메신저 맞춤법 교정 병렬 말뭉치의 구축과 쟁점

황은하, 안진산○, 남길임
배재대학교, 경북대학교, 경북대학교
behisson@gmail.com, siveking@naver.com, nki@knu.ac.kr

Construction of a Parallel Corpus for Instant Messenger Spelling Correction and Related Issues

HUANG YINXIA, Jin-san An○, Kil-im Nam
Pai Chai University, Kyungpook National University, Kyungpook National University

요 약

본 연구의 목적은 2021년 메신저 언어 200만 어절을 대상으로 수행된 맞춤법 교정 병렬 말뭉치의 설계와 구축의 쟁점을 소개하고, 교정 말뭉치의 주요 교정 및 주석 내용을 기술함으로써 맞춤법 교정 병렬 말뭉치의 특성을 분석하는 것이다. 2021년 맞춤법 교정 병렬 말뭉치의 주요 목표는 메신저 언어의 특수성을 살림과 동시에 형태소 분석이나 기계 번역 등 한국어 처리 도구가 분석할 수 있는 수준으로 교정하는 다소 상충되는 목적을 구현하는 것이었는데, 이는 교정의 수준과 병렬의 단위 설정 등 상당한 쟁점을 내포한다. 본 연구에서는 말뭉치 구축 시점에서 미처 논의하지 못한 교정 수준의 쟁점과 교정 전후의 통계적 특성을 함께 논의하고자 하며, 다음과 같은 몇 가지 하위 내용을 중심으로 논의하고자 한다.첫째, 맞춤법 교정 병렬 말뭉치의 구조 설계와 구축 절차에 대한 논의로, 2022년 초 국내 최초로 공개된 한국어 맞춤법 교정 병렬 말뭉치('모두의 말뭉치'의 일부)의 구축 과정에서 논의되어 온 말뭉치 구조 설계와 구축 절차를 논의한다. 둘째, 문장 단위로 정렬된 맞춤법 교정 말뭉치에서 관찰 가능한 띄어쓰기, 미등재어, 부호형 이모티콘 등의 메신저 언어의 몇 가지 특성을 살펴본다. 마지막으로, 2021년 메신저 맞춤법 교정 말뭉치의 구축 단계에서 미처 논의되지 못한 남은 문제들을 각각 데이터 구조 설계와 구축 차원의 주요 쟁점을 중심으로 논의한다. 특히 메신저 맞춤법 병렬 말뭉치의 주요 목표인 사전학습 언어모델의 학습데이터로서의 가치와 메신저 언어 연구의 기반 자료 구축의 관점에서 맞춤법 교정 병렬 말뭉치 구축의 의의와 향후 과제를 논의하고자 한다.

주제어: 한국어 맞춤법 교정, 메신저 말뭉치, 말뭉치 구축, 병렬말뭉치

1. 서론

메신저 말뭉치는 웹과 모바일상의 메신저 매체를 통해 둘 이상의 화자에 의해 이루어지는 의사소통을 수집, 가공한 말뭉치로, 국가 말뭉치로서는 [1]에서 원시 말뭉치로 처음 수집되었다. 이후, [2], [3], [4] 등에서 원시 말뭉치가 수집되었고, [5], [6] 등의 사업에서 각각 기구축된 메신저 말뭉치를 주석, 가공하는 시도도 이루어져 일부가 공개되기도 하였다. 이와 같이 최근 이루어진 메신저 말뭉치에 대한 일련의 구축 사업은, 웹 언어 장르로서 '메신저 언어'가 한국어 자원에서 차지하는 중요성을 보여준다.

2000년대 이후 비약적으로 증가한 웹 말뭉치에 대한 논의들이 역설하고 있듯이([6], [7], [8]), 구어와 문어의 이분법을 넘어서 제3의 언어로서의 웹 언어에 대한 연구, 특히 이미 전 국민의 95% 이상이 사용하고 있는[1] 메신저 언어에 대한 연구는 언어 이론 및 응용 분야 모두에서 중요하다. 최근 메신저의 형태 분석 방안 연구[9], 자동노이즈생성 기법을 활용한 맞춤법 오류 교정[10], 기계번역 방법론을 원용한 오류 교정 병렬 말뭉치의 활용[11] 등 메신저에 대한 전산언어학적 연구가 증가하는 추세이기는 하지만, 대다수의 연구가 소규모 말뭉치나 노이즈생성 데이터를 대상으로 한 것이어서([9], [10], [11]), 실제 메신저의 맞춤법 오류의 양상을 반영하고 있지 못하다.

