한글 문서에서 형태적 중의 오류의 교정

김 민 주*, 정 준 호*, 이 현 주**, 최 재 혁***, 김 항 준*, 이 상 조*
*경북대학교 컴퓨터공학과, **경북대학교 국어국문학과, ***신라대학교 컴퓨터교육과

A method for morphological correction of ambiguous error

* : Dept. of Computer Engineering, Kyungpook National University
** : Dept. of Korean Language & Literature, Kyungpook National University
*** : Dept. of Computer Education, Silla University

요 약

교정 시스템에 나타나는 오류 유형들 중에는 전체적인 교정률에 차지하는 비중은 적지만 출현할 때마다 틀릴 가능성이 아주 높은 오류들이 있다. 기존의 교정 시스템에서는 이러한 오류들에 대한 처리가 미흡한데, 철자 오류와 띄어쓰기 오류 중 형태가 비슷하거나 같은 형태가 다른 기능을 함으로써 발생하는 오류들이다. 이러한 오류는 일반 문서 작성자뿐만 아니라 한글 맞춤법에 대해 어느 정도 지식을 가진 사람의 경우에도 구분이 모호하다. 복합 명사와 미등록어를 제외한 오류 중 약 30%가 여기에 속한다. 따라서 본 논문에서는 이러한 오류 유형들을 분류하고, 이 중에서 빈번하게 출현하는 오류에 대한 교정을 시도하고, 오류 유형들이 문장 내에서 어떤 분포를 가지는지 알아본다. 약 617만 어절의 말뭉치를 이용하여 해당 형태와 다른 성분들과의 관련성을 조사하여 교정 방법을 제시하고, 형태소 분석을 하여 교정을 행한다. 코퍼스 655만 어절 대상으로 실험한 결과 84.6%의 교정률을 보였다. 본 논문에서 제시한 교정 방법은 기존의 교정 시스템에 추가되어 교정 시스템의 전체 교정률을 향상시킬 수 있다. 또한 이와 비슷한 유형의 다른 어휘 교정에 대한 기초 자료로 사용될 수 있을 것이다.

1. 서 론

컴퓨터의 대중화와 문서 편집기의 보급으로 컴퓨터를 이용한 문서의 작성이 증가하게 되었다. 이에 따라 문서의 오타나 오류를 교정해 주는 자동 오류 교정 시스템의 필요성이 증가하게 되었고, 이러한 교정 시스템은 기계 번역, 문자 인식, 정보 검색 등의 다른 응용 분야에서도 다양하게 응용되고 있다.

한글 문서에 나타나는 일반적인 오류는 맞춤법 오류, 구문 오류, 의미 오류로 분류된다[10]. 이 중에서 구문 및 의미 오류를 찾아내기 위해서는 구문 분석이나 의미 분석의 단계를 거쳐야 한다. 최근 들어 부분 구문 분석과 연어 정보에 기반한 교정 시스템이 연구되고 있지만[1,2,7], 일반적으로 한 어절이나 좌우의 한 어절을 대상으로 오류를 처리하고 있다. 대표적인 교정 방법으로는 자판 운지 거리 및 발음 유사성을 고려한 음소 대치 방법, 음절 간의 상호 정보를 이용하는 방법, 형태소 분석을 이용하는 방법 등으로 맞춤법 오류를 발견하여 교정하는 시스템들이 대부분이다[7,11,12].

기존의 오류 교정 시스템에서는 이러한 방법으로 전체적인 교정률이 향상되었으나, 일반 문서 작성자뿐만 아니라 한글 맞춤법에 대해 어느 정도 지식을 가진 사람들조차 자주 틀리는 특별한 어휘들에 대한 처리는 아직 미흡한 실정이다. 예를 들어 '로써/로서'는 용법이 다름에도 불구하고

형태가 비슷하여 문서상에서 잘못 사용되는 경우가 많으며, '-ㄴ데/-ㄴ 데'의 '데'는 같은 형태를 가지지만 각각 어미의 일부와 의존명사로 사용되어 띄어쓰기에 혼란을 가져온다.

이들 어휘는 다른 오류 어휘들과는 달리 출현할 때마다 틀릴 가능성이 아주 높기 때문에 전체적인 교정률에서 차지하는 비중은 적지만 이들의 교정은 중요한 의미를 지닌다고 할 수 있다. 실제로 교정 시스템에서 복합 명사와 미등록어를 제외한 오류 중 약 30%가 여기에 속한다.

