "id": "MDRW2100000005.40.11",
          "original_form": "꽤 친",
          "form": "꽤 친",
          "corrected_form": "hate-speech",
          "hate_exp_proc_form": "hate-speech",
          "speaker_id": "1",
          "emoticon": [],
          "meaningless_words": [],
          "hate_exp_infos": [
            {
              "begin": 1,
              "end": 2,
              "value": "꽤 친"
            }
          ]
        }
```

<표 8-2> 온라인 대화 맞춤법 교정 말뭉치의 JSON 양식

3. 맞춤법 교정 병렬 말뭉치의 구조 및 주요 오류 유형 연구

맞춤법 교정 말뭉치 사업은 온라인 대화의 특수성을 살림과 동시에 형태소 분석이나 기계 번역 등 한국어 처리 도구가 분석할 수 있는 수준으로 교정하는 다소 상충되는 목적을 구현하는 것이며, 맞춤법 오류 원시 값과 교정 값에 대한 정밀한 정보를 추출하기 위해서는 교정 전후 텍스트의 정렬 단위에 대한 논의가 필요하다. 나아가 교정 전후의 말뭉치를 비교해 주요 오류 유형을 귀납하는 것은 자동 노이즈 생성 말뭉치의 구축 등의 활용을 위해 매우 필요하다. 따라서 말뭉치의 원문과 교정문의 정렬 단위(3.1)의 쟁점, 교정 전후의 통계적 특성과 주요 오류 유형을 논의하고 분석함으로써(3.2) 향후 과제의 기획과 수행의 방향에 기초 자료를 제공하고자 한다.

3.1. 맞춤법 교정 병렬 말뭉치의 정렬 단위

병렬 말뭉치의 정렬 단위는 크게는 텍스트에서 문단, 문장, 작게는 절, 구, 단어 등 단위까지 다양한 언어 단위로 상정 가능하다. 말뭉치의 활용도 향상이라는 측면에서 보면 정렬하는 언어 단위가 작을수록 세밀한 언어 정보를 추출할 수 있겠으나, 작업의 가능성과 편의성이라는 관점에서는 언어 단위가 작아지면 작업의 부담 및 그에 따른 소요 시간과 노력은 증가한다. 따라서 정렬 단위의 결정은 말뭉치의 활용과 작업의 편의라는 상충하는 두 기준에서 출발하되, 절충점을 찾을 필요가 있다.

앞선 사례를 살펴보면, 언어 간 병렬 말뭉치는 문장 대 문장 정렬이 일반적이고, 언어 내 병렬 말뭉치의 경우는 텍스트의 특성상 정렬 단위에서도 말뭉치에 따라 다른 특성을 보인다.⁷⁾

언어 간 병렬 말뭉치들과는 달리, 온라인 대화 텍스트는 두드러진 언어적 특성으로 말풍선 단위를 고려할 필요가 있다. 그런데 말풍선을 기존의 문어 중심 문법의 문장, 즉 “생각이나 감정을 말과 글로 표현할 때 완결된 내용을 나타내는 최소의 단위.”(<표준국어대사전>, 1999)라는 관점에서 보면 그 실현 양상이 매우 다양하다는 데 쟁점이 있다.

7) 언어 간 병렬 말뭉치의 정렬 단위는 최초의 언어 간 병렬 말뭉치인 LOB(Lancaster-Oslo-Bergen) 말뭉치부터 시작하여 국내의 세종 계획의 한중, 한불, 한러 병렬 말뭉치, 중국의 베이징대 중영 이언어 말뭉치(汉英双语语料库), 중일(中日对译语料库), 난징대 영중 이언어 말뭉치(NJU_BDRBCB, 南京大学英汉双语平行语料库), “21세기 세종 계획” 병렬 말뭉치 등 알려진 병렬 말뭉치들은 대부분 텍스트, 문단, 문장 단위로 정렬되어 있으며, 이를 구축하거나 용례 검색을 위한 도구인 Paraconc, AntPConc, CUC_Parac, HepEditor.exe 등의 병렬 말뭉치 구축 도구 또한 텍스트-문단-문장 단위까지 정렬이 가능하도록 개발되어 있다(황은하, 2016).

예1) 화자1)-1 좋겠네

화자2)-1 잘해야하는데..

화자2)-2 나아지는듯 하더니

화자2)-3 다시켰어요

위의 예1)의 화자2)-1은 연결어미로 끝을 맺고 있기 때문에 중의적인 해석이 가능하다. 화자2)-2와 화자2)-3은 인과 관계의 문장으로 보기에 무리가 없는데, 화자2)-1이 하나의 절로서 이들과 같은 문장을 이루는지, 아니면 화자1)-1에 대한 응답으로서 말줄임표로 끝난 것인지 하는 두 가지 해석의 가능성이 있다. 이와 같이 온라인 대화 자료에서 문장 단위의 구획의 불명확성과 중의성, 온라인 대화 텍스트의 특성 유실, 작업의 일관성 준수 어려움 등을 고려하여 본 과제에서는 말풍선 단위 정렬을 수행하였는데, 향후 원문과 교정문의 보다 세밀한 대응 정보의 추출이나 기계 학습을 위해서는 여전히 말풍선의 경계를 보존하면서 문장, 어절, 형태소 등의 보다 작은 단위의 정렬을 고려해 볼 필요가 있다.

예2)

[문장 대응] “가진 않겠져?/가진 않겠쥬?”, “세번 먹었는데ㅋㅋ/세 번 먹었는데 ㅋㅋ”

[어절/청크 대응] “않겠져/않겠쥬”, “세번/세 번”, “먹었는데/먹었는데”

“안되요/안 돼요”, “아니었어/아니었어”

[형태소 대응] “여/요”, “ㅏ/ㅑ”, “는데/는데”

위의 예2)에서 보인 것처럼 어절 또는 청크(chunk)⁸⁾ 단위의 정렬은 띄어쓰기가 구분자 역할을 하므로 구획이 용이하고, 말풍선이나 문장보다 단위가 작아서 원문-교정문에 대한 보다 세밀한 대응 정보의 추출이 가능하기 때문에 그에 대한 시도가 요구된다. 한편 교착어로서의 한국어의 특성을 반영할 때, 형태소 단위의 대응을 통해야만 맞춤법 오류 정보를 보다 정확하게 관찰할 수 있는 경우도 있다.

정렬 단위에 대해 본 과제의 자문위원인 임희석 교수는 데이터 세트 구축의 속도와 품질 사이의 균형을 고려하면 ‘문장 대응’ 단위가 적절해 보이나, 어절 또는 형태소 단위의 효용성도 무시할 수 없다는 의견을 제시하였다.

8) 어절은 “문장을 구성하고 있는 각각의 마디. 문장 성분의 최소 단위로서 띄어쓰기의 단위가 된다.”(<표준국어대사전>, 1999)로 정의되며, 이에 따르면 ‘먹고싶어해요 침도흘리구용’, ‘두번’과 같은 교정 전 원시문은 띄어쓰기를 구분자로 각각 2어절, 1어절로 보는 것에는 무리가 있다. 따라서 이와 같은 단위를 자연어 처리에서 흔히 사용하는 용어인 ‘부분 구문’, 즉 ‘청크’로 하는 것이 더 타당하다고 하겠다.

- 기계 번역 모델은 문장 단위로 의미를 이해하여 번역을 수행하므로 데이터 세트 구축의 속도와 품질 사이의 균형을 고려하여 ‘문장 대응’ 단위 정도가 적절해 보임.
- 물론 어절 단위나 형태소 단위의 효용성은 무시할 수 없고, 어절 단위 대응을 통해 문장 단위 대응쌍도 생성 가능함. 어절 단위 대응의 구조를 갖는다면 문장 단위로 복원할 수 있도록 구조를 설계할 필요가 있음. 속도와 품질의 절충안으로 문장 단위가 적절함. 한편, 현재의 온라인 대화 말뭉치는 ‘말풍선 대 말풍선’의 대응 구조로 구축되어 있기 때문에 이를 문장 단위로 대응하기 위해서는 별도의 처리가 필요함.⁹⁾

여기서 ‘어절’은 원문의 띄어쓰기 오류로 인해 사실상 다:다 어절 대응, 즉 구 또는 자연어 처리에서 흔히 말하는 청크(chunk) 단위 대응으로 보는 것이 타당하다.

예3) “잘받아야하는데../잘 받아야 하는데...”, “나아지는듯 하더니/나아지는 듯하더니”

위의 예3)을 보면 띄어쓰기 기준으로 어절을 획분하면 1:3, 2:2이나, 원문의 ‘잘받아야하는데’나 ‘나아지는듯 하더니’를 단순히 띄어쓰기 기준으로 각각 1, 2어절로 보는 것은 문제가 있으며, 청크 단위의 대응으로 보는 것이 보다 타당하다고 하겠다. 또한, 청크가 문장의 하위 단위이기 때문에, 문장 또는 말풍선 대응 정보도 함께 제공될 수 있다는 장점이 있다.

다만, 청크 대응 정보로는 관찰이 불가능한 오류 유형이 있는데, 형태 단위의 오류와 교정형의 대응이다.

예4)

[어절/청크 대응] “않겠져/않겠쥬”, “돼여/돼요”, “씻는덴/씻는데”, “해용/해요”

[형태소 대응] “여/요”, “여/요”, “스/쓰”, “는덴/는데”, “용/요”

위의 예의 어절/청크 대응과 형태소 대응의 결과를 살펴보면, 형태소 단위의 대응 결과에서 얻을 수 있는 정보와 어절/청크 대응에서 얻을 수 있는 정보가 다를 수 있음을 확인할 수 있다. 한국어의 교착어의 특성상 형태소 단위의 정렬은 매우 필요하지만, 이런 작업은 형태

9) 즉, 고려해야 할 주요 사례로 (1) 하나의 말풍선이 여러 개의 문장으로 구성된 경우, (2) 여러 개의 말풍선이 하나의 문장으로 구성된 경우, (3) (2)와 같은 상황에서 타 화자의 개입을 처리해야 하는 경우 등이 있다. 이 경우 온라인 대화의 특성을 고려한 구체적인 문장 구획 지침을 작성할 필요가 있다.

소 분석이 선행되어야만 가능하며, 특히 오류가 들어 있는 원문에 대한 자동 형태소 분석은 정확도가 낮으므로 시간 소모적이며, 많은 비용이 발생할 수밖에 없다는 문제가 있다.

요약하면, 정렬 단위가 작을수록 오형태와 교정 형태의 대응쌍을 관찰하고 추출하고 언어학 연구, 언어 공학 연구 등 범용적으로 사용하는 데 효과적이다. 그러나 작업의 가능성이나 들이는 노력 대비 효과를 보았을 때, 현재의 말풍선 단위 정렬에 더해 문장 단위, 청크 단위의 정렬이 이루어진다면 한국어의 첨가어 및 자유어순 등의 형태·통사론적 특성상 보다 높은 활용 가치를 기대할 수 있다.

3.2. 맞춤법 관련 주요 오류 유형

맞춤법 관련 주요 교정 및 주석 내용에 대한 관찰과 분석은 맞춤법 오류를 인위적으로 생성하는 자동 노이즈 생성 말뭉치의 구축을 위한 기초 자료로서 중요한 의미를 지닌다. 한국어 어문 규범에서도 한글 맞춤법이나 외래어 표기법과 관련한 문자열의 주요 오류 유형은 문장보다는 작은 단위의 교정 전후 대응이 이루어질 때 비로소 가능한데, 어문 규범에 준한 교정 전후 결과는 현재의 원문과 교정문의 문장 단위 대응 구조로는 띄어쓰기의 변화 외에는 주요 교정 유형과 빈도에 대한 관찰이 어렵다. 따라서 본 과제에서는 추가 제안으로 전체의 약 0.5%인 15,726어절에 대해 청크 단위 정렬 교정 병렬 말뭉치를 구축하여 청크 단위의 교정 전후 비교 분석을 통해 주요 교정 유형을 관찰하고자 한다.

3.2.1. 청크 단위 정렬 말뭉치 구축 방법

청크 단위의 정렬 및 교정을 위해서는 일반 텍스트 편집기가 아닌 어절 단위 대응 정보 추출이 가능한 도구의 지원이 필요한데, 이를 위해 이르테크에서 개발한 Kronoth를 사용하였다.



[그림 12] Kronoth의 작업자 화면

또한, 비규범형과 규범형 간의 보다 면밀한 대응 관찰을 위해 교정 시 통제된 조작 규칙을 정하고 따르도록 하였다.

예5) 원시문: 귀를 막고있습니다
교정문: 귀를 막고 있습니다.

위의 예의 원시문을 교정문으로 교정하는 데는 가능한 조작 방법이 여러 가지가 있다. 교정의 양상은 크게 문자열, 빈칸, 문장부호 등의 삽입, 교체, 삭제 세 가지가 있다. 예5)에 대해 가능한 조작의 방법 두 가지를 보이면 다음과 같다.

예6) 1) 교체: 막고있습니다 → 막고 있습니다.
2) 교체: 고있을 → 고 있습, 삽입: 마침표

작업자에 따라 위의 예6)의 1)처럼 교체 1회를 통해 정확한 교정문을 얻는 경우도 있지만, 예6)의 2)처럼 선 문자열 교체, 후 마침표 삽입을 통한 교정도 가능하다. 그런데 이처럼 같은 유형의 오류에 대해 다른 방식의 교정을 수행하는 경우, 교정 기록이 달리 기록되고 따라서 교정의 내용에 대한 유형 분류 및 계량 분석에 어려움이 생긴다. 따라서 ‘통제된 조작’의 원칙을 정하여 따르도록 하였다. 여기서 통제된 조작이란, 오류 하나당 한 번의 교정 조작을 수행하는 것이다.

이에 따라 앞서 예5)의 원시문은 다음 예7)과 같은 세 번의 통제된 조작으로 최종 교정문을 얻게 된다.

예7) 삽입: ‘고’뒤 빈칸

교체: 읍 > 습

삽입: 마침표

한 번의 조작으로 가능한 교정 작업을 오류별로 나누어 작업을 하기 때문에 작업 소요 시간은 매우 길어져서 기존의 3~4배가 드는 것으로 보고되었다. 게다가 이처럼 사람에 의한 인위적인 통제된 조작은 여러 사람에 의해 수행되면서 절대적인 ‘통제’가 사실상 어려우므로, 이를 효과적으로 통제할 수 있는 작업 플랫폼의 개선 또는 개발이 매우 필요한 것으로 보인다.

이상에서처럼, Kronoth에서 통제된 조작을 통해 구축된 말뭉치는 기존의 JSON 구조에 correction_info와 segment_correction_info 두 가지 속성값을 추가해 구조화한다. 전자는 오류 교정 양상, 즉 삽입, 교체, 삭제에 대한 정보이며, 후자는 원문과 교정문 간의 청크 대응 정보이다.

```

{
  "id": "MMRW2100000067.50.4",
  "original_form": "저 어제 못봐가지고",
  "form": "저 어제 못봐가지고",
  "corrected_form": "저 어제 못 봐 가지고",
  "correction_tag": [],
  "correction_info": [
    {
      "type": "space",
      "begin": 6,
      "end": 6,
      "value": " "
    },
    {
      "type": "space",
      "begin": 7,
      "end": 7,
      "value": " "
    }
  ],
  "segment_correction_info": [
    {
      "begin": 5,
      "end": 10,
      "original": "못봐가지고",
      "correct": "못 봐 가지고",
      "correction_info": [
        {
          "type": "space",
          "begin": 6,
          "end": 6,
          "value": " "
        },
        {
          "type": "space",
          "begin": 7,
          "end": 7,
          "value": " "
        }
      ]
    }
  ],
  "correction_tag": []
}

```

<표 9> 청크 단위 정렬 교정 말뭉치의 JSON 양식

3.2.2. 주요 오류 교정 양상 및 오류 유형

먼저, 약 15,000어절에 대한 교정 양상 결과는 다음의 표로 요약된다.