본 연구의 목적은 2021년 실제 메신저 원시 말뭉치를 대상으로 수행된 맞춤법 교정 병렬 말뭉치 200만 어절의 설계와 구축의 쟁점을 소개하고, 이의 특성을 분석하는 것이다. 2021년 맞춤법 교정 병렬 말뭉치의 주요 목표는 메신저 언어의 특수성을 살림과 동시에 형태소 분석이나 기계 번역 등 한국어 처리 도구가 분석할 수 있는 수준으로 교정하는 다소 상충되는 목적을 구현하는 것이었는데, 본 연구에서는 말뭉치 구축 시점에서 미처 논의하지 못한 교정 수준의 쟁점과 교정 전후의 통계적 특성을 함께 논의하고자 한다.

1) 한국갤럽 <마켓70 2021>에 따르면, 2017년 이후 이미 전 국민의 메신저 이용률은 95%를 넘어섰고, 2020년 현재 70대를 제외한 전 연령층에서 90% 이상의 메신저를 사용하고 있는 상황이다. https://www.gallup.co.kr/gallupdb/reportContent.asp?seqNo=1205

본 연구는 다음과 같이 구성된다. 2장에서는 맞춤법 교정 병렬 말뭉치의 구조 설계와 구축 절차를 논의한다. [11]에서는 모델을 훈련하기 위한 맞춤법 교정 병렬 말뭉치가 존재하지 않는다고 보고하고 있으나, 2022년 초, 한국어 최초로 공개된 맞춤법 교정 병렬 말뭉치('모두의 말뭉치')가 공개되었는데, 여기서는 이 말뭉치의 구축 과정에서 논의되어 온 말뭉치 구조 설계와 구축 절차와 구조를 소개할 것이다. 3장과 4장은 맞춤법 교정 병렬 말뭉치의 쟁점과 관련된 부분이다. 3장에서는 문장 정렬 교정 말뭉치에서 관찰 가능한 띄어쓰기 오류 규모, 미등재어, 부호형 이모티콘 등의 고빈도 목록을 보임으로써, 향후 메신저 말뭉치의 구축과 활용의 방향성을 타진할 것이며, 4장에서는 대응쌍 정렬 단위와 교정의 수준의 쟁점을 사전학습 언어모델의 학습데이터 활용의 관점에서 논의하고자 한다.

2. 맞춤법 교정 병렬 말뭉치의 설계와 구축

2.1. 맞춤법 교정 병렬 말뭉치의 말뭉치언어학적 특성

교정형 병렬 말뭉치의 설계를 위해 고려해야 할 말뭉치의 특성은 아래 (1)과 같다.

(1) 맞춤법 교정 병렬 말뭉치의 특성
ㄱ. 언어학적 정보의 측면: 비규범형에 대한 교정 정보를 담은 '교정' 말뭉치
ㄴ. 말뭉치 구조적 측면: 원문과 교정문을 일정한 언어 단위로 정렬한 '병렬' 말뭉치
ㄷ. 메신저 언어의 장르적 측면: 메신저 텍스트의 언어적 특성을 보존한 '특수 장르' 말뭉치

우선, (1ㄱ)은 언어학적 정보의 측면에서 "교정 정보"를 담은 말뭉치를 전제로 할 때, 무엇을 교정 정보로 한정할 것이냐는 점이다. 원시 말뭉치에 나타난 비규범형을 대상으로 한다고 할 경우에도, 비규범형의 범주를 어디까지로 볼 것이냐는 말뭉치의 특성과 용도에 따라 달라질 수 있기 때문이다.