본 논문의 목적은 철자 오류와 띄어쓰기 오류 중에서 형태가 비슷하거나 같은 형태가 다른 기능을 함으로써 발생하는 오류들의 유형을 분류하고, 이들 중에서 빈번하게 출현하는 오류의 교정 방법을 제시하는 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 오류 유형별 분류에 대해서 알아보고, 3장에서는 2장에서 분류된 유형들의 교정 방법에 대해서 알아본다. 4장에서 코퍼스에서 나타난 오류 유형별 문장 분포와 실험을 통해 제시된 교정 방법의 교정률을 알아보고, 5장에서는 결론을 맺는다.

2. 오류 유형별 분류

2.1 중의 오류

이 장에서는 기존의 교정 시스템에서 다루어지지 않았던 철자 오류와 띄어쓰기 오류의 일부분을 모아 유형별로 제시한다.

기존의 교정 시스템에서는 주로 자판 오류나 맞춤법 지식 부족으로 인한 오류에 대한 전체적인 교정을 시도해 왔다. 사람들이 자주 틀리는 특별한 어휘에 대한 처리는 미흡한 실정이다. 형태가 비슷하거나 같은 형태가 다른 기능을 함으로써 혼동을 일으키는 어휘들은 다른 오류 어휘들과는 달리 출현할 때마다 틀릴 가능성이 아주 높음에도 불구하고 이들에 대한 명확한 구분 방법을 제시하기 어렵기 때문이다.

이러한 어휘의 처리에 대해서 [2]에서 '접미사/의존명사/어미'의 혼용 오류를 처리할 수 있지만 극히 제한된 경우에 한하고 있다.

표 1은 형태가 비슷하거나 같은 형태가 다른 기능을 함으로써 두 개 중에 하나를 선택하는데 있어서 발생하는 철자 오류나 띄어쓰기 오류의 유형들을 나타낸 것이다. 본 논문에서는 이러한 오류 유형들을 위해 중의 오류라는 명칭을 사용하기로 한다.

표 1 중의 오류의 유형과 용례

	유 형	용 례
철자 오류	의존형태소	'로써/로서', '(음)으로/므로', '던지/든지' 등
	자립형태소	'반드시/반듯이', '지그시/지긋이' 등
띄어쓰기 오류	의존명사/어미의 일부	데, 듯, 바, 지, 걸
	의존명사/조사	만큼, 대로, 만
	의존명사/접미사	뿐, 녘, 것, 나름, 체, 채, 적, 이래, 노릇, 무렵, 빨, 쪽, 차, 통, 품
	의존명사/부사/명사	족족
	동사/접미사	되다
	동사/보조동사	돌아가다, 돌아오다, 걸어가다, 걸어오다, 꾸며내다 등

맞춤법 오류는 크게 철자 오류와 띄어쓰기 오류로 구분할 수 있는데, 본 논문의 대상이 되는 철자 오류는 서로 다른 용법으로 쓰이지만 비슷한 형태를 가짐으로써 생기는 사용자의 중의 오류를 말한다.

철자 오류는 의존형태소 오류와 자립형태소 오류로 나눌 수 있다. 예를 들어 '로써/로서', '(음)으로/므로' 등의 의존형태소는 그 선행요소가 문장 내에서 어떤 의미로 사용되었는가에 따라 구별이 되며, '반드시/반듯이', '지그시/지긋이'와 같은 자립형태소는 그 문장의 다른 성분들과의 구문관계에 따라 하나를 선택해야 하는 것이다. 한국어는 조사나 어미와 같은 기능어가 발달되어 있으므로 의존형태소 오류가 자립형태소 오류보다 더 빈번히 발생하기 때문에 본 논문에서는 철자 오류 중에서 의존형태소 오류의 교정 방법을 제시하고자 한다.