연번	교정 대상	교정 양상	교정 빈도	비율
1	문자열, 문장부호	삽입	6,303	40.0%
2		교체	2,963	19.0%
3		삭제	689	4.0%
4	띄어쓰기	삽입	5,881	37.0%
5		삭제	112	1.0%
			15,948	100.0%

<표 10> 교정 대상별 교정 양상 규모

먼저, 15,726어절(청크)의 원문 말뭉치에 대해 15,948회의 교정 작업이 수행되었는데,

이는 1어절(청크)당 1회를 약간 상회하는 교정 작업이 수행되었음을 의미한다. 여기에 교정이 아닌 주석 성격의 미등재어, 이모티콘 등 특수표현, 혐오와 차별 표현 등에 대한 마크업 작업까지 더하면 거의 어절(청크)당 1회 이상의 교정 또는 주석 작업이 이루어진 셈으로, 작업 양이 매우 큰 것을 알 수 있다.

다음으로, 문자열과 문장부호에 대한 교정은 전체의 62%, 띄어쓰기에 대한 교정은 약 38%를 차지한다. 문자열과 문장부호의 교정은 조작별로 삽입이 가장 많아서 전체의 40.0%, 교체 19.0%, 삭제 4.0%를 차지한다. 띄어쓰기는 누락된 빈칸의 삽입이 압도적으로 많아서 전체 교정 작업에서는 37%, 띄어쓰기 교정에서는 약 97.4%(112/5,881)를 차지하며, 불필요한 띄어쓰기의 삭제는 전체 교정 작업에서 1.0%로, 그 규모가 미미하다.

아래에 문자열을 중심으로 주요 교정 내용을 관찰 및 분석하고자 한다. 문자열의 교체는 교정 값 기준으로 토큰(token) 규모는 2,779개, 타입(type) 규모는 783개로 집계되어 타입별 중복도는 약 3.55회이다. 교정 값을 기준으로 고빈도 상위 30개의 목록과 그 원시 값을 정리해 보이면 <표 11>과 같다.

	원시 값	교정 값	교정 빈도	비율	누적 비율
1	잇	있	99	3.56%	3.56%
2	ㅇ	응	83	2.99%	6.55%
3	-구	고	81	2.91%	9.46%
4	-여, -읍, -웃	요.	80	2.88%	12.34%
5	-여	야	63	2.27%	14.61%
6	햇	했	61	2.20%	16.80%
7	-갯-	갯	56	2.02%	18.82%
8	-디, -뎡, -뎡, -대	데.	55	1.98%	20.80%
9	-어, -애	아.	51	1.84%	22.63%
10	-자나	잖아.	51	1.84%	24.47%
11	걸(로)	거	48	1.73%	26.20%
12	-였, -엇	였	47	1.69%	27.89%
13	-영, -삼, -아	어.	45	1.62%	29.51%
14	ㄹㅇ	레알	42	1.51%	31.02%
15	-믄, -만, 하면	면	42	1.51%	32.53%
16	마자	맞아.	34	1.22%	33.75%
17	-앗-, -였-	앗	26	0.94%	34.69%
18	왓	왔	26	0.94%	35.62%
19	바, 벼, 박	봐.	26	0.94%	36.56%
20	봣	봤	25	0.90%	37.46%
21	ㅇㅇ	응응,	25	0.90%	38.36%
22	조	좋	24	0.86%	39.22%
23	-넹, -노, -넹	네	23	0.83%	40.05%
24	-당, -따, -닥	다.	21	0.76%	40.81%
25	되, 대	돼	21	0.76%	41.56%
26	능, 느, 눈, 은	는	20	0.72%	42.28%
27	갓	갯	18	0.65%	42.93%
28	랏	랏	17	0.61%	43.54%
29	대, 돼	되	16	0.58%	44.12%
30	-치, -제, -쥐, -디	지	14	0.50%	44.62%

<표 11> 문자열의 교정 값 기준 고빈도 목록

어절(청크) 단위 대응이기 때문에 형태소 단위의 원시 값과 교정 값의 대응쌍 추출은 약간의 후처리를 거쳐야만 가능하며, <표 11>은 말뭉치의 결과물에 대한 후처리를 통해 교정 값 기준으로 고빈도의 원시 값을 추출하여 제시한 것이다. 교정 값 기준 고빈도 상위 30개의 빈도 합계는 문자열 오류의 44.62%로 중복도가 매우 높은 것을 알 수 있다. 이는 바꾸어 말하면 온라인 대화 텍스트에 나타나는 주된 오류 유형이 존재함을 보이며, 이에 대한 원시 값과 교정 값의 1:1 대응쌍을 추출하여 그 유형을 정리하는 것이 가능한 작업

이며, 언어학적으로나 언어공학적으로 모두 매우 필요한 일임을 시사한다.

3.2.3. 고빈도 미등재어 및 이모티콘 유형

온라인 대화 말뭉치는 일반 사용자의 사적 언어 특성상 사전에 실리지 않은 신어, 유행어 등이 많이 등장하는데, 이는 사전에 대응어가 없는 경우도 있고, 있다 하더라도 그 빈도가 높은 경우에는 일일이 교정하기보다 사전에 신규 등재하는 방법으로 형태소 분석 등 컴퓨터 처리 정확률을 높이는 것이 바람직하다. 본 연구에서는 2021년 과제에서 이러한 미등재어에 대해 별도의 주석과 결과물에 대한 분석을 통해 그 목록을 확보하였으며, 그중 고빈도 목록을 보이면 <표 12>와 같다.

연번	미등재어	품사	빈도
1	두구두구두구두구	부사	8
2	아아아아아아아앙	감탄사	8
3	터키아이스크림	명사	7
4	와아아아아아	감탄사	7
5	거짓부렁쟁이	명사	6
6	에어프라이기	명사	6
7	헤마토코쿠스	명사	6
8	브레이크 타임	구	6
9	꾸리꾸리하다	형용사	6
10	목살스테이크	명사	6
11	미니멀리스트	명사	6
12	모짜렐라치즈	명사	6
13	플레이리스트	명사	6
14	드라이브스루	명사	6
15	개피곤하다	형용사	5

<표 12> 고빈도 미등재어 목록

<표 12>를 보면 새로운 문물이나 개념과 함께 생겨난 신어(3, 6, 7, 8, 10, 11, 12, 14)가 많은데 주로 명사 또는 구 형식이고, 그 외 부사, 감탄사, 형용사 등 순서이다. 형용사는 상징부사의 ‘-하다’ 파생 유형(9)이 있는가 하면 ‘개-’, ‘꿀-’과 같은 생산성이 높은 신어 접두사에 의한 파생어가 적지 않다.

또한, 온라인 대화에는 감정을 표현하는 준언어적 장치로 이모티콘이 자주 사용되는데, 이 중에 부호형 이모티콘, 즉 이미지가 아닌 문장부호와 특수기호 등으로 조합된 이모티콘의 빈도가 매우 높다. 그 사용 양상도 매우 다양하여 동일한 부호의 N회 반복의 변이형, 다양한 부호와의 혼합형 등이 있으며 출현하는 위치 또한 일정하지 않아서 자연어 처

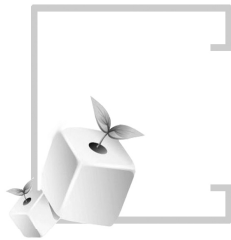
리에서 자동 인식의 어려움이 있을 것으로 예상된다. 또한, 이모티콘에 대한 정확한 분석은 화자의 의도나 감정 등을 분석하는 데 유용하게 쓰일 수 있다. 아래에 문장부호, 특수 기호, 자소 문자 등으로 이루어진 이모티콘 중에 일부 고빈도 목록을 보이면 <표 13>과 같다.

	이모티콘	빈도	비고
1	π	15,934	최다 58회 반복형
2	⌊	6,513	최다 13회 반복형
3	^^	2,557	
4	——	679	
5	;	661	최다 12회 반복형
6	π⌊	293	
7	⌊π	149	
8	〇^〇	116	
9	^^;	104	
10	><	90	

<표 13> 고빈도 부호형 이모티콘

이에 더해, 온라인 대화에는 입력의 편의를 위한 ‘ㅋㅋㅋ’와 같은 자소형 약어 형식의 비표준적 표기 또한 매우 다양하게 고빈도로 등장하며, ‘πππㅋ’처럼 부호형 이모티콘과 자소형 약어의 혼합형도 적지 않다.

이상으로, 어절(청크) 단위 정렬 교정 병렬 말뭉치 구축과 분석을 통해 주요 오류 유형을 살펴본 외에, 말뭉치 전체에서 추출한 고빈도 미등재어, 고빈도 부호형 이모티콘 목록 또한 일별하였다. 말풍선 대응에 더해 문장 대응, 청크 단위 대응 정보가 부착된 말뭉치는 언어학적 오류 유형 정보의 활용성, 어텐션(Attention)을 중시하는 최근 딥러닝 모델들에서는 중요한 힌트(hint) 데이터로 사용될 수 있다. 또한, 청크 단위 대응 말뭉치에 오류 유형에 대한 주석을 추가하는 경우, 현재의 문장 대응 주석 대비 구축 비용은 높을 수 있으나, 데이터의 활용도를 증대시킬 것을 기대해 볼 수 있겠다.



제 3 장

맞춤법 교정 말뭉치의 교정 지침 수립



1. 기본 지침과 지침 연구

온라인 대화 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 온라인 대화의 특성을 반영한 교정 지침을 수립하되, 언어학적 정밀성과 공학적 활용도를 고려한 맞춤법 교정 지침의 수립을 목표로 하였다. 교정의 수준은 한국어 형태소 분석기 적용이 가능한 정도를 목표로 한다. 특히 공공재로서의 말뭉치의 활용도를 높이기 위해 개인 정보 및 부적절한 표현의 범주를 엄격하게 분석하여 그 결과를 지침에 반영하였다. 맞춤법 교정 말뭉치의 기본 지침과 설계 방향은 다음과 같다.

1.1. 기본 지침

1.1.1. 한글 맞춤법 및 오타자 교정

온라인 대화 말뭉치에 나타나는 맞춤법 오류와 오타자를 교정한다. 이 경우, 온라인 대화 말뭉치에 특화되어 나타나는 표현과 표기의 처리 부분을 고려할 필요가 있다. 온라인 대화 말뭉치는 사용자 생산 콘텐츠(User Generated Content, UGC)이면서, 비격식적 글쓰기이기 때문에 문어와 구어의 특성을 동시에 지니고 있다. 온라인 대화 말뭉치에는 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 온라인 대화 말뭉치의 한글 맞춤법 및 오타자 교정 지침 설계 시에는 이러한 점을 고려하였다.

1.1.2. <우리말샘> 미등재어(외래어, 신어) 및 비표준어의 처리

온라인 대화 말뭉치에는 다양한 유형의 <우리말샘> 미등재어와 비표준어가 빈번하게 나타난다. 미등재어의 경우, 외래어, 신어, 구어, 방언 등으로 나누어 처리하도록 지침을 수립하였다. 아울러 미등재어의 경우, 작업을 하는 과정에서 OoV(Out of Vocabulary) 목록을 구축하여 관리할 수 있도록 하였다.

1.1.3. 개인 정보 및 부적절한 표현에 대한 엄격한 처리

개인 정보 및 부적절한 표현의 처리는 공공재로서의 말뭉치의 성격을 제고하기 위해 반

드시 필요한 작업이다. 앞서 언급한 바와 같이 사업의 전체적인 유기성을 고려해 2019년에 추진한 국립국어원 온라인 대화 자료 수집 및 말뭉치 구축 사업과 2021년에 수행한 맞춤법 교정 말뭉치 연구 분석 사업에서 제시한 범주에 따라 이 사업의 대상 말뭉치에 나타난 개인 정보와 부적절한 표현을 철저히 제거하여 비식별화하였다.

1.2. 지침 연구

이 절에서는 지침을 구축하기 위한 본 사업팀의 지침 연구 설계 단계별 세부 내용에 관해 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 수립, 2021년 작성된 지침의 개선 및 보완 작업 등을 위한 연구를 다음과 같이 진행하였다.

1.2.1. 선행 연구 검토

본 사업의 대상인 온라인 대화 말뭉치는 사용자 생산 콘텐츠로, 규범성의 결핍이 이러한 말뭉치 유형이 가지고 있는 중요한 특성이다. 기존의 전통적 의미의 텍스트는 일반적으로 생산 과정에서 전문가의 교정과 교열 작업이 선행되지만 사용자 생산 콘텐츠는 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 온라인 대화 말뭉치의 맞춤법 교정을 위한 지침 연구 단계를 사업 수행의 첫 번째 단계로 두었다.

1.2.2. 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.

<맞춤법 교정 말뭉치 기본 오류 유형>

[유형1] 한글 맞춤법 미준수형

[유형2] <우리말샘> 미등재어형(외래어, 신어 등)

[유형3] 문장부호

[유형4] 특수 표현

[유형5] 방언형

[유형6] 개인 정보 및 부적절한 표현

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, 검둥), 신어(예: 깨발랄), 방언(예: 개안타)과 전달 효과를 고려하여 개인적으로 사용하는 표현들이다(예: 응, 응응응).

1.2.3. 2021년 작성된 지침의 개선 및 보완 작업 수립

올해 사업의 맞춤법 교정 지침은 2021년 수행된 맞춤법 교정 병렬 말뭉치 연구 분석의 지침을 따르되, 기존 지침의 각 유형별 내용을 보다 정밀하게 정리하였으며, 자모형 문자, 다양한 유형의 이모티콘, 의미불명어 등 세 가지 특수 유형에 대해서 새롭게 지침을 수립하여 개선하고자 하였다. 이번 사업을 통해 중점적으로 개선한 부분은 다음과 같다.

첫째, 띄어쓰기, 오타자 등에 대한 교정 수준 및 개선 방향 제시

관형사, 대명사, 부정사, 의존명사 등을 포함한 고빈도 띄어쓰기 오류에 대한 목록을 정성적으로 작성한 후 자동 교정의 형식으로 반영하였으며, 이를 추후 검토하는 방식으로 작업을 진행하였다. 고빈도 띄어쓰기 오류 유형은 아래 그림과 같이 구글 문서를 통해 공유하여 작업자와 검토자가 확인하여 작업 및 검토할 수 있도록 하였다.