메신저 텍스트는, "말하는 것처럼 쓰며", "가능한 빨리 쓰기" 위한 여러 가지 전략을 통해 생성되며, 이 결과로 언어 단위와 형태론적 차원에서 특수성이 나타난다[10], [12]. 특히, 한국어 메신저 텍스트는 표음문자 특성상 "말하는 것처럼 쓴" 결과로, 형태를 살리지 않고 발음대로 적은 비규범형이 많으며, "가능한 빨리 쓴" 데서 비롯된 과도한 축약, 띄어쓰기 누락 등의 비규범적 형태가 다수 나타나는데,[2] 본 연구에서의 맞춤법 교정 말뭉치는 바로 이러한 메신저 텍스트의 비규범형을 맞춤법에 따라 교정한 말뭉치를 말하며, 다음의 항목을 교정의 주요 대상으로 분석한다.

(2) 주요 교정 정보의 유형
ㄱ. 한글맞춤법에 따른 띄어쓰기와 오탈자
ㄴ. <우리말샘> 미등재어(외래어, 신어)
ㄷ. 이모티콘, 자소 단위 표현 등 특수 표현
ㄹ. 방언
ㅁ. 문장부호의 교정

한편 위와 같이 하였을 경우에도, 교정 정보의 언어학적, 공학적 활용을 위해서는 메신저 텍스트 원문의 언어 정보를 보존하고, 원문과 교정문의 정보를 체계적으로 관리할 필요가 있다. 교정의 유형과 빈도 정보는 언어학적으로는 비규범형의 양적, 질적인 특성을 고찰하는 데 유용한 자료일 뿐만 아니라, 언어공학적으로 기계 학습과 자연어처리의 정확률을 향상시키는 데 중요한 자원이기 때문이다. 따라서 메신저 원시 말뭉치의 비규범적 형태를 규범적 형태로 전환하되, 원문과 교정문을 대응시켜 제시하는 형식이 필요하다.

다음으로 (1ㄴ)의 말뭉치 구조적 차원의 특성이다. 맞춤법 교정 말뭉치는 원시 언어와 주석되는 교정 언어가 모두 '언어'이며, 내용상 일치하다는 점에서 병렬 말뭉치(parallel corpus)의 일종이다. 단 일반적으로 병렬 말뭉치가 "둘 이상의 언어에 대해 원문과 번역문을 문장 등의 단위로 나란히 정렬한(align) 말뭉치"[13]로서, 둘 이상의 언어를 대상으로 한 것이었다면, 본 연구는 '단일 언어의 비규범형과 규범형에 대해 일정한 언어 단위로 나란히 정렬한 말뭉치'로 그 의미를 확장해 사용하기로 한다. 한편 일반적으로 이개어 말뭉치가 대조언어학, 번역학, 기계번역 등을 위해 구축되고 활용되어 온 것을 고려할 때, 본 연구의 말뭉치 역시 형태 분석과 구문 분석, 감성분석, 기계번역 등 메신저 텍스트의 확장된 활용을 염두에 두고 설계되어야 할 것이다. 이때 원문과 교정문의 대응 언어 단위의 구획의 문제, 교정 정보의 유형과 주석 정보의 층위와 설계의 문제 등이 쟁점이다.[3]

마지막으로 (1ㄷ)은 말풍선 단위의 구조를 가지는 텍스트 구조상의 특수성이나 온라인 담화를 기반으로 하는 메신저의 특성을 어디까지 고려할 것인가에 대한 문제로, 문어 또는 구어와는 다른 제3의 매체로서의 온라인 말뭉치의 특성, 새로운 장르로서의 메신저 텍스트의 특성을 어디까지 고려할 것인가의 문제이다. 이는 교정의 대상, 교정 범위의 문제이기도 하지만, 다른 한편 앞서 (1ㄴ)의 병렬 단위와 관련되기도 한다. 원문 정보가 병렬 말뭉치 내에 존재하기는 하지만, 극단적인 문어, 구어의 수준으로 텍스트를 모두 정제할 경우, 메신저 말뭉치의 특수성이 손상될 수 있다. 본 연구에서는 표준형 대응쌍의 존재 유무, 기계학습사전의 규칙 사전 활용 가능성, <우리말샘> 등재 여부 등에 따라 지침을 수립함으로써, 메신저 텍스트의 특수성을 유지하고자 하였다.

2) 안의정 외(2020)[9]에서는 이를 언어 단위의 특수성과 형태적 특수성으로 구분한 바 있다.