띄어쓰기 오류는 가장 빈번하게 나타나는 오류의 형태로서[3] 띄어 써야할 것을 붙여 쓴 오류와 붙여 써야할 것을 띄어 쓴 오류의 형태로 나타난다. 즉, 본 논문에서 의존명사의 경우에는 앞말과 띄어야 하나 붙여 쓴 오류에 속하고, 어미의 일

부, 접미사, 조사 등은 붙여 써야 할 것을 띄어 쓴 오류에 속한다. 예를 들어 '데'는 단독 혹은 조사나 어미와 함께 쓰여 한 어절을 이룰 때 의존명사로 사용되고, 앞말과 함께 어미 '-ㄴ데'의 일부로 사용되면 앞말에 붙여 쓴다.

'의존명사/조사'는 의존명사와 조사의 역할을 공유하는 유형들을 모은 것이다. 예를 들어 "그만큼 열심히 해라"에서 '만큼'은 체언 뒤에 나타나 조사로 사용되었으므로 앞말과 붙여 써야 한다. "내가 생각하는 만큼"에서 '만큼'은 용언의 관형사형 어미 뒤에 나타나 선행문의 형편, 처지에 한정되어 의존명사로 사용되었으므로 앞말과 띄어 써야 한다. 마찬가지로 "좋을 대로 해라"에서 '대로'는 내포문이 한정하는 형편·처지·상황과 동일한 의미로 의존명사로 사용되었으므로 앞말 '좋을'과 띄어 써야 한다. 그러나 "마음대로 하십시오"에서는 조사로 사용되어 앞말과 붙여 써야 한다[5].

'의존명사/접미사'는 의존명사와 접미사의 역할을 공유하는 유형들을 모은 것이다. 예를 들어 '녘'은 "해가 질 녘에 출발했다."에서 어떤 형편·처지·상황에 놓여 있는 시공간의 방향을 나타내는 뜻을 가질 때는 의존명사의 역할을 하므로 앞말과 띄어 쓴다. 그러나 "들녘, 아랫녘"에서는 어떤 쪽(방향)을 나타내는 접미사로 사용될 때는 앞말과 붙여 써야 한다.

'되다'는 동사가 되는 경우가 있고, 접미사가 되는 경우가 있다. "결정하다", "해결하다"에서처럼 '-하' 접미사가 붙어 동사가 되는 말에 '-하' 대신 '-되'가 붙는 "결정되다, 해결되다"의 '-되'는 동사 파생 접미사이므로 붙여 쓴다. 그러나 "문제 되는"은 명사 뒤에 '되다'가 나오는 위의 "결정되다, 해결되다"와 같은 형태를 가지지만, "문제 되는"에서의 '문제'는 동사인 '되다'의 보어로서 조사가 생략된 형태이므로 '문제'와 '되다'는 띄어 써야 한다[3].

3. 유형별 교정 방법

본 장에서는 2장에서 설명한 중의 오류 중 출현 빈도가 높다고 생각되는 철자 오류 '로써/로서'와 띄어쓰기 오류 중에서 '의존명사/어미의 일부'로 쓰이는 '데, 듯', '의존명사/조사' 유형인 '뿐, 만큼, 대로', '의존명사/접미사' 유형인 '녘, 나름, 채, 노릇, 무렵, 쪽' 그리고 '동사/접미사' 유형의 형태적 교정 방법을 제시한다. 코퍼스를 이용하여 위의 형태들이 출현한 문장들을 추출하고, 해당 형태와 다른 성분들과의 관련성을 조사하여 교정 방법을 제시한다.

본 논문에서 교정 방법을 찾기 위해 이용한 코퍼스는 97년 한국과학기술원(KAIST)에서 만든 대한민국 국어 정보베이스의 한국어 텍스트 코퍼스 약 615만 어절과 동아일보 사설 1만 6천 어절이다.

3.1. '로써/로서'의 교정 방법

이 절에서는 일반적으로 틀리기 쉬운 철자 오류의 예인 '로써/로서'의 교정 방법을 제시한다. 코퍼스에서 추출한 '로써/로서'를 포함한 문장 약 2만 개를 분석한 교정 방법은 다음과 같다.

교정 1. 용언의 명사형 + '로써'
교정 2. 사람의 의미자질을 가지는 명사 + '로서'
교정 3. 상용구

교정 1에 나타난 모든 형태에 조사 '로서'가 잘못 사용된 경우 '로써'로 교정한다. 교정 2는 자격을 나타낼 수 있는 사람의 의미자질을 갖는 명사가 선행될 때 '로서'로 교정하는 방법이다. 모든 명사에 대한 의미자질을 부여하는 데에는 한계가 있다. 따라서, 본 논문에서는 모든 명사에 의미자질을 부여하지 않고 사람의 의미자질을 가지는 명사에 대한 정보를 사전에 두고 이를 이용한다. 교정 3은 국어사전에는 나타나지 않는 어휘이나 일반적으로 많이 사용하는 형태를 묶은 것이다.