대분류	중분류	원 어절	교정 어절(오류)	정확한 교정	비고
일괄 교정	띄어쓰기 - 구	담달	담달	담 달	
일괄 교정	띄어쓰기 - 구	버스기사	버스기사	버스 기사	
일괄 교정	띄어쓰기 - 구	폐소공포증	폐소공포증	폐소 공포증	
일괄 교정	띄어쓰기 - 구	영상매체	영상매체	영상 매체	
일괄 교정	띄어쓰기 - 구	봉사활동	봉사활동	봉사 활동	
일괄 교정	띄어쓰기 - 구	영상통화	영상통화	영상 통화	
일괄 교정	띄어쓰기 - 구	직장생활	직장생활	직장 생활	
일괄 교정	띄어쓰기 - 구	영업사원	영업사원	영업 사원	
일괄 교정	띄어쓰기 - 구	룸패딩	룸패딩	룸 패딩	
일괄 교정	띄어쓰기 - 단어	까먹어	까 먹어	까먹어	까먹다 001-004
일괄 교정	띄어쓰기 - 단어	그만 돌	그만 돌	그만돌	그만두다 001, 002
일괄 교정	띄어쓰기 - 단어	집밥/집 밥	집 밥	집밥	집밥001
일괄 교정	띄어쓰기 - 단어	논리정연한	논리 정연한	논리정연한	
일괄 교정	띄어쓰기 - 단어	떡볶이 집	떡볶이 집	떡볶이집	집 009 참고
일괄 교정	띄어쓰기 - 단어	불꽃놀이	불꽃 놀이	불꽃놀이	불꽃놀이 001
일괄 교정	띄어쓰기 - 단어	비행기 표	비행기 표	비행기표	비행기표 001
일괄 교정	띄어쓰기 - 동사 + 동사 구성		타먹-(나/여/음/있어/고)	타 먹-	
일괄 교정	띄어쓰기 - 문장부호 뒤의 띄어쓰기 누락		.ㅋㅋ/.쿠기...	띄어쓰기	
맥락 교정	띄어쓰기 - 부사 '못' 뒤		못하/못해/못했/못한/못할.../못보-/못만	띄어쓰기	동사/형용사/보조동사 '못하다'와 구분이 필요함.
맥락 교정	띄어쓰기 - 부사 '안' 뒤		안돼/안됐/안되/안됨/안했/안하/안해...	띄어쓰기	동사/형용사 '안되다'와 구분이 필요함.
맥락 교정	띄어쓰기 - 부사 '잘' 뒤		잘자/잘 잤나요/잘 잤어/잘바르코/잘노	띄어쓰기	잘생기다/잘하다... 등 제외해야 함

[그림 13] 주의해야 할 띄어쓰기 목록(구글 문서)

’22년 맞춤법 교정 말뭉치 사업에서는 아래 그림과 같이 ‘의미불명어’를 별도로 수집하였다. 오타자나 사용자의 실수로 잘못 입력된 명백한 오류인 경우에는 모두 교정하되, 원래 형태를 확신할 수 없는 형태에 대해서는 과도하게 추측하여 교정하지 않고, 의미불명어로 분류한 후 수집하여 그 정보를 반영하였다.

구분	연번	형태	구분	연번	형태
의미 불명	1	2 f 2 — ㄱ 나	의미 불명	31	그이뱅방에
의미 불명	2	88목처	의미 불명	32	그구늬나
의미 불명	3	a2	의미 불명	33	금굴첼로
의미 불명	4	SSG	의미 불명	34	금뿡
의미 불명	5	가배	의미 불명	35	기
의미 불명	6	각전	의미 불명	36	기문
의미 불명	7	간게	의미 불명	37	기행이
의미 불명	8	갓따따	의미 불명	38	길뿡
의미 불명	9	갓케	의미 불명	39	깨바가지
의미 불명	10	강츄츄	의미 불명	40	결뿡
의미 불명	11	개인	의미 불명	41	꼬방이
의미 불명	12	갠통	의미 불명	42	꽃음
의미 불명	13	경보하데끼	의미 불명	43	꾸까
의미 불명	14	고고리	의미 불명	44	꾸지꾸지
의미 불명	15	고웁	의미 불명	45	ㄱ ㄱ
의미 불명	16	고웁	의미 불명	46	나니가
의미 불명	17	곰도	의미 불명	47	나라캉
의미 불명	18	공게스앵	의미 불명	48	내배고
의미 불명	19	공뽀	의미 불명	49	냐
의미 불명	20	공아동아	의미 불명	50	냐미
의미 불명	21	공아지	의미 불명	51	남
의미 불명	22	곽도	의미 불명	52	너티너티
의미 불명	23	관대회차	의미 불명	53	노넬놈
의미 불명	24	구두미	의미 불명	54	누누
의미 불명	25	구람	의미 불명	55	누누습
의미 불명	26	구지비	의미 불명	56	눈뜨스
의미 불명	27	국고거거꺽꺽	의미 불명	57	눈새캐캐캐시리플렛아
의미 불명	28	귀엽비보	의미 불명	58	니니
의미 불명	29	그러누	의미 불명	59	니빠
의미 불명	30	그런뎃니	의미 불명	60	니팅모블

[그림 14] 수집한 의미불명어 목록 예시

둘째, 우리말샘 미등재어(외래어, 신어 등) 및 비표준어 처리 방안 정교화

외래어, 신어 등의 미등재어는 가급적 교정하지 않고, OoV로 표시한 다음, 해당 목록을 아래 그림과 같이 추출하여 정리하였다.

구분	연번	형태	구분	연번	형태
OoV	1	n뽕	OoV	31	개관찮다
OoV	2	OMG	OoV	32	개귀엽다
OoV	3	RGRG	OoV	33	개귀찮다
OoV	4	st	OoV	34	개극혐
OoV	5	TMI	OoV	35	개깜놀하다
OoV	6	가나슈	OoV	36	개꿀
OoV	7	가마보꼬	OoV	37	개꿀개꿀
OoV	8	가방순이	OoV	38	개꿀따리개꿀따
OoV	9	가오	OoV	39	개꿀딱
OoV	10	가오피	OoV	40	개꿀맛
OoV	11	가즈아	OoV	41	개꿀잼
OoV	12	가차	OoV	42	개냥이
OoV	13	가통문	OoV	43	개노맛
OoV	14	각	OoV	44	개다행
OoV	15	간계밥	OoV	45	개당황
OoV	16	간접	OoV	46	개대박
OoV	17	간지나다	OoV	47	개대충하다
OoV	18	감튀	OoV	48	개답다
OoV	19	갑	OoV	49	개땡큐
OoV	20	갑분-	OoV	50	개마르다
OoV	21	갑분잡채	OoV	51	개막하다
OoV	22	갑분주	OoV	52	개많다
OoV	23	갓갓맛	OoV	53	개많이
OoV	24	갓뚜기	OoV	54	개맛없다
OoV	25	갓벽	OoV	55	개맛있다
OoV	26	갓성비	OoV	56	개망하다
OoV	27	갓직히	OoV	57	개모차
OoV	28	개	OoV	58	개무리
OoV	29	개간지	OoV	59	개무섭다
OoV	30	개갓먹방	OoV	60	개무시하다

[그림 15] 수집한 OoV 목록 예시

비표준 형태의 경우 <우리말샘>의 등재 표제어를 기준으로 교정하며, 외래어 표기법, 국립국어원의 한국어 어문규범 용례 찾기 등을 활용하여 정해진 규범형으로 교정하였다. 그리고 <우리말샘>에 미등재된 접사, 어미와 같이 띄어쓰기에 영향을 미칠 수 있는 요소에 대한 처리 방안을 지침에 마련하여 정리하였다. ([지침2] 참조)

셋째, 온라인 환경에서 나타나는 특수 표현 등에 대한 처리 방안 보완

자모음 문자 연쇄, 로마자·숫자 혼용, 문자 모양의 유사성에 기반한 단어 사용 등과 같은 의도적인 표기 변형이나 줄임말 등에 대한 구체적인 처리 방안을 지침에 마련하여 정리하였다. 그리고 자모음 문자 연쇄에 대한 처리 방안을 개선하였으며, 이미지, 자모와 문장부호 혼합형 등 다양한 유형의 이모티콘에 대한 처리 방안을 새롭게 수립하여 지침에

정리하였다. ([지침4] 참조)

그리고 구어에서 흔히 나타나는 단어 내 축약 및 모음 삽입 등을 통한 확장 등 단어 내부의 형태 변화에 대한 처리 방안을 마련하였으며,([지침1] 참조) 자모로만 이루어진 표현은 원 문자로 복원하였다. 아울러 “ㅋㅋ, ㅏㅏ, ㄷㄷ” 등과 같이 동일 자모가 무한히 덧붙을 수 있는 의성의태어의 처리 방안에 대한 개선안을 제시하였다. 마지막으로 기타 식별 불가능한 문자열 등에 대한 처리 방안을 새롭게 마련하여 지침에 정리하였다. ([지침4] 참조)

넷째, 개인 정보 및 부적절한 표현의 비식별화

개인 정보는 국립국어원 사업의 전체적 유기성을 고려하여 국어원 온라인 대화 말뭉치(버전 1.0) 구축 시 제시된 범주에 따라 비식별화하였다. 그리고 부적절한 표현은 연구진 협의와 전문가 자문의 단계 등을 거쳐 철저하게 판단 작업을 진행하였다. ([지침6] 참조)

비식별화 지침 수립 시, 혐오 차별 표현의 경우, 작업자와 검토자의 이해를 돕기 위해 유형별로 구체적인 사례를 문맥과 함께 제시하였으며, 혐오 및 차별 표현의 다양한 유형을 풍부히 수집하여 지침의 마지막에 [부록]으로 실었다. (지침 중 [부록] 참조)

2. 맞춤법 교정 말뭉치의 교정 지침

이 장에서는 본 사업에서 온라인 대화의 맞춤법 교정 말뭉치를 구축하기 위해 적용한 지침에 대해 기술하고자 한다.

I. 사업의 목적과 교정의 방향

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] 문장부호의 처리

[지침4] 특수 표현의 유형 분류와 유형별 처리

[지침5] 방언형의 처리

[지침6] 개인 정보 및 부적절한 표현(욕설, 혐오 표현 등)의 비식별화 처리

<부록> 욕설, 혐오 차별 표현의 비식별화 사례

1. 사업의 목적과 교정의 수준

이 사업의 목적은 온라인 대화의 맞춤법 교정 말뭉치를 병렬 말뭉치로 가공함으로써, 자동 형태소 분석, 기계 번역 등 한국어 처리 도구의 온라인 대화 분석 효율을 높이고, 온라인 대화의 언어학적 연구에 기여하는 말뭉치를 구축하는 것이다. 또한 구어, 문어와 다른 온라인 대화의 특수성을 살린 교정 병렬 말뭉치를 구축함으로써, 온라인 대화 연구의 기초 자료를 제공하고자 한다.

이러한 목적을 고려한 교정의 수준은 다음과 같다.

첫째, 완벽한 문어나 구어의 수준을 지향하기보다 제3의 매체로서의 온라인 대화의 특수성을 살리는 수준에서 교정 말뭉치를 구축한다.

둘째, 온라인 대화의 자연어 처리에서, 자동 형태소 분석기, 기계 번역 등의 한국어 처리 도구의 효율에 기여하는 분석의 수준을 모색하되, 규범 사전이나 교과서 수준의 정제를 지향하지 않는다.

셋째, 이론 연구의 관점에서, 원시 말뭉치를 대상으로 했을 때 다소 어려움이 있는, 어휘 빈도의 추출, 형태/통사/담화 층위의 연구를 용이하게 하는 말뭉치로 가공한다.

2. 기본 원칙

2.1. 유형별 지침

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

2.2. 기본 작업 과정

교정 작업 시 작업 도구는 이르테크의 크로노스(Kronoth)와 엑셀(Excel)을 이용한다. 작업의 효율을 위해 MS Excel을 사용할 경우, ‘최종 교정 문장’ 칼럼에 직접 교정하고, OoV(미등재어), 의미불명어, 자소형 이모티콘과 비식별화 대상(혐오 표현, 개인 정보 등)은 교정문의 우측에 있는 해당 칼럼에 복사해서 붙인다.

<작업 화면 예시>

회사 ID	원 문장(개행 문자 삭제 X)	최종 작업 문장(일괄 교정 Ver 6.0)	OoV	의미 불명어	소형 이모티콘	작업자 메모	검토자 메모	현오 및 지	전화번호	계좌번호	이름	주소
1	물치실 전화해서	물치실 전화해서	물치실									
1	그냥 맛있음 평맛	그냥 맛있음. 평맛	평맛									
1	미조미조	미조미조	미조미조									
1	그래서 메시가 공뽕주면	그래서 메시가 공뽕 주면	공뽕									
2	2+2-7나...	2+2-7나...		2+2-7나...								
1	엄마~	엄마~.										
2	name3됐다?	name3 됐나? ^^			^^							
2	개객끼...	개객끼...						개객끼...				
1	2019.10.31. 사무장이 이상한 가면 주고 쓰라고 함.	2019년 10월 31일 사무장이 이상한 가면 주고 쓰라고 함. 1,818원.						1818				

2.3. 병렬 구조를 고려한 단위 설정과 교정

맞춤법 교정 병렬 말뭉치의 병렬 구조는 기본적으로 21년 온라인 대화 말뭉치의 상당 부분을 차지하고 있는 온라인 대화(카카오톡)의 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 3] 문장부호 지침’을 따른다.

3. 2021년 지침과 2022년 지침의 비교

2021년 지침과 비교하였을 때, 2022년 지침은 여러 사항에 대해 지침이 보완되거나 추가되었는데, 구체적인 목록을 보이면 다음과 같다. 각 항목에 대한 자세한 사항은 각 목록에서 표지가 부착된 해당 지침을 참고하기 바란다.(1건의 경우 2021년 지침에서 수정된 경우가 있다.)

3.1. 2021년 지침에서 보완된 경우

- [22년 보완] 어절이 지나치게 길어지는 경우의 띄어쓰기
- [22년 보완] 신조어 파생어의 띄어쓰기
- [22년 보완] 명사+명사의 구 구성과 합성 명사의 띄어쓰기
- [22년 보완] <우리말샘>에 등재된 유형에 대한 교정 사항
- [22년 보완] 구어체 말뭉치 통용 방언 목록과 교정
- [22년 보완] 널리 사용되는 틀린 표현, 문법 오류 등의 교정
- [22년 보완] 의미 불명 표현의 처리
- [22년 보완] 미등재어의 처리
- [22년 보완] 자소형 이모티콘의 교정

[22년 보완] 숫자가 문자열과 결합한 경우의 교정(<우리말샘>에 등재된 어휘로 대체)

[22년 보완] 방언 종결어미의 처리

3.2. 2021년 지침에 추가된 경우

[22년 추가] 보조용언이 거듭 나타나는 경우의 띄어쓰기

[22년 추가] 기타 띄어쓰기의 허용 지침과 주의 사항

[22년 추가] ‘-고 하-’가 생략되어 녹아 붙은 꼴의 복원 여부

[22년 추가] 인명, 지명, 영화 제목 등과 같은 고유명의 OoV 목록

[22년 추가] 의존명사 줄의 방언(경상)

[22년 추가] <우리말샘>에서 대응 방언을 특정하기 어려운 경우의 처리(통용 방언으로 처리)

[22년 추가] 유튜버 이름의 비식별화

[22년 추가] 인명 중 이름만 비식별화된 경우의 비식별화 작업

3.3. 2021년 지침에서 수정된 경우

[22년 수정] 이미지로 된 이모티콘의 처리 관련

[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 이루어지므로, 맞춤법 검사기의 오교정과 과교정에 유의하여야 한다.

온라인 대화 원시 말뭉치	자동 검사기 처리	최종 작업(수작업 교정)
아이구 힘들었겠다	아이고 힘들었겠다.	아이구 힘들었겠다.
아침에8시에나와서두유만먹어가지구	아침에 8시에 나와서 두유만 먹어서	아침에 8시에 나와서 두유만 먹어 가지고
끝났웅?	끝났어?	끝났어?