3) 실제로 이개어 병렬 말뭉치의 병렬 단위의 문제는 상당히 복잡한 문제로, 언어 간 병렬 말뭉치의 경우 활용도를 고려해 단어 단위의 정렬이 제안되기도 하였으나 작업의 효율과 언어 간 차이로 인한 대응의 현실적 가능성을 고려해 문장을 단위로 하는 것이 일반적이다[14].

2.2. 말뭉치 구축의 절차와 양식

위 2.1의 언어학적 특성, 말뭉치 구조적 특성, 메신저 장르 속성 등의 특성을 고려하여 말뭉치의 구축은 1) 지침 수립, 2) 메신저 원시 말뭉치의 전처리, 3) 자동 교정, 4) 수작업 교정, 5) 개인정보, 혐오 및 차별 표현 비식별화, 6) 검수, 7) JSON 구조화의 7단계로 이루어진다. 이를 도식화하면 다음의 그림과 같다.

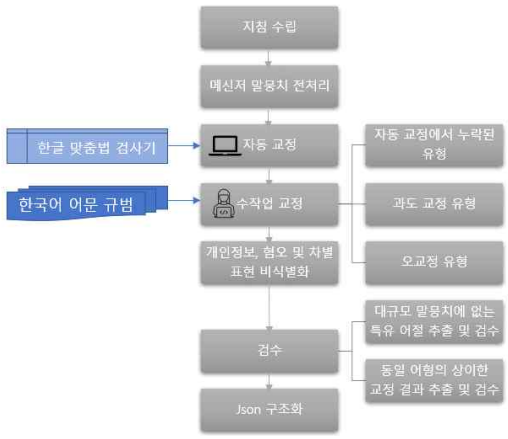

그림 1. 메신저 교정 병렬 말뭉치의 구축 절차

말뭉치의 결과물은 JSON 형식으로 구조화하는데, 원시 말뭉치의 구조를 계승하면서, 교정 정보를 표상할 수 있는 속성 “corrected_form”을 추가하고, 해당되는 값으로 원 문장에 대응되는 교정문을 제시한다. 구체적인 말뭉치 결과물의 양식은 다음과 같다.

```
  "correction" : [ {
    "form" : "엄마~",
    "speaker_id" : "1",
    "original_form" : "엄마~",
    "id" : "MDRW1900000012.1.1.1",
    "time" : "20191128 16:55",
    "corrected_form" : "엄마~."
  }, ...
  {
    "form" : "먹고싶어해요 침도흘리구용",
    "speaker_id" : "1",
    "original_form" : "먹고싶어해요 침도흘리구용",
    "id" : "MDRW1900000012.1.1.14",
    "time" : "20191128 16:57",
    "corrected_form" : "먹고 싶어해요. 침도 흘리고요."
  },
...
]
```

3. 맞춤법 주요 교정 및 주석 유형

맞춤법 주요 교정 및 주석 내용에 대한 관찰과 분석은 맞춤법 오류를 인위적으로 생성하는 자동 노이즈 생성 말뭉치의 구축을 위한 기초 자료로서 중요한 의미를 지닌다. 맞춤법 교정 말뭉치의 구축과 가공 내용에는 한국어문규범에 준해 맞춤법 오류를 교정하는 외에, <우리말샘>의 등재 여부를 기준으로 미등재어와 이모티콘, 자소 단위 특수 표현에 대한 주석도 포함한다. 여기서는 우선적 문장 단위 병렬 구조에서 추출할 수 있는, 띄어쓰기, 미등재어, 자소 단위 특수 표현을 중심으로 교정과 주석이 이루어진 내역을 살피고, 소규모 말뭉치를 대상으로 한 어절 단위 병렬 구조 테스트 결과도 함께 논의하기로 한다.

먼저, 본 과제에서 구축한 메신저 말뭉치는 461,365개의 말풍선이며, 교정 전후의 어절이 각각 1,493,275개, 1,789,982개로 교정 후 21.4% 정도가 증가했음을 알 수 있으며, 표 1.과 같다.