예 1 ① 저 돌을 움직임으로써 문을 열 수 있다.
② 교육자로서 일생을 보내다.
③ 현재로서, 당시로서, 때로서

그러나 사람의 의미자질을 가지는 명사를 제외한 일반 명사, 추상명사들, 예를 들어 '말(言) 것, 이치, 이성' 등 10,995개의 문장들은 선·후행 문

장 성분을 보고, 구문 분석을 통해서 교정해야 한다.

3.2. '-ㄴ데/-ㄴ 데'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '의존명사/어미의 일부'의 예인 '-ㄴ데/-ㄴ 데'의 교정 방법을 제시한다. 코퍼스에서 추출한 '-ㄴ데/-ㄴ 데'를 포함한 문장 약 7,500개를 분석한 교정 방법은 다음과 같다.

교정 1. '데[조사]'+'있다/없다'의 활용형 또는 그 파생어
교정 2. '데' + 격조사 혹은 어미
교정 3. '데[조사]' + 문장 부호
교정 4. 체언류 + '인데'
교정 5. 상용구

교정 1에서 '데'는 의존명사로 사용되어 앞말과 띄어 쓴다. '데'가 '것에, 곳에'로 바꾸어 사용해도 말이 될 때 의존명사로 사용된 경우인데 '-에 있다/없다'가 선행되는 말은 주로 장소에 관련된 말이 오기 때문이다. 교정 2는 '데'의 역할에 따라 나타나는 조사 혹은 어미를 묶은 것이다. '데'에 조사인 '서, 로, 가, 를, 만, 는, 에' 등이 결합하면 의존명사로서 앞말과 띄어 쓴다. '데'에 종결형 어미 '-요'와 특수 조사 '-도' 등이 결합하면 어미의 일부로 사용되어 앞말과 붙여 쓴다. 교정 3에서 '데'는 어미의 일부로 사용되어, 문장 부호 ', . !' 등이 결합되면 절(문장)의 끝을 의미하여 앞말과 붙여 쓰는 방법이다. 교정 5는 문서에서 일반적으로 많이 나타나는 상용구를 묶은 것이다.

예 2 ① 목적은 지식을 얻는 데 있다.
② 그를 설득하는 데에 며칠이 걸렸다.
③ 곰곰이 생각해 봤는데, 그 일은 …
④ 오늘은 휴일인데 회사에 왜 나왔니?
⑤ '- 데[에] 대한', '- 데[에] 비해',
'- 데 [에] 반해',
'-하는 데[에] 대해',
'- 데[에] 대해서',
'-데[에]도 불구하고' 등

그러나 '데'가 복합문의 절의 경계를 나타낼 때, 문맥상 주어와 서술관계가 성립하는데 아래의 예 ⑥과 ⑧에서 어미의 일부로 사용된 '데'와 의존명사 '데'를 분석할 수 있다. 그러나 예 ⑦에서처럼 우리말의 생략 현상으로 인해 주어가 생략된 문장은 교정할 수 없다. 따라서 2,960개의 문장에 대해서 문장 성분만을 고려한 구문 분석 방법으로는 교정하기 어렵다.

⑥ 눈이 펑펑 쏟아지는데 그가 왔다.
⑦ 형편이 어려운데 계속 공부한다.
⑧ 그것을 이해하는 데 도움이 되지 않는다.

3.3. '듯'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '의존명사/어미의 일부'의 예인 '듯'의 교정 방법을 제시한다. 코퍼스에서 추출한 '듯'을 포함한 문장 약 6,500문장을 분석한 교정 방법은 다음과 같다.

교정 1. 듯 + '-하다/-싶다'의 활용형
교정 2. 용언의 관형사형 어미 아래에서 띄어 쓴다.