오늘할수있는일은곳	오늘할수있는일은곳	오늘 할 수 있는 일은 끝
적응잘하려는지거경되죽겠다.	적응잘하려는지거경되죽겠다.	적응 잘하려는지 걱정돼 죽겠다.

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

예) 끝났웅? → 끝났어?

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

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

- 3) 맞춤법 검사기에서 구어 표현을 다른 단어로 대체하는 등 과하게 수정한 경우, 원래의 단어로 복원한 후 표준형으로 수정한다.

예) <1> 먹어가지구 → <2> 먹어서 → <3> 먹어 가지고 / 먹어가지고

<1> 아이구 → <2> 아이고 → <3> 아이구

※ 여기서 <2>는 과교정, <3>은 복원한 형태를 뜻함.

2. 띄어쓰기

1) 한글 맞춤법의 띄어쓰기 규정을 따른다.

2) 보조용언의 띄어쓰기

(1) 보조용언은 맞춤법의 규정에 따라, 띄어 쓸을 원칙으로 하되, 경우에 따라 붙여 쓸도 허용한다.

이는 현재의 맞춤법 규정을 따르는 동시에, 현재 형태소 분석기, 맞춤법 검사기 등의 **한국어 처리 도구의 정확도가 보조용언 띄어쓰기에 영향을 받지 않는 것**을 고려한 것이다.

예1) 가보고 싶어(○), 가 보고 싶어(○)

예2) 가보니(○), 가 보니(○)

(2) **[22년 보완]¹⁰⁾ 단, 어절이 4음절 이상으로 지나치게 길어지는 경우, 특히 ‘명사+하다/되다’의 용언의 경우는 띄어 쓴다.**

예1) 이야기해봐야겠구먼(x) → 이야기해 봐야겠구먼(○)

예2) 공부해봐.(x) → 공부해 봐.(○)

(3) 맞춤법 규정에 따라 **‘-어지다’, ‘-어하다’의 구문의 경우는 붙여 써야 한다.**

예) (낙서가) 지워진다, (아기를) 예뻐한다, (놀고) 싶어한다

(4) **[22년 추가]¹¹⁾ 보조용언이 거듭 나타나는 경우는 앞의 보조용언만을 붙여 쓸 수 있다.**

예1) 적어 둘 만하다. / 적어둘 만하다.

예2) 해 와 줘. / 해와 줘.

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

예) 가야겠어요, 가봐야겠어요, ...

4) **[22년 보완]¹²⁾ 신조어 파생어의 경우 붙여 쓴다. 이들은 <우리말샘>에 등재되지 않은 새로운 의미의 접사가 붙어 만들어진 단어들로 다음과 같은 것들이 있다.**

ㄱ. 갓- 예) 갓동원, 갓뚜기, 갓보검, ...

ㄴ. 꿀- 예) 꿀나은, 꿀잼, 꿀목소리, ...

ㄷ. 핵- 예) 핵공감, 핵노잼, 핵인싸, ...

ㄹ. 개- 예) 개간지, 개꿀잼, 개귀엽다, 개좋아, ...

ㅁ. 캐- 예) 캐감동

<주의> 다음처럼 접사가 아닌 부사처럼 사용된 ‘개/캐-’는 띄어 쓴다.

예) 시험 개 잘 봤네, 개 푹 자 버렸다

ㅂ. 급- 예) 급-: 급만남, 급생각해서, 급어필했지...

<주의> 다음처럼 접사가 아닌 부사처럼 사용된 ‘급-’은 띄어 쓴다.

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

10) 자료의 통일을 위한 예시를 포함한 명확한 기준 추가한 것이다.

11) 보조용언의 띄어쓰기 통일을 위한 어문 규범의 내용을 추가한 것이다.

12) 작업자의 이해를 돕기 위해 작업 과정에서 수집된 예를 추가한 것이다.

스. -각(角)

<우리말샘>에 각(角)은 “「의존 명사」 「030」(명사 뒤에 쓰이거나 어미 ‘-을’ 뒤에 쓰여)) 어떤 일이 일어날 조짐이나 분위기.”로 등재되어 있으므로 **띄어 씀**.

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, ...

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, 땅땅땅, 그래그래, 그치그치, 그쵸그쵸, ...

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

예) 기대기대, 인정인정, ...

(3) 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.

(4) 용언의 반복은 띄어 쓰고, 가운데에 쉼표를 넣는다.

예) 알아알아 → 알아, 알아.

조아조아 → 좋아, 좋아.

(5) 대명사 등을 반복한 경우는 띄어 쓰고, 가운데에 쉼표를 넣지 않는다.

예) 어디 어디, 그거 그거, 뭐 뭐, ...

6) 전문용어의 띄어쓰기

전문용어의 띄어쓰기는 허용 규정에 따라 띄어 쓰는 것과 붙여 쓰는 것 모두 허용한다. (자동 교정된 결과를 반영하는 것도 허용함.)

예) 영상통화(○), 영상 통화(○)

※<우리말샘>에 표제어로 ‘영상^통화’가 등재되어 있음.

전문용어의 기준은 <우리말샘>의 전문어 표지 부착 여부에 따른다.

7) [22년 보완] 명사+명사의 구 구성과 합성 명사의 띄어쓰기

전문용어가 아닌 일반 명사의 연쇄는 띄어 쓰는 것이 원칙이나, 명사+명사의 구 구성과 합성 명사의 경계는 자유로이 띄거나 붙일 수 있다. 단 인접한 맥락에서의 일관성을 고려하는 방향으로 한다.¹³⁾

13) 이는 명사+명사의 띄어쓰기가 사전 간의 상이성, <우리말샘>의 역동성으로 인해 쟁점이 있고, 실제 언어 처리 도

예) 단체 사진, 남자 친구, 코인 노래방, 커피 향, 감상 평 등

아울러 ‘명사+명사’ 합성어의 경우 <우리말샘> 등재 여부와 관련 없이, 역시 붙여 쓰는 것과 띄어 쓰는 것 모두 허용한다. (이유는 위의 (1)과 동일)

예1) 등재 유형 : 양꼬치(○)/양 꼬치(○), 염통구이(○)/염통 구이(○)

예2) 미등재 유형 : 염통꼬치(○)/염통 꼬치(○), 순대꼬치(○)/ 순대 꼬치(○)

예3) 미등재 유형 : 양꼬치집(○)/양꼬치 집(○)

8) [22년 추가]¹⁴⁾ 기타 띄어쓰기의 허용 지침과 주의 사항

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

예1) 알바하다(○)/알바 하다(○), 공부시키다(○)/공부 시키다(○), 충격받다(○)/충격 받다(○)

예2) 리트윗하다(○)/리트윗 하다(○)

9) 비식별화 기호가 포함된 경우

비식별화 기호가 들어가 있는 경우라도 다음과 같이 조사, 의존명사와 단어를 구별하여 띄어쓰기한 다.

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

10) [22년 수정]¹⁵⁾ 이모티콘 중에 아래의 예문처럼 이미지로 된 이모티콘은 별도의 처리가 없이 그대로 둔다.(일괄 처리)

예) 딸기□랑 바나나□사서 집으로...

11) ‘+, =’ 등의 부호가 포함된 경우 앞뒤를 한 칸 띄어서 쓴다.

예) 보습효과+미백.주름개선+피부진정+메이크업베이스 이모든게 한번에 가능하다고

→ 보습효과 + 미백, 주름 개선 + 피부 진정 + 메이크업 베이스. 이 모든 게 한 번에 가능하다고

3. 구어형(축약형 포함)과 온라인 대화 비표준형의 교정

1) 구어 표현과 비표준형의 처리

구어 표현이나 온라인 대화에서 많이 나타나는 다음과 같은 비표준형은 괄호와 같이 교정한다

(1) 음운이 탈락된 경우

예) 마이→ 많이

암꺼나→ 아무거나

개안아→ 괜찮아

(2) 음운이 첨가된 경우

구의 효율 향상과도 무관하다는 데 주된 이유가 있다.

14) 작업의 효율성을 위해 띄어쓰기의 허용 규정을 적용한 것이다.

15) 일괄 작업이 가능한 경우를 명시하기 위해 수정한 것이며, 이는 작업의 효율성을 위한 것이다.

예) 맛있어잉 → 맛있어, 가야징 → 가야지, 너동 → 너도, 그럴깁 → 그렇게
 고마워염 → 고마워요
 제법 해웃 → 제법 해요, 싫어웃 → 싫어요
 줄일려고 → 줄이려고
 싫으다 → 싫다, 꼬옥 → 꼭
 감좌 → 감사, ...

(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) 썩돈 → 생돈, 사주께 → 사줄께, 짤라야 → 잘라야, 할꼬얌 → 할 거야, 체험하구 → 체험하고, 별루 → 별로, 지대로 → 제대로, 이뿌다 → 이쁘다, 슬푸당 → 슬프다, 귀웁다 → 귀엽다, 모야 → 뭘야, ...

2) 군더더기 표현

다음의 ‘달, 날’과 같이 구어 특유의 군더더기 표현은 교정하지 않는다.(띄어쓰기와 ‘날’, ‘달’의 삭제 주의)

예1) 1월달, 2월달, 3월달 월요일 출근은 죽음인데, ...

예2) 1일날, 2일날, 토요일날, 일요일날도 일하는데!, ...

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 교정

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

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

ㄷ. 방언형 ‘니’의 경우는 교정이 필요하므로 [지침5]에 기술된 유형에 따라 교정을 한다.(참고: 지침5의 ‘니’ 관련 기술)

예) 방언형 ‘니’는 → 너는(‘니’를 교정하지 않는 경우는 ‘니가(네가)’ 또는 ‘니 친구(네 친구)’일 때뿐임. 따라서 니를(×), 니는(×), 니처럼(×)으로 함.)

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

(2) ‘거’의 처리(‘거’는 우리말샘 등재어임을 고려)

ㄱ. ‘ㄹ’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 교정하지 않는다. 단, ‘모’는 ‘뭉’로 교정한다.

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?(찾아싸아?), 싶 어지나봉가.(싶어지나 봐.), ….

※ 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.).

예1) 효진쓰 → name1쓰

예2) 선호짱 → name1짱

의미 없이 붙는 ‘-쓰’의 경우 위의 예와 같은 고유명사에 붙는 경우 이외는 모두 생략한다.

예) 위험쓰 → 위험, 꿀잤쓰 → 꿀잤, 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 교정하지 않는다. 그리고 <우리말샘>에 없더라도 고빈도로 쓰이고 자주 쓴다고 판단될 경우, 별도의 목록을 관리하고 교정하지 않는다.

(1) [22년 보완]¹⁶⁾ <우리말샘>에 등재된 유형은 교정하지 않는다. (가나다순)

예) 갓다, 갠, 그쵸, 그지, 그쵸, 그치, 근데, 글로, 글서, 글치, 글쿤, 글쿤요, 깜놀, 넬, 념, 념나, 담, 맘, 섬, 스탈, 애, 얘기, 어캐, 어케, 얼집, 여따가, 왜캐, 왜케, 웰캐, 웰케, 요새, 일로, 일욜, 잼, 재밌다, 절로, 젤, 줌, 첨, 큰맘, …

(2) [22년 보완]¹⁷⁾ <우리말샘>에 방언으로 되어 있지만, 실제 통용되는 구어이고 널리 자주 쓰이는 유형을 ‘구어체 말뭉치 통용 방언’으로 지칭하고, 교정하지 않는다. ‘구어체 말뭉치 통용 방언’은 구글에서 해당 방언형을 “ ” 안에 넣어 검색한 결과, 같은 의미로 해당 어형이 만 개 이상 검색되는 경우로 한정한다(※ 이하 7-1)-(3) 구어체 통용 방언형 관리 지침 참조).

16) 작업 중 수집된 예시를 추가한 것이다.

17) 통용 방언형에 대한 규정을 명확히 하고, 작업 중 수집된 예시를 추가한 것이다.

예) 강, 그니까, 이케, 냅두다, 함, 여튼, 코자다...

(3) [22년 수정]¹⁸⁾ <우리말샘>에 등재된 준말 외에는 모두 교정한다.

예1) 짐(지금), 어캄(어떡함), 어카지(어떡하지), 왜냐면(왜냐하면), 그치만(그렇지만)

예2) 여깃다(여기 있다), 어딴다(어디 있다)

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

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

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

(5) [22년 추가]¹⁹⁾ ‘-고 하-’가 생략되어 놓아 붙은 꼴은 복원하지 않고 그대로 둔다.

예) 먹으랬는데 : ‘먹으라고 하였는데’로 교정하지 않고 그대로 둔다.

[비교] 비슷한 꼴로 “(내가) 먹으려고 했는데”의 줄인 말인 “먹을랬는데”가 있는데, 이 역시 교정하지 않고 그대로 둔다(르도 빼지 않음).

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

예) 밖에서 사 먹음 국밥도 7천 원씩 하더라고요.

(7) 3중 모음의 축약은 축약 전의 형태로 교정한다.

예) 바껴서 → 바뀌어서(○), 사겨서 → 사귀어서(○)...

4. 감탄사와 의성의태어의 교정

1) 감탄사와 의성의태어의 경우 온라인 대화의 감정 표현의 다양성과 특수성을 고려하여 다양한 변이형이 존재하고, <우리말샘>에 소극적으로 등재되어 있으므로, 교정하지 않는다.

예) 냅, 냅, 아앗, 와아, 크으, 호오, 아우, 오웅, 와아, 우와와, 우앙, 웅, 웅웅, 까짓거, 까아, 까아아아아아, 까아아아아악, 까아아아악, 까아앙, 까악, 까오, ... 네에, 네에네, 네에에, 네, 네에, ...아아니, 아아아, 아아아아, 아아아아아, 아아아아아악, 아아앗, 에그마, 에또, 에라이, 에에에, 카하, 크으, 크크, 크흐, 크흑...

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

예) 오먼아 → 어머니, ㅏㅓㅓ → 우와

※ 욕설 및 혐오표현: 아쌔/아씨

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 교정하지 않는다(예: 괜찮아(○), 괜찮아요(○), 괜찮다(×)).

예) 아래와 같은 문맥에서 밑줄 친 ‘괜찮-’의 처리

[문맥] A: 병원 어디 갈지 고민 중

B: 응 심해?

18) 작업 결과물의 통일성을 기하기 위해 <우리말샘>을 기준으로 삼기로 한 것이다.

19) 작업 중에 수집된 예시를 추가한 것이다.

A: 오웬
오늘 계속 나서
B: 괜찬?

[교정 방법]

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 교정하지 않는다.

6. 외래어/외국어의 교정

<우리말샘>의 등재어를 기준으로 규범에 맞게 교정한다. 다만, 규범 표기가 미확정인 경우라도 일관성을 위해 현재 등재된 형태를 따른다.

예1) 요거트 → 요구르트

예2) 패스트리(pastry) → 규범 표기는 미확정이지만, 패스트리로 교정함.

2) 비규범형일지라도 고유명은 교정하지 않는다.

예) 너무 보챌 때는 끌라미엘 하고

요새 줘비스 유명하잖아.