표 1. 연구 대상 메신저 말뭉치의 규모

	말풍선	어절	어절/말풍선
원문	461,365	1,493,275	3.24
교정문	461,365	1,789,982	3.88

교정 전후의 어절수 증가는 띄어쓰기 교정으로 인한 것으로, 띄어쓰기가 296,706개 늘어난 것으로 풀이된다. 여기에 띄어쓰기 오류에 불필요한 띄어쓰기를 제거하여 붙여 쓰는 경우까지 감안하면 띄어쓰기 교정 횟수는 더 증가한다.

메신저 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 이러한 미등재어에 대해 별도의 주석과 결과물에 대한 분석을 통해 그 목록을 확보하였으며, 그 중 5음절 이상 단위에 대한 고빈도 목록을 보이면 표 2.와 같다.

표 2. 고빈도 미등재어 목록

	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 2.를 보면 새로운 문물이나 개념과 함께 생겨난 신어(3, 6, 7, 8, 10, 11, 12, 14)가 많은데 주로 명사 또

는 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, 메신저 언어에는 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 자연어처리에서 자동 인식의 어려움이 있을 것으로 예상된다. 또한, 이모티콘에 대한 정확한 분석은 화자의 의도나 감정 등을 분석하는 데 유용하게 쓰일 수 있다. 아래에 문장부호, 특수기호, 자소 문자 등으로 이루어진 이모티콘 중에 일부 고빈도 목록을 보이면 표 3.과 같다.

표 3. 고빈도 부호형 이모티콘

	이모티콘	빈도	비고
1	ㅠ	15934	최다 58회 반복형
2	ㅜ	6513	최다 13회 반복형
3	^^	2557	
4	ㅡㅡ	679	
5	;	661	최다 12회 반복형
6	ㅠㅜ	293	
7	ㅜㅠ	149	
8	ㅇㅅㅇ	116	
9	^^;	104	
10	><	90	

이에 더해, 메신저 언어에는 입력의 편의를 위한 'ㅋㅋㅋ'와 같은 자소형 약어 형식의 비표준적 표기 또한 매우 다양하게 고빈도로 등장하며, 'ㅠㅠㅋ'처럼 부호형 이모티콘과 자소형 약어의 혼합형도 적지 않다.

한편 이와 같은 문장 단위 병렬 구조의 한계를 보완하고 오류 유형을 보다 면밀히 살피기 위해, 본 연구에서는 15,000어절에 대해 chunck 단위[4]에 대한 병렬 및 교정하는 작업을 시험적으로 수행하였는데 그 결과를 간략히 소개하면 다음고 같다.

교정의 유형을 문자열 또는 문장부호에 대해 누락된 것을 추가하는 '삽입(insert)', 불필요한 것을 '삭제(delete)'하는 것, 잘못된 것을 '교체(change)'하는 것으로 구분하고, 띄어쓰기(space)와 붙여쓰기(link)를 적용한 교정 작업의 결과는 다음과 같다.

표 4. 맞춤법 교정 작업의 유형별 규모

교정 대상	교정 유형	교정 빈도	비율
띄어쓰기	space	3,685	40.6%
	link	64	0.7%
문자열 또는 문장부호	insert	2,841	31.3%
	change	2,005	22.1%
	delete	474	5.2%
합계		9,069	100.0%

이를 통해 '띄어쓰기>삽입>교체'의 순으로 교정의 분포가 이루어진다는 것을 알 수 있다. 특히 아래 표 5.에서 제시된 바와 같이 삽입 교정의 주요 내용은 주로 문장부호로 나타났다는 점은 문장 단위 구획에서 교정 말뭉치의 활용 가능성을 보여준다.

표 5. 삽입 교정의 주요 유형 및 유형별 규모

	교정 내용	교정 빈도	비율
1	.	1971	69.4%
2	,	636	22.4%
3	?	182	6.4%
4	"	17	0.6%
5	."	4	0.1%
	...	...	0.0%
총합계		2841	100.0%

한국어문규범에서도 한글맞춤법(띄어쓰기 외)과 외래어표기법과 관련한 문자열의 주요 오류 유형은 문장보다는 작은 단위의 교정 전후 대응이 이루어질 때 비로소 가능하며, 그 필요성에 대해서는 4장에서 다루기로 한다.

4. 맞춤법 교정 병렬 말뭉치 설계의 쟁점

4.1 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 노력과 시간은 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.