'듯'은 의존명사와 어미의 일부로 함께 사용되어질 수 있기 때문에 띄어쓰기에 혼란을 가져온다. 의존명사 '듯'은 용언의 관형사형 어미 '-ㄴ, -ㄹ' 아래에 쓰여 '그럴 것 같기도 하고, 그렇지 않을 것 같기도 하다'는 뜻으로 사용된다. '듯'이 어미의 일부로 쓰일 때는 '듯이'의 준말로 각 어간에 붙어, '비슷하거나 같은 정도'의 뜻을 나타낸다. 교정 1에서 '듯'이 '-하다/-싶다'와 결합하여 보조 용언이 되므로 '듯'의 선행 요소와 띄어 쓴다.

예 3 ① 비가 올 듯싶다.
② 자는 듯 마는 듯.
③ 새가 하늘을 날듯이 …

교정 1을 적용했을 때, '날다, 살다, 길다'와 같

이 어간의 끝음절 종성이 'ㄹ'인 단어들은 형태소 분석시 모호성이 발생한다. 예를 들어 예3의 ③에서 '날듯이'는 '나(出)+ㄹ듯이'와 '날(飛)+듯이, 날(飛)+ㄹ듯이'로 형태소 분석된다. 어간에서 발생하는 형태소 분석 결과의 모호성은 구문 분석을 이용할 경우 해소될 수 있다. 그러나 아래 예 ④과 ⑤에서 '날듯'은 각각 '날(飛)+듯'과 '날(飛)+ㄹ듯'으로 구분된다. 이것은 ⑤의 '날'이 미래에 일어날 일에 대한 예정의 의미가 함축되어 있기 때문이다. 따라서 '날(飛)+듯'인지 '날(飛)+ㄹ듯'인지를 결정하기 위해서는 앞뒤 문맥이나 상황을 고려해야 한다.

④ 새가 공중을 날듯이.
⑤ 날 듯이 가벼운 기분.

3.4. '뿐, 만큼, 대로'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '의존명사/조사'의 예인 '만큼, 대로'의 교정 방법을 제시한다. 여기서 '뿐'은 '의존명사/접미사'의 예지만 '만큼'의 처리 과정과 유사하여 같이 설명한다. 코퍼스에서 추출한 '뿐, 만큼, 대로'를 포함한 각각의 문장 8,472개, 3,573개, 3,875개를 분석한 교정 방법은 다음과 같다.

교정 1. 용언의 관형사형 어미 아래에서 띄어 쓴다.

'뿐, 만큼, 대로'는 의존명사와 접미사 혹은 조사의 역할을 공유한다. 의존명사인 '뿐, 만큼, 대로'는 용언의 관형사형 어미 '-ㄴ, -ㄹ'이 선행되면 반드시 띄어 쓰고, 체언 뒤에 나타날 경우 접미사와 조사로서 체언과 붙여 써야 한다. 예를 들면 다음과 같다.

예 4 ① 끊임없는 투쟁만이 있을 뿐이다.
② 살아남은 사람은 그뿐이다.
③ 그녀를 연상할 만큼 닮았다.
④ 어머니라는 호칭만큼 그리운 말도 드뭅니다.
⑤ 그가 바라는 대로 행동한다.
⑥ 네 뜻대로 해라.

예 5 ① 그건 하나의 표현일 뿐이야.
② 온통 집안일뿐이야.
③ 재산 공개가 주 이슈인 만큼…
④ 미술 애호인만큼 미술을 사랑하는…

그러나 예5의 ①과 ②의 '표현일'과 '집안일'에서 '일'이 형태는 같지만 각각 '이-(서술격 조사)+-ㄹ'과 '일(事)'로 사용된다. 마찬가지로 예5의 ③, ④에서도 '인'이 '이-(서술격 조사)+-ㄴ'과 '인(人)'으로 사용될 수 있기 때문에 띄어쓰기에 모호성이 발생한다. 그런데 '집안일일 뿐이야'나 '그녀는 미술 애호인인 만큼'과 같이 '일'이나 '인'이 연속되어 나타나면, 두 번째 '일, 인'은 분명히 '이-(서술격 조사)+관형사형 어미(-ㄴ, -ㄹ)'이므로 '뿐, 만큼'과 띄어 쓰면 된다. 그러나 예 5와 같은 문장들(약 1,295문장)의 정확한 띄어쓰기를 위해서는 별도의 처리 과정이 필요하다. 참고로 '뿐/만큼'을 포함한 문장 중 약 11%가 다음의 처리루틴으로 교정해야 한다.