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

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

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

예) 그웨잇(온라인 대화 60만에서 9회 출현) … 그웨잇(구글에서 50만 회 출현)
 스투핏… 스투핏(구글에서 23만 회 출현)
 땡큐
 프로틴

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

예) 외래어 고유명사 교정 사례

원문	교정문	교정 근거
스타듀밸리	스타듀밸리	지침을 따라 그대로 둠
스타듀 벨리	스타듀 벨리	지침을 따라 그대로 둠
스타듀밸리	스타듀 벨리	자동 교정 결과
스타듀 벨리	스타듀 벨리	자동 교정 결과

6) 응답 표현의 처리: 감탄사 처리 원칙에 따라 교정하지 않는다.

예) 노, 놉, 예스, 노우, 노놉, 노우노우, 노노, 예스예스, …

7) 외래어/외국어를 순화어 또는 대응되는 어휘로 교정하지 않는다.

예1) 타코 와사비 ('와사비'를 '고추냉이'로 교정하지 않음.)

예2) 기스가 찌네 ('기스'를 '흠, 흠집'으로 교정하지 않음.)

예3) 익스펜시브 푸드 ('비싼 음식'으로 교정하지 않음.)

8) 특정 음식명을 줄여서 부르는 경우 형태를 복원하지 않고 그대로 둔다.

예) 까르보 ('까르보나라'를 줄인 경우) => OoV에 추가

7. 고유명사의 띄어쓰기, 맞춤법 교정

1) 상호명, 상품명 등의 고유명사는 맞춤법 검사기 결과를 반영하고 더 이상의 교정을 하지 않는다²⁰⁾.

2) 단, 인접 문맥에 한하여 일관성을 고려하여 교정하며, 문맥의 고려 범위는 다음과 같다.

(1) 한 문맥 안에서 표기나 띄어쓰기가 일치되지 않은 경우는 일관성을 고려하는 선에서 교정한다.

예) 스타듀 벨리/스타듀 벨리 → 스타듀 벨리로 교정(통일함)

(2) 일부 외래어 표기 지침과 상충되는 경우 또는 규범형 표기로 교정하는 것이 합당하다고 판단되는 경우, 맞춤법 검사기에서 일관되게 교정한 경우 등의 사례는 일관성을 고려하여 교정한다.

3) 고유명이 줄어든 경우도 교정하지 않는다.

예) 파바('파리바게트'가 줄어듦), 흠플('흠플러스'가 줄어듦), 필핀('필리핀'이 줄어듦)...

8. [22년 보완]²¹⁾ 표현상의 오류 처리

20) 이들의 띄어쓰기를 교정하거나 통일하는 것이 자연언어 처리의 관점에서 효율적이지 못하기 때문이다.

21) 작업 시 수집된 다양한 예를 추가한 것이다.

널리 사용되는 틀린 표현, 문법 오류 등은 교정하지 않는다.

예1) 이거랑 저거랑 틀리다. → ‘다르다’로 교정하지 않음.

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

예3) 소개시키다, 주차시키다, ... → ‘소개하다, 주차하다’로 교정하지 않음.

예4) 잊혀지다, 보여지다, ...(이중 피동) → ‘잊히다, 보이다’로 교정하지 않음.

예5) 맞다고 → ‘맞는다고’로 교정하지 않음.

예6) ‘기존에’와 같이 조사 ‘의’를 ‘에’로 잘못 쓴 경우 교정하지 않음.

예7) ‘발등의 불’과 같이 조사 ‘에’를 ‘의’로 잘못 쓴 경우 교정하지 않음.

예8) ‘들어간다고’가 문맥상 ‘들으러 간다고’일 경우 → ‘들으러 간다고’로 교정하지 않음.

[문맥] 내가 친구랑 같이 그 가수 노래를 들어간다고(문맥상 ‘노래를 들으러 간다고’의 의미임)

9. [22년 보완]²²⁾ 의미 불명 표현의 처리

<주의>

의미 불명 표현은 사전에 등재되지 않은 형태로, 앞뒤 문맥에 근거하더라도 그 의미를 분명히 알 수 없는, 주로 어휘 단위로 실현된다. 구체적으로 의미 불명 표현으로 처리되는 사례는 다음과 같다.

예) 만떡첩, 니병, 력부섬야, 학약 등

의미 불명 표현은 교정의 대상이 아니며, MS Excel의 의미불명어 칼럼에 그대로 복사해 붙여서 추후 일괄 추출할 수 있도록 한다. 복사해 붙일 때, 공백이나 마침표 등 특수 기호가 추가되지 않도록 주의한다.

10. 기타

화자가 오타를 인지하고 스스로 교정한 경우에도 오타는 교정한다.

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~

22) 데이터 수집의 일관성을 위해 도구 사용과 관련한 주의사항을 추가한 것이다.

[지침2] <우리말샘> 미등재어(외래어, 신어 등)의 처리

1. 미등재어의 범위

미등재어는 <우리말샘>에 등재되어 있지 않은 신어와 외래어 등을 이른다.(*일부는 미등재에서 등재로 바뀐 예도 있음)

신어 예) 맵부심, 반려인, 소떡소떡, 옷짜, 주린이, 코로나 블루, 비대면 강의, 공적 마스크, 낀낀세대, K방역, 배달거지 ...

외래어/외국어 예) 오렌지 브라운, 이너피스, 타임세일 ...

2. [22년 보완]²³⁾ 미등재어의 처리

미등재어는 향후 사전 등재 또는 기계 처리의 가능성을 고려하여 OoV(Out of Vocabulary) 목록으로 관리한다. 특히, ‘웨이팅, 푸드’처럼 사전에 형태는 있으나 해당 의미가 사전에 기술되어 있지 않은 경우(의미적 신어)에도 OoV로 표시한다. 구체적으로 교정하지 않는 유형과 사례는 다음과 같다.

1) 신어

예1) 나 진짜 프로 혼밥러임.

예2) 그럴 땐 궁디팡팡이야.

예3) 너무 깨발랄해.

예4) 홈트로빅, 헬욕아, 핵인싸, 개귀엽다, 대환장, 인기템 (꿀-, 핵-, 개-, 대-, -템 등의 접사가 붙은 다양한 어형 고려)

2) 줄임말

줄임말은 풀어서 쓰지 않고 그대로 둔다.

예1) 나는 컴알못이라 의성

예2) 꾸안꾸 스타일

예3) 겉바속촉에 커피 향도 나고

예4) 떡알못, 힙알못, 할말하얏, 토달볶, 내돈내산, 꾸안꾸, 아묻따, 음쓰봉

예5) 꾸안꾸 스타일 12가지야 → 꾸민 듯 안 꾸민 듯 스타일 12가지야 (x)

3) 미등재 외래어/외국어

예) 폰캠, 포텐, 페디, 팔로잉, 팀업, 트윈룩, 타임세일, 크레페, 쿠키박스

‘웨이팅, 웨이팅 리스트’처럼 사전에 형태는 있으나 해당 의미가 없는 어휘도 OoV 목록에 포함한다.

4) 미등재된 감탄사와 의성의태어

예1) 나란히 누워서 똥구르르르.

예2) 과연 가능할까? 두둥!

예3) 끄아, 냅글냅글, 두구두구, 잇힝힝, 해헹, 아고, 오옹, 호다닥...

5) [22년 추가]²⁴⁾ 단, 인명, 지명, 영화 제목 등과 같은 고유명: 미등재어이지만 OoV 목록에 포함하

23) 작업 중 수집된 예시를 추가한 것이다.

24) 기준을 명확히 하기 위해 OoV에 포함되지 않는 유형을 제시한 것이다.

지 않는다.

예) 이스너, 케이윌, 임영웅, 보라카이, 칸쿤, 조커, 블랙스완, 겨울왕국, 엑시트, 왓챠 ...

[지침3] 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 위해서?? → 다른 거 쓰기 위해서?

술도 먹고!!!! → 술도 먹고!

그럴 수도?/ → 그럴 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, 바뀐 지.

아, 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 말 차례가 바뀐 채로 문장이 이어진 경우에 맥락을 고려하여 발화 단위 마지막이라도 문장부호를 생략할 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 추가한다. 단 이모티콘 등 특수 표현이 단독으로 나타나면 문장부호를 추가하지 않는다.

예) 샀지 → 샀지.

같은뎡 → 같은데.

맞아 → 맞아.

ㅌㅌㅌ → ㅌㅌㅌ

ㅋㅋㅋㅋㅋㅋㅋ → ㅋㅋㅋㅋㅋㅋㅋ

연결어미로 말풍선이 끝나는데, 이후의 말풍선에 같은 화자의 발화가 이어지면 맥락에 따라 침표를 추가할 수 있다.

예) 말풍선 1: 책도 샀고 → 책도 샀고,

말풍선 2: 펜도 샀어.

5) 복수의 줄임표 문장부호가 나타나면 맥락에 맞게 하나만 남긴다. 또한 말줄임표의 역할을 하는 복수의 마침표는 규정에 맞게 세 개의 마침표로 교정한다.

예) 맞아.... → 맞아...

그래,,, 좋아. → 그래, 좋아.

6) 문장부호가 나타날 자리에 물결표만 나타난 경우는 발화 단위 마지막에 나타나야 할 문장부호를 추가한다.

예) 시퍼요~~ → 싶어요~~.

안녕하세요~~~ → 안녕하세요~~~?

거예요~ → 거예요~?

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

예) 내가해야해?! (의문문이 확실한 경우) → 내가 해야 해?

20분짜리니까...? → 20분짜리니까?

8) 어절 단위 내에 나타나는 말줄임표는 삭제한다. 어절 단위 내에 “?”를 써서 발화 내용에 대한 의문이나 불확실성을 나타내는 경우, “(?)”로 교정하고 앞뒤에 빈칸을 두지 않는다.

예) 내...가 → 내가

출산전 두달?부터 → 출산 전 두 달(?)부터

9) 부호가 규정에 맞지 않게 사용된 경우 또는 누락된 경우는 교정한다.

ㄱ. 마침표가 잘못 사용된 경우

예) 미백.주름개선 → 미백, 주름 개선 (마침표를 침표로 교정)

ㄴ. 문장 종결된 것이 분명한 경우임에도 문장부호가 없는 경우

예1) 오늘 3시 좋아 → 오늘 3시 좋아. (문장 종결이므로 마침표 추가)

예2) 정말 맛있다 이모든게 → 정말 맛있다, 이 모든 게. (도치문으로 침표와 마침표 추가)

10) 빗금이 두 개 이상의 어구를 묶어 나타낼 때나 수식어와 기준 단위 사이 등에 사용된 경우 빗금을 그대로 두고 앞뒤를 붙여서 쓴다.

예) 둘째 임신/출산으로 인해

11) 단어 사이에 붙임표가 있는 경우 삭제한다.

예) 크림치이-즈 → 크림치즈
아-무리 → 아무리
저엉-말 → 정말

12) 물결표의 사용

물결표는 음의 길이, 억양 등을 나타내는 초분절음소 기능을 하는 것으로 보아 문장부호 규정과 상관없이 유지한다.

(1) 단어 중간에 나오거나 복수의 물결표가 나와도 물결표는 유지시킨다.

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

(2) 물결표는 초분절 음소로 이해하여 문장부호보다 선행하여 표시한다.

예) 맞이하면 되겠네!!!!~ → 맞이하면 되겠네~!

13) 문장부호가 아닌 기타 특수 기호

문장부호가 아닌 기타 특수 기호가 사용된 경우에는 앞뒤를 한 칸씩 띄어서 쓴다.

예) 물어보는 거지~# → 물어보는 거지~. #
→ 물결표 다음에 마침표를 넣고 # 앞에 한칸 띄운다.

14) 이모티콘 기능을 하는 문장부호와 특수 기호

문장부호, 특수 기호 등을 이용해 얼굴 표정을 표현한 이모티콘의 경우는, 교정하지 않고, 별도의 목록으로 관리한다.

예1) 섭섭하네^~~ → “섭섭하네.”로 교정하고, “^~~”를 하나의 이모티콘으로 보아 별도의 목록으로 관리한다.

예2) 그랬어 ~~ → 그랬어. ~~ → “그랬어.”로 교정하고, “~~”를 하나의 이모티콘으로 보아 별도의 목록으로 관리한다.

[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 복원한다고 하더라도 우리말샘에 등재되지 않은 표현은 복원하지 않는다. 이러한 유형도 목록을 만들어 둘 필요가 있는데, 특수한 어절 구성을 보므로 자동으로 목록을 만들어 관리할 수 있다.

1. [22년 보완]²⁵⁾ 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 않고 별도의 목록으로 관리한다.

예) ^^, ^^::, ^~ㅠㅠ~, ^^~, >_<, ○^○, ^.^, ^0^, -_- , _ㅍ, ㅠㅠ, ㅠ, ㅍㅍ, ㅍ, ㅍㅍ, ;ㅍㅍ, ㅍㅍㅍㅍ, _ㅍ, ㅍㅍ; , @@@@.

2) 형태를 교정하지 않는다. 문자열에 붙어 있는 이모티콘은 앞 어절과 띄어 쓴다.

예) 참 조아^0^ → 참 좋아. ^0^

개꿀이지>_< → 개꿀이지. >_<

미안해요ㅍㅍㅍㅍ → 미안해요. ㅍㅍㅍㅍ

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

예) 눈치는 대리님이^^..... → 눈치는 대리님이... ^^

name1님^^~ 반갑습니다 → name1 님~, ^^ 반갑습니다.

4) 주로 한 단위로 쓰는 자소형 문자가 띄어져 있을 경우, 붙여 쓴다.

예) ㄷ ㄷ ㄷ → ㄷㄷㄷ

2. 감탄사나 응답 표현 기능을 하는 자소형 표현의 교정자음

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

예1) ㅋㅋㅋㅋㅋㅋㅋ, ㅎㅎㅎ (교정하지 않음.)

※ 이들 뒤에는 별도의 문장부호를 추가하지 않는다.

예2) 아ㅏ 아아ㅏㅏ 아 → 아아아아아아아

예3) ㅇㅋ → 오키

ㅎㅇ → 하이

ㅎㅇㅎㅇ → 하이하이

ㅇㅈ → 인정

ㅁㅈ → 맞아. (용언의 활용형이므로 마침표 부여)

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

25) 작업 중 수집된 예시와 유형을 추가한 것이다.

ㄹㅇ → 레알 (<우리말샘> 등재어 ‘레알’)

ㅇㅇ → 응응

2) 문자열에 붙어 있는 자소형 표현은 한 칸을 띄워서 교정하되 필요한 경우 문장부호를 추가하고 한 칸을 띄어 교정한다.

예1) 역시ㅋ → 역시 ㅋ

ㅎㅎㅎㅎ그래도 어제 → ㅎㅎㅎㅎ 그래도 어제

ㅎㅎㅎㅎ어머니아버지께 → ㅎㅎㅎㅎ 어머니, 아버지께

ㅋㅋㅋㅋ혼나? → ㅋㅋㅋㅋ 혼나?

예2) 아ㅋㅋ큰거면 → 아, ㅋ ㅋ 큰 거면

아이고 ㅎㅎㅎ → 아이고. ㅎㅎㅎ

꽤있어서ㅋㅋ → 꽤 있어서. ㅋ ㅋ

예3) [22년 추가]²⁶⁾ 아늑ㅋㅋㅋ → 아니. ㅋ ㅋ ㅋ ㅋ

ㅈ | ㄴㅈ ㅈ → 진짜

3. 기타 특수 표현의 처리

1) [22년 보완]²⁷⁾ 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.