먼저, 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_BDRCBC, 南京大学英汉双语平行语料库),

4) 어절은 "문장을 구성하고 있는 각각의 마디. 문장 성분의 최소 단위로서 띄어쓰기의 단위가 된다."(표준국어대사전, 1999)로 정의되며, 이에 따르면 "먹고싶어해요 침도 흘리구용"과 같은 교정 전 원시문은 띄어쓰기 하나로 분리되어 있지만, 2어절로 보는 것에는 무리가 있다. 따라서 이와 같은 단위를 자연어처리에서 흔히 사용하는 용어인 '부분구문', 즉 'chunck'로 이르기로 한다.

“21세기 세종 계획” 병렬 말뭉치[5] 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[15].

한편 메신저 텍스트는 앞선 병렬 말뭉치의 원문들과 구별되는 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.

(3) 화자 1-1): 그래서 약이 또 달라 좋겠네
화자 2-1): 약빨이 잘받아야하는데..
화자 2-2): 나아지는듯 하더니 더심해지는거같아서
화자 2-3): 가습기 다시켰어요

(3)의 2-1)은 2-1)이 연결어미로 끝을 맺고 있기 때문에 중의적인 해석이 가능하다. 2-2)와 2-3)은 인과 관계의 문장으로 보기에 무리가 없는데, 2-1)이 하나의 절로서 이들과 같은 문장을 이루는지, 아니면 1-1)에 대한 응답으로서 말줄임표로 끝난 것인지 하는 두 가지 해석의 가능성이 있다. 이와 같이 메신저 텍스트에서 문장 단위의 구획의 불명확성과 중의성, 메신저 텍스트의 특성 유실, 작업의 일관성 준수의 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.

(4) [문장 대응] “날리진 않겠져?/날리진 않겠죠?”, “난 두번 썻는뎁ㅋㅋ/난 두 번 썼는데 ㅋㅋ”
[어절/청크 대응] “않겠져/않겠죠”, “두번/두 번”, “썻는뎀/썼는데”
[형태소 대응] “여/요”, “ㅅ/ㅆ”, “는뎀/는데”

(4)에서 보인 것처럼 어절 또는 청크(chunck)[6] 단위의 정렬은 띄어쓰기가 구분자 역할을 하므로 획분이 용이하고, 말풍선이나 문장보다 단위가 작아서 원문-교정문에 대한 보다 세밀한 대응 정보의 추출이 가능하기 때문에 그에 대한 시도가 요구된다. 이는 특히 최근 어텐션을 중시하는 딥러닝 모델들에서 유용한 힌트 데이터로 활용될 수 있다. 한편 교착어로서의 한국어의 특성을 반영할 때, 형태소 단위의 대응을 통해 맞춤법 오류 정보를 보다 정확하게 관찰할 수 있다. 이밖에, 엔그램 등을 활용한 구나 청크 단위의 대응쌍도 고려할 수 있는데, 언어의 유형론적 특성을 고려한 단위 설정에 대한 연구가 필요하다.

4.2 교정 범위와 수준

맞춤법 교정 병렬 말뭉치의 ‘맞춤법’은 실제로는 한글 맞춤법, 표준어 규정, 외래어 표기법까지 포함한 한국어 어문 규범[7]이라고 하는 것이 타당하다. 다만, 표준어 규정의 사정 원칙이나 외래어 표기법에는 표기 규칙과 더불어 전형적인 어휘만 예시로 보이고 있으므로, 구체적인 낱말의 표기에 대해서는 사전을 기준으로 삼을 필요가 있는데, 본 사업에서는 <우리말샘>을 그 기준으로 삼고 있다. 기존의 자연어 처리 도구가 일련의 언어 규칙 외에 사전 및 통계적 기법을 활용하는 하이브리드한 방법으로 개발된 점을 감안해도, 본 말뭉치는 교정의 기준을 맞춤법 준수 여부뿐만 아니라 사전으로 확대할 필요가 있다. 그런데 어문규범의 보수성으로 인해 엄격한 맞춤법 교정을 하는 경우, 메신저 언어의 창조성이 훼손되기도 하고, <우리말샘>의 불완전함과 더불어 수시로 업데이트되는 역동성 때문에 완전하고 일관된 교정이 사실상 매우 어렵다. 여기에 여러 작업자의 협업이라는 작업의 특성상 일관된 띄어쓰기 준수가 현실적으로 어려운 부분도 있다.