['뿐/만큼'에 대한 '일'과 '인'의 처리루틴]

① if '일'과 '인'이 단독으로 나타나면
: '일/인'을 '뿐/만큼'과 붙여 쓴다.
② else if '일'과 '인'이 겹쳐서 출현하면
: '일/인'을 '뿐/만큼'과 띄어 쓴다.
③ else if '일'과 '인'이 단일어이면
: '일/인'을 '뿐/만큼'과 붙여 쓴다.
④ else if '일(事)/인(人)'과 복합어가 될 수 없는 명사
: '일/인'을 '뿐/만큼'과 띄어 쓴다.
⑤ else if '일'과 '인'의 선행 요소가 부호일 때
: '일/인'을 '뿐/만큼'과 띄어 쓴다.
⑥ else if 'Noun$_1$+조사(혹은 어미) Noun$_1$인(일)'일 때
: '일/인'을 '뿐/만큼'과 띄어 쓴다.
else if .....
⑦ else
: '일/인'을 '뿐/만큼'과 띄어 쓴다.

위의 처리루틴에 대한 각각의 예는 다음과 같다.

예 6 ① 그런 일만큼, 얻는 일만큼.
② 원인일 뿐 아니라,고된 일인 만큼.
③ 아랍인만큼, 잡일뿐.
④ 것일 뿐만 아니라, 경우인 만큼.

⑤ <자의적인 강제행사>일 뿐.
⑥ 처지가 처지인 만큼.
⑦ 표현일 뿐, 평가인 만큼.

'대로'는 예5에서 언급한 형태들 외에 명사 '대로(大路)'와 중의성이 발생할 수 있다.

예 7 ① 올림픽대로를 따라 내려가세요.
② 너 갈 대로 가라.

예7의 ①에서처럼 특정 지역을 나타내는 '올림픽대로, 천호대로' 등과 같이 '대로'가 명사 '대로(大路)'로 쓰일 수 있기 때문에 '대로(大路)'가 포함된 복합어를 사전에 포함시켜 처리한다. 그런데 예7의 ②에서 '갈 대로'는 동사의 어간 '가-'에 관형사형 어미 '-ㄹ'과 의존명사 '대로'로 분석될 수 있고, 명사 '갈대'와 조사 '로'의 분석도 가능하다. 이러한 경우에 구문 분석을 이용하여 교정해야 한다.

3.5. '되다'의 교정 방법

이 절에서는 띄어쓰기 오류 유형 중 '동사/접미사'의 예인 '되다'의 교정 방법을 요약하면 다음과 같다.

교정 1. 명사 + 동사 파생 접미사 '되'

'되다'는 그 자체로 동사이면서 '하다'처럼 동사 파생 접미사로도 쓰인다. 이때, '되다'가 동사일 경우 앞 명사와 띄어 쓰고, 동사 파생 접미사일 경우 앞 명사와 붙여 써야 한다.

'결정되다, 해결되다'가 '결정하다, 해결하다'처럼 '-하' 접미사가 결합될 수 있는 명사 '결정, 해결'에 '-되'가 붙은 경우 이 '-되'는 접미사이고, '-하'가 붙을 수 없는 명사 뒤에 '되다'가 결합된 경우 이 '되다'는 보어인 앞 명사의 동사가 된다. 따라서 이를 처리하기 위해 사전에 '-하'가 결합될 수 있는 명사에 대한 정보를 두어 이용한다.

예 8 ① 문제하다(X) -> 문제 되는
② 주목하다(ㅇ) -> 주목되는

예8의 ①과 ②에서처럼 '하다' 와 붙을 수 있는 단어일 때 붙여 쓰고, 그렇지 않으면 띄어 쓴다.

4. 실험 및 고찰

본 논문에서 제안한 중의 오류의 형태적 교정 방법으로 코퍼스에서 나타난 해당 오류의 문장을 분석하고, 제시된 교정 방법을 적용하여 실험하였다.

본 실험에 앞서 97년 한국과학기술원에서 만든 대한민국 국어 정보베이스의 한국어 텍스트 코퍼스 약 615만 어절을 대상으로 문장을 분석하였다.