ㅈㅈ하셈 → 수고하세요.

3) 신어를 알파벳으로 표기한 것은 그대로 둔다.

예) JMT(존맛탱 뜻으로 쓴 것)

4. 야민정음의 처리

야민정음은 한글 자모를 모양이 비슷한, 다른 자모 등으로 교체하여 단어를 표기하는 방법이다. 야민정음이 두루 쓰이어 언중 사이에 단어로 정착이 되면 사전에 등재될 수 있다. 현재 많은 목록이 <우리말샘>에 등재가 되어 있는데, <우리말샘>에 등재된 어형은 교정하지 않는다.

예) 땡땡이, 머머리, 땡곡, 땡반, 커엣다

5. 언어유희의 처리

아래와 같은 언어유희는 교정하지 않고 OoV로 처리한다.

예) 망실(‘실망’의 의미로), 아깝지 않다람쥐.

26) 작업 중 수집된 예시를 추가한 것이다.

27) 작업 중 수집된 예시를 추가한 것이다.

[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. 단, <우리말샘>에 미등재된 어형 또는 방언으로 기재되어 있으나 규범 표기가 제시되어 있지 않은 경우에는 표준어형을 추측해 교정하지 않고, 'OoV'로 처리한다.

2. 방언 종결어미의 처리

대화 상황에서 종결어미는 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 종결어미가 둘 이상의 교정 결과를 가질 수 있다.

[22년 보완]²⁸⁾

1) 동남 방언 의문형 종결어미 '-나' → 최대한 원형에 가깝게 교정하되, 맥락에 따라 '-어/어요'를 우선 적용하며, 맥락을 고려해 교정한다.

2) 어미의 경우, 방언형이 표준형 하나로 대응되지 않는 경우가 많으므로, 맥락을 고려하여 적절한 표준어형으로 교정한다.

예) 우리 점심때 영화 한 편 보고 저녁에 고기 먹을라하는데 엄마는 어떻노?

→ 우리 점심때 영화 한 편 보고 저녁에 고기 먹으려 하는데 엄마는 어때요?

※ 이 예문에서 '어떻노'는 표준어형 하나에 대해 1:1 대응이 되지 않는다. 이런 경우, 문맥에 따라 교정형을 선택하되, '-어/어요'형을 우선으로 하여 교정한다.

3) 종결어미의 경우, 사전에서 제시한 표준어형을 주로 하되, 화계에 따라 적절한 종결어미를 선택하도록 한다.

예) 맛있었슈? → 맛있었어요?

※ 화계에 따라 문맥을 고려하여 “맛있었어요?” 또는 “맛있었어?” 중 하나를 선택해 교정한다.

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

예) 지금 머라카노 → 지금 뭐라고 하니?

3. 방언형에서 음운을 변용한 경우의 처리

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

예1) 너무 많이 지난 것 갈애. → 너무 많이 지난 것 갈아.

예2) 그런 생각은 아예 하지를 말어. → 그런 생각은 아예 하지를 말아.

예3) 커능기 → 하는 것이

28) 작년 지침에서 문구를 보다 정확히 표현하고 예시를 추가한 것이다.

예4) 전부 수매를 하꺼인대 → 전부 수매를 할 것인데

4. 방언형에서 띄어쓰기의 처리

방언형에서는 띄어쓰기가 무시되었더라도 표준어형으로 교정 시 어문 규범에 준하여 띄어 쓴다.

예) 뭐라카노? → 뭐라고 하니?

5. 의미 불명 방언의 처리

방언의 의미를 정확하게 해석하지 못해 정확한 교정형을 판단하기 어려운 경우에는 ‘의미불명어’로 처리하여 방언 교정 지침의 참고 목록으로 제공한다.

예) 찌간거 딸고 다니기 힘들 → 찌간거 딸고 다니기 힘들어.

※ 위의 예시의 밑줄 친 부분은 ‘조그마하다’의 전라도 방언인 ‘찌간허다’를 사용하고 있는 것으로 보인다. 그러나 해당 대화는 서울 출생, 서울 거주 화자의 대화이며, 대화 문맥상 밑줄 친 부분의 의미를 정확하게 파악하기 힘들기에 ‘의미불명어’로 처리한다.

6. 미등재어 방언의 처리

1) 교정하지 않는 경우

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

예) 오늘 같은 날엔 패딩...에 찌부돼서

※ ‘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. 뜨시다 → 따뜻하다 (‘따뜻하다’의 방언(강원, 경상).)

예) 뜨신 물에 → 따뜻한 물에

ㄷ. 아짜리 → 차라리 (‘차라리’의 방언(경상))

예) 아짜리 다 줘 버려. → 차라리 다 줘 버려.

ㄹ. [22년 추가]²⁹⁾ 지 → 줄(의존명사 줄의 방언(경상))

예) 기장떡은 다 동그란지 알았는데 → 기장떡은 다 동그란 줄 알았는데

7. 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 기

29) 작업 중 수집된 예시를 추가한 것이다.

준에 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 교정하지 않는다. 구어체 말뭉치 통용 방언은 아래와 같은 기준에 따라 판단한다.

- ‘구어체 말뭉치 통용 방언’의 판단 기준

(1) 기존의 비교정 구어체 말뭉치 통용 방언형인 경우

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

(2) [22년 추가]³⁰⁾ <우리말샘>에서 ‘~의 방언’과 같이 대응하는 표준어형이 제시되지 않고 해당 단어의 뜻을 기술하고 있어 대응 표준어를 특정하기 어려운 경우

예) 발꼬랑내

※ 발꼬랑내는 사전에서 [방언]으로 대응 표준어 없이 발에서 나는 고약한 냄새(경상))를 가리킨다고 설명만 제시하고 있어 ‘발꼬랑내’로 그대로 둔다.

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

예) 이케

※ 273만 개 이상의 용례가 검색되므로 그냥 둔다.

2) 맥락에서 표준형으로 교체하여 어색한 경우, 해당 방언을 그대로 두기로 한다.

예) 와냐, 이 머스마.

※ 이 예문의 경우, ‘아냐, 이 사내아이’로 바꾸면 대화의 맥락이 어색해지므로 ‘머스마’를 그대로 둔다.

8. 교정하는 경우 ‘방언 : 표준어’ 대응쌍 예시

교정하는 방언형의 경우, 아래 예시와 질의응답 시트의 방언형 목록을 참고하여 교정하되 해당 어형이 목록에 없을 경우, 해당 시트에 교정한 방언형을 기록한 후 교정한다.

예1) 방언형 ‘니’는 → 너는

※ ‘니’를 교정하지 않는 경우는 ‘네(‘너+-의’의 줄임말, ‘니 것’, ‘니 동생’, ‘니 심정’)의 의미일 때뿐이므로 주의한다.

예2) 경상 방언 동의의 뜻인 ‘맞아’

화자 1. 맛있어 보여 ㅋㅋㅋㅋㅋㅋ

화자 2. 아, 맞나?

→ 맞아?

※ 경상 방언의 ‘맞나?’는 ‘맞아?’로 교정.

예3) 경상 방언 ‘~다 아니가/ ~다 아이가/ ~다이가

화계에 맞게 종결 어미를 교정함.

내가 일부러 아는 척했다이가 → 내가 일부러 아는 척했어.(또는 아는 척했지 않았어?)

※ 맥락에 맞는 것으로 교정하되 가장 원 형태의 의미를 살린 형태로 고른다.

30) 작업 중 수집된 예시를 추가한 것이다.

예4) 방언에서 비롯된 속어

얼탱 없다 → 어처구니 없다.

예5) 제주 방언

-멘: 화계와 시제에 따라 ~지? ~해? 체로 교정

나는 왜 그거 안주멘?: 나는 왜 그거 안 주지?

-클: -(으)르게/해, 해요 중 시제와 화계를 고려해 교정

그럼 좀 빠칠클: 그럼 좀 빠치는데?

비싸클: 비싸.

-연: -여서에 대응

듀랑고가 최고연: 듀랑고가 최고여서

-헨: -했어

예약헨: 예약했어.

-겐: -겠어

예약해야겐: 예약해야겠어.

예6) 따시다: '따뜻하다'로 교정

우째: '어찌'로 교정

예7) 땡기다: '당기다'로 교정

예8) 세다: '세다'로 교정

예9) 줌: '좀'으로 교정

9. 방언처럼 보이지만 방언이 아닌 경우의 처리

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

[지침6] 개인 정보 및 부적절한 표현(욕설, 혐오 표현 등)의 비식별화 처리

1. 비식별화 방법

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

<주의>

해당 칼럼에 OoV, 의미불명어, 자소형 이모티콘, 비식별화 대상(혐오 및 차별 표현, 개인 정보 등)의 항목을 복사해 붙여 넣을 때 “원 문장의 항목 형태” - “최종 교정 문장의 항목 형태” - “해당 유형별 칼럼의 항목 형태”를 일치시킨다. 공백이나 마침표 등 특수 기호가 추가되지 않도록 주의한다. 칼럼에 포함되는 항목 외의 교정 작업은 그대로 진행한다.

예) 뭐라고하는거야 미친새끼가

→ 뭐라고 하는 거야, 미친새끼가. (“미친 새끼”로 교정하지 않고 “미친새끼” 그대로 칼럼에 복사해 붙여 넣는다. 그 외 띄어쓰기, 마침표 부여 등 교정 작업은 그대로 진행한다.)

2) 입력할 때는 “원 문장 - 최종 교정 문장”을 일치시킨다.

3) 부적절한 표현(욕설, 혐오, 차별 등)은 조사를 떼고 입력하되, 용언의 활용형 및 형태소 경계를 파악하기 어려운 경우에는 어절 단위로 비식별화한다.

예1) 개빡쳐 → 개빡쳐(○), 개빡치(×)

예2) 근데 그 색히? → 색히(○), 새끼, 새끼(×)

예3) 좀 있다 쥘겨버라 → 쥘겨버라(○), 죽여버려(×)

예4) 정말 염병이었다 → 염병(○), 염병이었다(×)

예5) 시벌노마 → 시벌노마(○), 시벌놈아(×)

4) 이외의 비식별화 대상은 해당 부분만 입력한다.

예1) 양희가 곱창볶음 사뒀대. → 양희(○), 양희가(×)

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

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

2. 개인 정보의 비식별화

국어원 온라인 대화 말뭉치 구축 시 다음 범주에 대해 아래와 같이 비식별화를 진행하였다. 본 사업 팀에서는 이전 사업과의 유기적 연계성을 위해 본 과제의 대상 자료 전체에 대해 아래의 내용을 비식별화하기로 한다.

1) 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)

[22년 추가]³¹⁾ 유튜버의 경우 유명인도 있고 그렇지 않은 경우도 있어, 일관성을 위해 모두 비식별화함.

2) 비식별화하지 않는 대상은 다음과 같으니 주의한다.

일반 애칭 별명, 공인 실명(유재석, 조승우 등), 만화 주인공(신노스케, 짱구, 펑수 등), 드라마 주인공(김삼순, 길라임 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)

3) [22년 추가]³²⁾ 주의 사항1

기구축된 말뭉치에서 성과 이름으로 구성된 인명 중 이름만 비식별화된 경우에는 별도의 비식별화 작업을 하지 않는다.

예) 윤name14

4) 주의 사항2 : 인명이 호격조사와 엉겨 붙은 경우

(1) 일반인 인명(비식별화 대상)의 경우

예) 성후나 → ‘성훈아’인 것으로 보고 ‘성훈’을 비식별화

(일반인의 경우 직접 부르는 경우가 일반적일 것이므로, ‘성훈’을 단독으로 떼서 비식별화한다.)

(2) 연예인 인명(비식별화 대상이 아님)의 경우

예) 성후냐, 성후니, 후나, 성후니야 → 교정하지 않음.

(연예인의 경우 직접 부를 일은 없을 것이므로, 그 자체로 고유명사로 봄.)

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, 출신 지역, 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

31) 작년 데이터와는 달리 유튜버의 이름이 많이 등장하여 지침에 추가한 것이다.

32) 작년 데이터와는 다른 새로운 유형의 오류가 발견되어 추가한 것이다.

갖는 표현 (국가인권위원회, 혐오표현 리포트, 2019)

부적절한 표현은 욕설, 차별, 혐오, 성적인 표현을 이른다. 연구 등의 특수한 목적으로 욕설이나 혐오 표현에 대해 연구할 경우, 원시 말뭉치에서 원문이 보존되므로 일반 사용자를 고려하여 이들 표현은 비식별화한다.

비식별화 대상은 우리 사회의 혐오를 조장할 가능성이 있는 경우, 당사자들에게 과도한 불쾌함이나 모욕감을 줄 수 있는 경우, 잘 알려져 있지 않던 혐오 표현이 오히려 알려지는 계기가 되는 경우 등이다.

연구 등의 특수한 목적으로 욕설이나 혐오 표현에 대해 연구하고자 하는 경우가 있더라도, 원시 말뭉치에서 원문이 보존되므로, 일반 사용자를 고려하여 아래와 같은 유형에 대해 비식별화한다.(비식별화의 보다 다양한 예는 ‘<부록> 욕설, 혐오 차별 표현의 비식별화 사례’ 참고)

1) 욕설의 비식별화

(1) 다음과 같은 욕설(밈줄)은 비식별화한다.

예1) 아 씨, 기억이 안 나. ㅋㅋㅋㅋㅋㅋ

예2) 아쉽다고 전해줘 ○ㄷㄹ

(2) 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, 찐다 등

(3) 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 ㅋㅋ

ㄴ. 비식별화하는 차별 및 혐오 표현(밈줄)의 예는 다음과 같다.

예) 사우디 위험하다.

종교 경찰 있고

ㄷ 보수적 사회다.

3) 특정 대상에 대한 부정적 평가 표현의 비식별화

공인이나 기관의 경우에는 비식별화 대상이 아니지만, 부정적인 내용이 포함될 경우에는 해당 대상을 비식별화한다.

예1) 백예대는 양아치 소굴이라는 게 학계 정설

→ ‘백예대’에 대한 부정적인 내용이라서 ‘백예대’ 비식별화, ‘양아치’는 욕설이라서 비식별화

함.

예2) 근데 이재용 님 똑똑해 보이진 않는데. 돈 많게 자라서 그런가 착해 보이긴 한다면 ㅋㅋㅋ

→ ‘이재용’에 대해 ‘똑똑해 보이진 않는다’라는 부정적 평가 내용을 포함하고 있으므로, ‘이재용’을 비식별화함. (유명인은 비식별화 대상이 아니지만)

예3) 오투기가 청크 카레 너무 실망. 비싼데 맛이 없었어.

→ 상품에 대한 부정적 평가는 비식별화하지 않는다.

예4) 겨울왕국도 봤어? 끔찍해, 진짜. 완전 유해해

→ 겨울왕국과 같은 영화에 대한 부정적 평가는 비식별화하지 않는다.

예5) 너 나훈아의 누나가 되고 싶어?