(5) 띄어쓰기 판단이 필요한 유형
- 사전 등재 어형: 양꼬치, 염통구이, 삼겹살집, 감자탕집...
- 사전 미등재 어형: 염통꼬치, 순대꼬치, 양꼬치 집, 소고깃집...

(5)는 모두 ‘명사+명사’의 합성어 또는 구의 문제로, 여기에서 두 가지 문제가 제기된다. 첫째, 사전에 ‘양꼬치’가 등재되어 있고 ‘염통꼬치’가 등재되어 있지 않다고 하여 사전 등재 여부라는 기준을 잣대로 ‘염통 꼬치’로 띄어 쓰는 것이 과연 타당한가 하는 문제이다. 둘째, 띄어쓰기의 일관성 준수가 현실적으로 매우 어렵고 시간 소모적이라는 문제점이 있다.

표 6. 사전 미등재 고유명사의 변이형과 그 형태의미 분석 및 기계번역 결과

5) 문장 단위로 정렬하더라도 1:1 대응 외에 0:1, 1:0 1:다, 다:1, 다:다의 다양한 대응쌍이 만들어지는 것이 현실이다. 세종 한영 병렬 말뭉치의 경우 한국어와 영어의 문장 대응이 1:26인 경우도 있는 것으로 나타났다.

6) 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 자연어처리에서 흔히 사용하는 용어인 ‘부분구문’, 즉 ‘청크’로 하는 것이 더 타당하다고 하겠다.

7) 한국 어문 규범은 로마자 표기법까지 포함해 4칙 규범이나, 로마자 표기법은 한글 텍스트를 로마자화하는 규칙이므로 본 연구에서는 적용의 대상이 아니다.

어형	형태의미 분석	기계번역	
	Utagger	한중	한영
겨울왕국2	겨 울 / N N G + 왕 국 __000001/NNG+2/SN	冰雪奇緣2	Frozen 2
겨울 왕국2	겨 울 / N N G + 왕 국 __000001/NNG+2/SN	冰雪奇緣2	Frozen 2
겨울 왕국 2	겨 울 / N N G + 왕 국 __000001/NNG+2/SN	冰雪奇緣2	Frozen 2
겨울왕국 2	겨 울 / N N G + 왕 국 __000001/NNG+2/SN	冰雪奇緣2	Frozen 2

표 6.은 메신저 말뭉치에서 다양한 띄어쓰기 변이형으로 실현된 영화 제목 '겨울왕국 2'의 예인데, 형태의미분석 결과는 모두 일치하며, 기계번역은 모두 정확하다. 기계번역을 위한 전자사전이 인간 독자를 대상으로 한 사전과는 달리 고빈도 고유명사와 그 대역 정보에 대한 수록을 적극적으로 하고 있고, 언어 간 병렬 말뭉치에서도 고빈도로 출현하기 때문에 정확한 대역 결과가 출력 가능한 것으로 풀이된다.

이에 대해 맞춤법 검사기와 같은 특수 활용을 고려하면 고빈도 어형에 대해 특히 일관성을 준수하는 것이 타당하다고 하겠으나, 교정의 일관성에 대한 관점은 말뭉치의 활용 목적에 따라 상이한 관점이 있을 수 있다. 예로 맞춤법 검사기 개발이나 규범문법적 관점에서는 정답 세트로서의 고도의 일관성을 추구하는 반면, 기계번역의 공학적 활용이나 기술언어학적 관점에서는 완벽한 일관성의 추구가 노력 대비 효과가 크지 않은 것으로 판단한다. 오히려 기계번역의 관점에서는 교정 데이터를 이용한 맞춤법 교정기를 기계번역 엔진의 전처리기로 사용할 경우, 명사+명사 띄어쓰기, 고유명이나 외래어에 대한 완벽한 일관된 교정은 오히려 효용성의 측면에서 제한적일 수 있다. 교정된 단어를 기준으로 구축한 번역 말뭉치로 기계번역 엔진을 학습한 경우 도움이 될 수도 있겠지만 기구축한 번역 말뭉치에서 교정 단어와 다르게 표현되어 있었다면 오히려 더 악영향을 미칠 가능성도 있기 때문이다.