아니, ㄹㅇ

땅에 묻혀야 되는 거 아냐?

ㄹㅇ냐.

글고 나훈아보다 어리셔. ㅋ

ㅈㅈ

글고 성훈 개빡치는 거

유튜브에 성훈 치면

배우 성훈만 ㅈㄴ 나옴.

→ ‘나훈아, 남진, 성훈’에 대해 부정적인 평가를 포함하고 있으므로 이들 인명을 비식별화함.

예6) 하이라이트 네 명이야?

스타트.

응.

name3, 손절.

더 많지 않았나?

개 똘 일 있었음?

name4이랑 같이 놀았음.

아, 그래서...

그룹에서 쫓겨남?

→ 연예인의 실명은 비식별화되어 있으나, 그룹명/노래명이 공개되어 있어 사건을 알면 비식별화된 인물이 누구인지 쉽게 알 수 있음. 이 경우 ‘그룹명, 노래명’을 비식별화함.

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.

예1) 연예인을 대상으로 한 ‘유사 연애’ 설정 대화

동성애 코드(알페스)가 두드러지지는 않으나 연예인을 실명을 거론하며, 실제 연애를 하는 것처럼 상황을 설정하고 대화를 나누는 내용으로, 작업 대상에서 제외함.

예2) 이재현 이번 학기 인턴십 나가서 돈 많다고 자기가 쓴다 할 때 그때서야 오빠 짱~ 해주는
 name1 언니
 ㅋㅋ 개웃기다.
 뭘 이재현의 수난시대... 아님?
 아, 이 사람 웰케 재밌지.
 존나 흥미가 생길 거 같다.존나...

진짜 성격은 모르지만...
 내가
 밀빡픽 몇 개 본 거론
 이 정도 캐해가 맞는 것 같아.
 아, 개웃겨.
 오랜만에 유사하니까 재밌다.

5. 비식별화의 대상이 아닌 경우

1) 비속한 표현이기는 하나, 강조의 의미를 갖는 경우

예) 진짜 연출력 미쳤다.

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

2) 비속한 표현이지만 욕설이라고 보기 어려운 경우

예) 존맛탱, 대존맛, 까먹다, 개좋다, 개꿀...

3) ‘인종, 종교’ 등과 관련되나 차별 및 혐오와 무관한 맥락인 경우

예1) 일본은 그런 거 잘하잖아. (상품에 대한 선호와 관련된 맥락임.)

예2) 동남아 쪽도 괜찮을 거 같음. (여행지 추천의 맥락임.)

<부록> 욕설, 혐오 차별 표현의 비식별화 사례

1) 욕설의 비식별화

예1) 이런 쌍것 / 상것들

예2) 아 씨, 기억이 안 나. ㅋㅋㅋㅋㅋㅋ

예3) 보여 주냐? 시팔 / 에이, 시벌 / 조까 시팔. / 씨발롬아 / 조팔놈아. / 벌써 몇 년. 스바 / 앓, 슈발 / 계란탕면 없어, 스바.

예4) 개못생긴 찌인싸.

예5) 내가 그걸 모르겠니? 시팡조팔. / 시팔조팔. / ㅋㅋㅋㅋ 스벌텐. / 시부럴. / ㅈㄴ시바라아랄 / 욕 먹는 기분 1818

예6) ㅋㅋ 존나 입에 붙네. / 한 명씩 죽일 거야. 존나. / 나는 존나 빨라서 / 물 존나 쳐먹고 자서. / 근데 저는 존나게 건치라서 / 중나 인상 깊었음. / 손톱 존나 아기자기. / 올 엄마한테 조온나 혼났거든. / ㅈㄴ 웃겨

예7) 시봉탱염불 외면서 올라갔잖아...

예8) 주차 자리가 없네, 젠장.

예9) 완전 많이 마셨는데도 그 새끼 안 취하더라고. / 턱형 새까... / 졸려, 이 색기야. / name6 스키.

예10) 쉬벌... 미친 새끼네. / ㄹㅇ놈

예11) 좇간이 미안해! / 좇도 연습을 안 하니까 그렇죠. / 조까튼 황야의 마녀.

예12) 너한테 기대한 내가 병신 / 내가 이해력 병신인 건가?

예13) 왜 설레는데? 이 정신병자야.

예14) 인터넷 소설 이 지랄하면 / 찌도 귀여울 거야 이 지랄하지만 / 지랄이라니?/가서 개지랄

예15) 양아치네. 사실 부럽다...

예16) 근데 여자 친구 있거나 결혼했거나 성격이 개씹이거나.

예17) 똥년똥.

예18) 그니까 나 년이 포라이지.

예19) 얼어 똤져라. / 너무 어려움. 똤짐. / 똤지는 수가 있어.

예20) 원래 그런 거부터 조져야 됴.

예21) 똤청아! 간발 새끼야.

예22) 시발놈이네, ㄴㅇ 개자식.ㅈㅈ 예) 아가리로 대장 뽑아버린다. / 아가리 열지 마.

예23) 너라고 하면, 입 찢는다.

예24) 나 걱정해 주는 거야? 그냥 죽어.

예25) 모가지 관리 잘해. ㄲ

예26) 갑자기 뽀큐를 날려요

예27) 쓰레기 새기도 모자람.

예28) 아쉽다고 전해줘 ○ㅈㄴ

예29) ㄴㅇ

2) 차별 및 혐오 표현의 비식별화

예1) 추접시러워 중국집 여자같은

감히 중국 여자와 비교질을

너가 중국 사람처럼 생겨서 그래

예2) 동네 사람은 냄새나는 그 사람들.

예3) 인도보다 더 더러운 나라가 이집트인

예4) 그래...유럽서는 소매치기 당하고 동남아서는 죽는다.

예5) 기독교는 다른 종교 무시하잖아.

예6) 시골탱이 주제에.

- 예7) 거지 껏껏이야.
- 예8) 대만 뻗속까지 친일파다 반한에...
- 예9) 여튼 청소 아줌마랑 좀 다른 느낌인데
청소 여사보단 지능이 좀 더 높긴 한데
- 예10) 어휴, 스비 하여튼 한남 새끼.
- 예11) 회사 영감탱이들 회의를
너무 좋아해 가지고.
회의 없애자고 건의하면,
회의를 어떻게 없앨지,
회의합시다~~.
한다고. ㅋㅋㅋ
- 예12) 경상도 분들이 많이 그러신다고. ㅋㅋ
어디, 여자가!
- 예13) 저도 아침에 을지로까지는 지하철 타는데,
노약자석 쪽에서 타면은
뭔가 이상한 냄새 나요.
거기 아저씨들이 몰려 있어서.
저도 은근 냄새 민감해서 코 막아 버려요. ㅠㅠ
- 예14) 잼이 영국 거네.
그래서 맛이 없네.
영국. ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
야, 이 정도면 메이드 인 영국은
안 받아야 하는 거 아니냐?
내 말이.
양심 없는 놈들 수출 왜 해?

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.

- 예1) 백예대는 양아치 소굴이라는 게 학계 정설
→ ‘백예대’에 대한 부정적인 내용이라서 ‘백예대’ 비식별화, ‘양아치’는 욕설이라서 비식별화함.
- 예2) 근데 이재용 님
똑똑해 보이진 않는데.
돈 많게 자라서 그런가
착해 보이긴 한다면 ㅋㅋㅋ
→ ‘이재용’에 대해 ‘똑똑해 보이진 않는다’라는 부정적 평가 내용을 포함하고 있으므로, ‘이재용’을 비식별화함. (유명인은 비식별화 대상이 아니지만)
- 예3) 오투기가 좀 고급화 전략을 선택한 거 같은데
저번에 청크 카레 너무 실망.
비싼데 맛이 없었어.
→ 상품에 대한 부정적 평가는 비식별화하지 않는다.
- 예4) 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.

예1) 니가 걸 그룹 슴골을 검색하지 않았겠지?

난 걸 그룹 취향 아닌 거 알지 않니?

개네야 다 보라고 나오는 거니까 그냥 보는 거고

내가 좋아하는 건 따로인데.

예2) 그냥 가슴 변태라는 거밖에 몰라.

5) 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 결정한다.

해당 부분은 교정하지 않고, ‘혐오’ 열에 해당 대화를 복사하여 넣는다.

예1) 그리고 요새는

중국보다 태국인가?

거기 불체자가 더 많다며.

헐.

진짜 싫어.

아니, 여기 와서 여기 법 따르는 것도 아니고.

자기들 멋대로 하는 애들을 자꾸 왜 받아 줘.

그니까.

길거리 간판에

한국말로 안 쓰고 중국 말로만 써 두는 것도

존나 개 꼴 보기 싫음.

꼭 중국인들만 그래요.

그니까.

개네 병원에서 간병인 쓸 때도

중국인만 씬. ㅋㅋ 씨발.

ㅁㄷ

존나 싫어.

아, 진짜 중국인들 다 걸렸으면 좋겠어.

다시 재조사해서

불체자 다 쫓아 내고.

건보료 굶어 먹는 애들 돌려 보내고. 혐오

그니까.

건보료가 제일 빠쳐.

건강 보험 왜 빠세게 안 잡냐고, 씨바.

내 세금.

물론 지금은 안 내지만. ㅎㅎ

ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ

근데 우리 세금 다 야금야금 굶어먹는 거 맘에 안 들어.

그니까.

정작 우리나라에서 진짜 받아야 할 사람들은
 못 받고 있는데.
 씨발럼으 외국인 새끼들.
 좀 잡아라.
 맞아, 맞아.
 진짜.
 개억울해.
 존나 필요한 건 안 하고
 필요 없는 건 함.
 국회의원들 대체
 월급 왜 받는지.
 ㅋㅋㅋㅋㅋㅋㅋㅋ
 맞아.
 시발 국회 의원 최저 시급으로 돌려라.
 레알 최저 시급 너무 많다며.
 그럼 자기들도 최저 시급 받으면.
 행복할 거 아님?
 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
 너무 많은 돈 받잖아. ㅋㅋ
 맞아, 맞아.
 같이 하자.
 국민을 위해 봉사하는 사람들이
 돈을 왜케 많이 받나.
 그니까.
 그리고 그 보좌관인가?
 개네 월급도 국가 돈이라며.
 그래서 지인들로 풀로 채운다며. ㅋㅋ 스님
 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
 스님
 나도 지인이면 좋겠다.
 그니까. ㅋㅋ 자기들 월급에서 돈 나가야.
 인력 제대로 쓰지. ㅋㅋㅋ 시발.
 ㅋㅋㅋㅋㅋㅋㅋㅋ
 국회 의원 하고 싶다.
 자한당이라도 되어야 할까요?

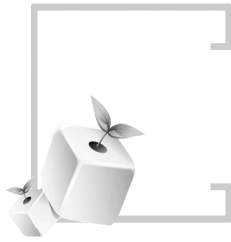
예2) 어, ㅋㅋㅋㅋㅋㅋ 내 거도 아닌데.
 아니, 내가 남의 남친 힘 좋은 거랑 고추 큰 거 알아서
 얻다 쓰냐?
 ㅋㅋㅋㅋㅋㅋㅋㅋ
 상상하기도 싫은데. ㅋㅋㅋㅋ
 내 말이. ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
 막, 아니, 나타나자마자
 토 나온다고.
 name6 만나자마자 앉기도 전부터
 징그러워. ㅋㅋㅋㅋ

"오늘 너무 피곤하다."
 이러더래.
 막 개네듯이.
 어. ㅋㅋㅋㅋㅋㅋ
 ㅋㅋㅋㅋㅋㅋㅋㅋ
 그래서 name6가 "왜? 어제 뭐 했는데?"
 이러니깐
 "아니, 네 번이나 했잖아, 어제 나..."
 이러면서
 어제 했대?
 아,
 헐.
 아니, 무슨
 "어찌나 밝히는지."
 가능하나?
 이러면서.
 ㅋㅋㅋㅋㅋㅋㅋㅋ
 왜 네 번은 가능하지.
 피곤하긴 하겠다.
 하루 종일이면.
 난 안 될 거 같은데.
 ㅋㅋㅋㅋㅋㅋ
 병날 거 같은데.
 그러면서 마지막은 그냥 그랬는데 세 번은 너무 좋았다고
 이런 말 하던데. ㅋㅋㅋㅋ

예3) 아까 낮에 말한 30대 여배우
 봤어?
 호빠 맞잖아. ㅋㅋㅋㅋ
 내가 호빠랬지? ㅋㅋ 유흥업소
 부킹해서 만난거면 나이트라고 했을 거야
 ㅋㅋ
 난 몰랐어, 드레스도.
 검색하고 알았네.
 일부러라고... 사람들은...
 나도 그 생각 했는데
 좀 일상 자체가...
 인스타 올라온 거 보니
 좀...
 ㅋㅋㅋㅋ
 응.
 사진들이 죄다 좀...
 보면서 나도 드... ㄹ 이랬거든. ㅋㅋ
 사람들 눈은 다 비슷해.
 찐티 아주 짹짹...
 여하튼 내 스타일 아니야.

저런 분
 저게 과감이니?
 ㅋㅋㅋ
 저건 그냥 벗는 거야.
 과감 섹시 뭐 이래야지.
 노출 더럽 이럼 안 되잖니?
 ㅋㅋㅋ
 그니까.
 걸릴 게 뭐 있어.
 가렸잖아. ㅋㅋㅋ
 ㅋㅋㅋㅋㅋ
 내가 더러... ㅂ
 했잖아. ㅋㅋ
 아니, 좀 뭐랄까.
 저러면서 섹시하고 예쁜 것도 많잖아.
 한혜진 그 누드 사진 얼마나 멋져.
 이분은 아니야.
 오우, 노노~~~.
 응, 나도 싫어. ㅋㅋㅋㅋ

예4) 혈~, 그분 선도하러 가나?
 첨엔 터키 가려고 하더니...
 나라서 위험 지역이라 했는데.
 교회 사람들이라 순례길 같은 거.
 위험 지역에는 가지 말아야 하는데.
 꼭 기독교가 이상한 곳에 가서 선교 많이 하더라.
 그래서 내가 잡혀서 울면서 티브이 나와 구해 달란다 할 거다 했다. ㅋ
 상황 봐 가면서 선교해야지.
 선교 말고 순례
 가지 말라고 해 놓고는.
 순례하면서 선교도 하겠지.
 그 사람들이 선교를 포기하겠니?
 그래도 간다더니.
 다시 이집트 가기로 했단다.
 혈~, 잘못하다가 큰일 날 텐데.
 기독교는 다른 종교 무시하잖아.
 몰라~, 선도는 위험하기도 해서 잘 안 한다더라.
 선교하는 사람은 만나고 오는 듯.
 으이구~, 답 없다.
 암튼 마인드가 좀 그렇더라, 위험 지역인데도 간다는...



제 4 장

결 론



1. 사업 요약

이 사업은 2021년에 국립국어원에서 구축한 온라인 대화 말뭉치를 대상으로 125만 발화를 선별하고, 이를 자동 형태소 분석, 기계 번역 등의 한국어 처리 도구가 분석할 수 있는 수준으로 교정한, 교정 병렬 말뭉치를 구축하는 사업이다. 따라서 이 사업에서는 완벽한 수준의 맞춤법 교정보다는 자연어 처리 도구의 적용이 가능한 수준의 말뭉치를 구축하는 것, 공공재로서의 말뭉치의 특성을 고려하여 맞춤법 교정 및 개인 정보 및 혐오 표현 처리 지침과 처리 방안을 마련하는 것을 사업의 주요 목적으로 삼았다. 이러한 목적 하에 수행된 이 사업의 결과를 요약하면 다음과 같다.