5. 결론

본 연구에서는 메신저의 특수성을 고려한 메신저 맞춤법 교정 병렬 말뭉치의 구축 과정과 메신저 말뭉치의 언어공학적 활용을 고려한 구축의 쟁점을 논의하였다. 현재 맞춤법 교적 병렬 말뭉치는 <국립국어원> 메신저 말뭉치, 웹 말뭉치, 온라인 대화 말뭉치를 대상으로 700만 어절 가량이 구축되고 있으며, 이 중 일부는 '모두의 말뭉치'로 공개되었다.

본 연구를 통해 문자의 영역으로 넘어온 대화, '메신저'의 언어학적 특성을 고려한 맞춤법 병렬 말뭉치의 구축에서 주요 쟁점이 무엇인지에 이에 대한 윤곽은 어느 정도 확인이 되었다고 생각된다. 병렬 구조, 병렬 대응쌍 단위, 교정 수준의 문제 등은 이 말뭉치가 메신저 언어를 대상으로 하고 있다는 점, 이개어가 아닌 특수 어형을 포함한 한국어 병렬 말뭉치라는 점을 고려하여 보다 정교화되어야 할 것이다. 향후 문어, 구어 수준과 평행한 형태, 통사, 의미, 화용 주석 말뭉치의 구축, 공학적 활용도를 고려할 때, 메신저 말뭉치의 병렬 구조, 교정의 수준 등에 대한 논의가 지속될 필요가 있다. 특히 현재까지 국립국어원 말뭉치의 구조적 일관성을 고려하여 말풍선 단위의 병렬 구조로 구축된 구조는 향후 말뭉치의 활용도를 고려하여 어절 대 어절, 문장 대 문장 등의 가능성이 모색될 필요가 있다.

참고문헌

[1] 국립국어원, 메신저 대화 자료 수집 및 말뭉치 구축, 2019.

[2] 한국지능정보사회진흥원, 한국어 SNS 데이터셋, 2020.

[3] 국립국어원, 2021년 온라인 대화 말뭉치 구축, 2021.

[4] 국립국어원, 2020년 어휘의미분석 말뭉치, 2020.

[5] 국립국어원, 2021년 맞춤법 교정 말뭉치, 2021.

[6] Baroni, M., & Ueyama, M., Building general-and special-purpose corpora by web crawling. In Proceedings of the 13th NIJL international symposium, language corpora: Their compilation and application, pp.31-40, 2006.

[7] Gatto, M. Web as corpus: Theory and practice. A&C Black, 2014.

[8] 남길임, 이수진, 최준, 웹 말뭉치를 활용한 의미적 신어의 연구 동향과 쟁점, 한국사전학 31, 한국사전학회, pp.55-84, 2018.

[9] 안의정, 송현주, 김진웅, 형태 분석을 위한 메신저 텍스트 처리 방안, 텍스트언어학 49, 한국텍스트언어학회, pp.27-52, 2020.

[10] 구선민, 박찬준, 소아람, 임희석, 딥러닝 기반 한국어 맞춤법 교정을 위한 오류 유형 분류 및 분석, 한국융합학회논문지 12, 한국융합학회, pp.65-74, 2021.

[11] Park, C., Park, S., & Lim, H., Self-Supervised Korean Spelling Correction via Denoising Transformer. In 7th International Conference on Information, System and Convergence Applications, 2020.

[12] Vandekerckhove, R., & Nobels, J., Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers 1. Journal of sociolinguistics, 14(5), pp.657-677, 2010.

[13] Granger, S., & Tyson, S., Connector usage in the English essay writing of native and non-native EFL speakers of English. World Englishes, 15(1), pp.17-27, 1996.

[14] 홍윤표 외, 21세기 세종계획 국어정보화 중장기 발전계획 연구보고서, 국립국어원, 1997.

[15] 황은하, 말뭉치 기반 한외(韓外) 대조언어학 연구에 대한 일고찰, 어문론총, 69권, pp.39-72, 2016.