첫째, 이 사업의 대상을 확정을 위해, '21년 온라인 대화 말뭉치에서 맞춤법 교정 대상 대화를 일정한 기준에 의해 선별하였다. 온라인 대화 말뭉치는 시스템 메시지(사진 공유, 동영상 공유 등), 자모 연쇄 발화, 이모티콘만으로 이루어진 대화 등, 온라인 대화 특유의 대화 유형을 포함하고 있다. 이 사업에서는 질적으로 우수한 온라인 대화 맞춤법 교정 병렬 말뭉치를 구축하기 위해, 이러한 무의미 대화를 제외한, 온라인 대화 대화의 특성이 잘 드러나며, 하나 이상의 주제에 대해 실질적인 대화 내용을 담고 있는 발화 125만 발화 이상을 선별하였는데, 이는 부적절한 표현에 대한 비식별화 맥락상의 혐오 및 차별 표현 등이 드러난 대화 파일의 삭제 등에 대비하기 위한 것이다. 이렇게 추출된 맞춤법 교정 대상은 어절 기준으로 약 400만 어절에 달한다.

둘째, '21년 맞춤법 교정 말뭉치 연구 분석 사업의 결과물인 맞춤법 교정 지침을 비판적으로 검토하고 개선하였다. 이 사업의 대상인 온라인 대화 말뭉치는 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 자연어 처리 도구의 적용을 어렵게 하며, 인공지능 데이터의 윤리적 활용에서도 문제를 일으킨다. 본 사업에서는 온라인 대화 말뭉치의 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, 21년 맞춤법 교정 말뭉치의 지침을 교정 보완하고 내용을 더욱 정교화하여 수립하였다. 표준형과 비표준형의 판별은 <우리말샘>을 주요 기준으로 하되, 유형에 따라 별도의 지침을 수립하여 목록을 관리하였다. 특히 이번 사업에서는 2021년부터 도입된 'OoV'(Out of Vocabulary)의 항목 정보와 함께 '의미불명어' 범주를 새롭게 도입하여 관련 분류 항목을 세분화하였다. 또한 온라인 환경에서 나타나는 특수 표현은 원문자로 복원하는 방향으로 지침을 개선하였으며 이모티콘은 별도의 범주를 부여하여 JSON 구조에

반영하였다.

마지막으로, 맞춤법 교정 병렬 말뭉치의 구축은 다음과 같은 절차와 방법, 도구 활용을 통해서 이루어졌다. 이 사업은 자동 검사기 처리를 거친 후 수작업으로 맞춤법과 띄어쓰기를 교정하는 방식으로 이루어졌으며, 교정 작업의 효율화와 검수 일관성 작업을 위해 교정 병렬 말뭉치 구축 도구인 Kronoth와 마이크로소프트사의 Excel 프로그램을 병행하여 사용하였다. 또 말뭉치의 형식 검수 단계에서는 (주)이르테크의 말뭉치 검증 시스템을 활용해 분석 결과의 정확도를 확보하였다. 맞춤법 교정 말뭉치의 구축은 (1) 맞춤법 교정 대상 대화의 선별, (2) 텍스트 전처리를 통한 맞춤법 교정용 말뭉치 변환, (3) 자동 맞춤법 교정 도구를 이용한 1차 자동 교정, (4) 수작업 전수 교정, (5) 개인 정보와 부적절한 표현의 비식별화, (6) 세 차례의 품질 검수, (7) JSON 구조화, (8) 최종 형식 검수의 과정으로 구축되었다.

이상의 과정을 통해 이 사업은 125만 개 발화(대화 파일별 최대 어절 수 20,000어절, 대화 파일별 최소 발화 수 25개 기준)의 온라인 대화 맞춤법 교정 병렬 말뭉치를 구축하였다. 이 사업으로 구축된 말뭉치는 자연어 처리의 관점에서 인공지능 학습용 데이터로서의 수준을 만족하는 수준으로, 자연어 처리 도구의 적용이 가능하다. 특히, 온라인 대화가 가지는 특수한 언어적 성격을 충실하게 반영하여 수립된 지침에 따라 구축되었다는 점에서, 향후 맞춤법 검사기의 정밀도와 정확도 향상을 비롯한 자연어 처리 응용 기술뿐만 아니라 언어 연구와 언어 교육 분야에서도 폭넓은 활용이 기대된다.

2. 향후 연구 및 정책 제안

본 과제 수행 결과를 바탕으로 향후에 진행하여야 할 연구 과제 및 정책에 대해 제안하면 다음과 같다.

2.1. 맞춤법 정렬의 단위 관련 제안

올해 사업은 원문과 교정문의 정렬 단위를 말뭉치 단위로 설정하여 구축하였고, 보다 세밀한 언어 단위의 원문과 교정문의 대응 유형 및 오류 유형을 살피기 위해 전체의 0.5%인 15,726어절에 대해 어절 단위의 정렬 작업을 수행하였다. 그러나 향후 원문과 교정문의 보다 세밀한 대응 정보의 추출이나 기계 학습을 위해서는 여전히 말뭉치의 경계를 보존하면서 문장, 어절, 형태소 등의 보다 작은 단위의 정렬을 고려해 볼 필요가 있다.

예1)

[문장 대응] “가진 않겠져?/가진 않겠쥬?”, “세번 먹었는데ㅋㅋ/세 번 먹었는데 ㅋㅋ”

[어절/청크 대응] “않겠져/않겠쥬”, “세번/세 번”, “먹었는데/먹었는데”

“안되요/안 돼요”, “아니였어/아니었어”, “한 잔해요/한잔해요”

[형태소 대응] “여/요”, “스/쑈”, “는데/는데”

위의 예1)에서 보인 것처럼 어절 또는 청크 단위의 정렬은 띄어쓰기가 구분자 역할을 하므로 구획이 용이하고, 말풍선이나 문장보다 단위가 작아서 원문-교정문에 대한 보다 세밀한 대응 정보의 추출이 가능하기 때문에 그에 대한 시도가 요구된다. 또 위 예1)의 [형태소 대응]에서 볼 수 있듯이, 교착어로서의 한국어의 특성을 반영할 때, 형태소 단위의 대응을 통해야만 맞춤법 오류 정보를 보다 정확하게 관찰할 수 있는 경우도 있다.

한편, 오류 유형에 따라서는 관찰이 가능하거나 추출 및 식별이 훨씬 용이한 정렬의 언어 단위가 다를 수 있다.

예2) 너도 한 번 해 봐./너도 한번 해 봐.

위의 예2)에서 ‘한 번’의 띄어쓰기 오류에 대한 교정은 기계 학습의 관점에서는 청크 단위보다는 ‘해 봐’와의 결합 관계가 포착 가능한 문장 단위 정렬이 보다 적합할 수 있다. 그러나 말뭉치의 정렬 작업에서 보다 작은 언어 단위의 정렬은 상위 언어 단위의 정렬을 전제하고 상위 단위의 정렬 정보를 계승하는 점을 상기할 필요가 있다. 다시 말해, 문장 단위의 정렬은 말풍선 단위의 정렬을 전제로 하고, 어절 또는 청크 단위의 정렬은 그보다 큰 문장 단위의 정렬을 전제로 한다.

예3)

[어절/청크 대응] “않겠져/않겠쥬”, “돼여/돼요”, “씻는데/씻는데”, “해용/해요”

[형태소 대응] “여/요”, “여/요”, “스/쑈”, “는데/는데”, “용/요”

위의 예3)의 어절/청크 대응과 형태소 대응의 결과를 살펴보면, 형태소 단위의 대응 결과에서 얻을 수 있는 정보와 어절/청크 대응에서 얻을 수 있는 정보가 다름을 확인할 수 있다. 한국어의 교착어의 특성상 형태소 단위의 정렬은 매우 필요하지만, 이런 작업은 형태소 분석이 선행되어야만 가능하다. 그러나 오류가 들어 있는 원문에 대한 자동형태소 분석은 정확도가 낮기 때문에 시간 소모적이며, 상당히 많은 비용이 발생할 수밖에 없다는 문제가 있다.

요약하면, 정렬 단위가 작을수록 오형태와 교정 형태의 대응쌍을 관찰하고 추출하고 언어학 연구, 언어 공학 연구 등 범용적으로 사용하는 데 효과적이다. 특히 맞춤법 교정 말뭉치의 활용 범위를 인공지능의 기계 학습이나 맞춤법 교정기 개발 등의 차원을 넘어, 언어학, 언어교육학적인 활용까지 고려할 경우, 궁극적으로 형태 단위의 정렬까지 고려해야 할 필요가 있다.

2.2. 맞춤법 오류 유형 연구 관련 제안

본 연구에서는 추가 제안으로 전체의 약 0.5%인 15,726어절에 대해 청크 단위 정렬 교정 병렬 말뭉치를 구축하여 청크 단위의 교정 전후 비교 분석을 통해 주요 교정 유형을 관찰하였다. 이를 통해 청크 단위의 교정 양상을 삽입/교체/삭제로 나누어 빈도와 비율³³⁾을 관찰하고, 말뭉치에 대한 후처리를 통해 상위 30위까지의 고빈도 형태소 단위의 교정 목록³⁴⁾을 추출하여 제시하였다. 이러한 시도는 실제 한국어 모국어 화자가 실생활에서 생성하는 오류의 경향성을 제시할 수 있다는 점에서, 최근 자연어 처리 분야에서 주로 인위적으로 생성된 오류를 활용해 온 것과는 변별되는 연구의 가치가 있다. 기계적으로 생성된 노이즈 생성 데이터와 모국어 화자가 실제 생성하는 오류 경향성은 각기 유형과 빈도 등 양상이 전혀 다르기 때문이다.

실제 모국어 화자의 오류 유형을 주석한 데이터는 자연 언어 처리뿐만 아니라 한국어학, 국어교육, 한국어교육학 분야에서도 광범위하게 활용될 수 있다. 최근 자연어 처리 분야에서 시도되고 있는 자동 노이즈 생성 말뭉치의 구축과 활용, 한국어 어문 규범 교육에서도 실제 오류에 기반한 객관적 데이터에 대한 관찰과 분석은 필수적이며, 맞춤법 오류의 유형별 빈도 목록은 기초 자료로서 중요하다. 따라서 추후 사업에서는 맞춤법 말뭉치의 구축 외에 맞춤법 오류 유형에 관한 연구를 사업의 범위로 고려할 필요가 있다.

물론 위의 사업 제안 2.1과 2.2는 모두 현행 말뭉치 구축 작업에 추가로 인력과 작업 시간을 소요하는 작업이다. 올해 추가 제안으로 시도된 연구로, 어절/청크 단위의 정렬 및 교정 작업은 말뭉치 단위의 정렬 및 교정 작업에 비해, 동일 분량 어절 대비 5~6배의 시간이 소요됨을 확인할 수 있었다. 또, 작업자들이 형태 단위의 일관된 교정 작업을 위해서는, 오류 교정 작업에 최적화된 효율적인 작업 도구의 개발 및 기능 개선 또한 매우 중요하다는 것을 확인하였다. 따라서 추후 문장, 어절 또는 청크 내지는 형태 단위의 정렬 및 교정 사업을 위해서는 인력과 시간이 훨씬 더 많이 소요된다는 점을 고려하여 적절

33) 34쪽 <표 10>의 교정 대상별 교정 양상 규모를 참조 바람.

34) 36쪽 <표 11>의 문자열의 교정 값 기준 고빈도 목록을 참조 바람.

한 사업 계획을 수립할 필요가 있다. 궁극적으로 맞춤법 교정 말뭉치의 구축 목표는 인공지능 기계 학습을 넘어, 언어학, 언어 교육학적 활용까지도 고려되어야 하는바, 교정 말뭉치의 정렬 단위와 오류 유형에 대한 고찰, 이러한 정교한 주석 작업을 고려한 사업 계획 설계가 필요하리라고 본다.

참고문헌

- 김진웅(2021), 자연언어 처리에서 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 외(2021), 『2021년 맞춤법 교정 말뭉치 연구 분석』, 국립국어원.
- 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 - 유튜브 댓글을 중심으로, 한말연구 57, 63~96.
- 박일섭 외(2019), 『메신저 대화 자료 수집 및 말뭉치 구축』, 국립국어원.
- 박일섭 외(2021), 『2021년 온라인 대화 자료 수집 및 정제』, 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), 병렬 말뭉치 구축과 응용, 언어정보와 사전편찬 25, 5~40.
- 윤은정, 김진호, 남길임, 송현주, 옥철영, 최준, 박운배(2018), 교육용 과학언어 연구를 위한 범용 자료로서 과학교과서 말뭉치 K-STeC(Korean Science Textbook Corpus) 구축, 한국과학교육학회지 38(4), 575~585.
- 이기황, 김수경(2022), 『한국어 대화 요약 데이터 구축 가이드라인』, 한국지능정보사회진흥원.
- 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.
- Baron, N. S. (2010). *Always on: Language in an online and mobile world*. Oxford University Press.
- 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., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk*, 41(5-6), 715~737.
- Flint E., Ford E., Thomas O., Caines A., Buttery P.(2017). A text normalisation system for Non-Standard English words. *Proceedings of the 3rd Workshop on Noisy User-generated Text*. 107~115.
- 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.

<기획·연구>

국립국어원 강미영 언어정보과장

국립국어원 이선영 연구원

<연구 참여자>

연구책임자 남길임(경북대학교 국어국문학과 교수)

공동연구원 곽용진((주)이르테크)

안미애(경북대학교 국어국문학과 교수)

송현주(경북대학교 국어교육과 교수)

안의정(연세대학교 문과대학 강사)

황은하(배재대학교 국어국문·한국어교육학과 교수)

연구보조원 심난희(배재대학교 주시경교양대학 교수)

현영희(경북대학교 국제교류처 강사)

강신아(연세대학교 박사 수료)

백미경(경북대학교 국제교류처 강사)

강윤희(경북대학교 국어교육과 박사과정)

강민지(경북대학교 국어국문학과 박사과정)

황지윤(경북대학교 국어국문학과 박사과정)

안진산(경북대학교 국어국문학과 박사과정)

박시온(경북대학교 국어국문학과 박사과정)

고예린(경북대학교 국어국문학과 석사과정)

박아름(배재대학교 한국어교육학과 석사과정)

정나현(배재대학교 한국어교육학과 석사과정)

김수지(배재대학교 한국어교육학과 석사과정)

정희연(배재대학교 한국어교육학과 석사과정)

성민규(경북대학교 국어국문학과 학사과정)

장희선(경북대학교 국어국문학과 학사과정)

조은실(경북대학교 국어국문학과 학사과정)
배수종(경북대학교 국어국문학과 학사과정)
안효민(경북대학교 국어국문학과 학사과정)
한도연(경북대학교 국어국문학과 학사과정)
이현영(경북대학교 국어국문학과 학사과정)
이담허((주)이르테크)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금남화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2022년 12월 12일
발행일: 2022년 12월 12일
인 쇄: 경대디지털

※ 이 보고서는 국립국어원의 용역비로 수행한 ‘2022년 맞춤법 교정 말뭉치 연구 분석’ 사업의 결과물을 발간한 것입니다.