아래 표에서 문장수는 각각의 교정방법의 대상 문장수를 말하고, 분포율과 분석률은 해당 오류의 총 문장수에 대한 각각의 대상 문장 수와 모든 문장이 교정 대상일 때 각각의 교정 방법으로 교정할 수 있는 교정률이다.

표 2 '로써/로서'의 문장 수, 분포율, 분석률

교정 방법	문장수 (문장)	분포율 (%)	분석률 (%)
[체언류] + '하-' + ㅁ(명사형 어미) + 으로써	2707	13.5	100
[체언류] + '되-' + ㅁ(명사형 어미) + 으로써	30	0.2	100
일반용언의 명사형(명사형 어미 'ㅁ') + 으로써	2216	11.1	81
사람의 의미를 갖는 명사 + (으)로서	3448	17.2	100
상용구	205	1.0	100
기 타	11417	56.9	0
합 계	20023	100	41

표 3 '-ㄴ데/-ㄴ 데'의 문장 수, 분포율, 분석률

교정 방법	문장수 (문장)	분포율 (%)	분석률 (%)
데[조사] + 있다(없다)의 활용형 또는 그 파생 단어	860	11	100
데 + 조사 혹은 어미	1060	14	100
데[조사] + 문장 부호	1940	26	100
체언류 + '인데'	360	5	100
상용구	220	3	100
기 타	3060	41	0
합 계	7500	100	59.2

표 4 '뿐'의 문장 수, 분포율, 분석률

교정 방법	문장수 (문장)	분포율 (%)	분석률 (%)
용언의 관형사형 어미 + '뿐'	6303	74.4	96
'일/인'의 처리루틴	1148	13.6	98
기 타	1021	12	98
합 계	8472	100	96.5

표 5 '만큼'의 문장 수, 분포율, 분석률

교정 방법	문장수 (문장)	분포율 (%)	분석률 (%)
용언의 관형사형 어미 + '만큼'	3076	86.2	96
'일/인'의 처리루틴	147	4	98
기 타	350	9.8	98
합 계	3573	100	96.2

실험은 문장 분석에 사용된 코퍼스 615만 어절, 동아일보 사설 20만 어절과 기타(통신 게시판의 문서) 20만 어절 등 총 655만 어절을 대상으로 하였다. 다음의 표 6은 중의 오류 유형과 교정률을 나타낸 것이다. 이것은 일반적으로 가장 많이 나타나는 오류 중의 일부에 대한 교정률이다.

표 6 선택 유형과 교정률

유 형	교정률 (%)
'로써/로서'	70.4
'-ㄴ데/-ㄴ 데'	72.9
'듯'	70.0
'뿐'	99.5
'만큼'	94.5
'되다'	100
평 균	84.6

5. 결론

본 논문에서는 철자 오류와 띄어쓰기 오류 중에서 기존의 교정 시스템에서 해결할 수 없었던 중의 오류들의 유형들을 분류하고 이들에 대한 형태적 교정 방법을 제시하였다. 이러한 오류는 일반 문서 작성자뿐만 아니라 한글 맞춤법에 대해 어느 정도 지식을 가진 사람의 경우에도 구분이 모호한 것들로서 이들에 대한 처리는 기존의 교정 시스템의 교정률 향상에 기여할 수 있다.

따라서 이러한 중의 오류의 교정에 대한 기초 연구로 기존의 교정 시스템에서 교정하고 있는 부분 이외의 철자 오류와 띄어쓰기 오류의 일부 즉, '로써/로서', '-ㄴ데/-ㄴ 데', '듯', '뿐', '만큼', '대로', '되다'에 대한 형태적 교정 방법을 제시하였다. 약 655만 어절의 코퍼스를 이용하여 실험한 결과, 84.6%의 교정률을 보였다.

특히 일반 사용자들의 통신 게시판 문서에서 다른 코퍼스보다 높은 교정률을 보였다. 이는 본 논문에서 제안하는 방법이 일반 사용자가 작성하여 전문가에 의한 교정을 거치지 않은 문장에서 효율적임을 보여준다. 그리고 이 교정률은 전체 교정 시스템에서의 교정률이 아니라 위 오류들에만 해당하는 것이므로 본 논문에서 제시한 교정 방법을 기존의 전체 교정 시스템에 추가했을 때 시스템 전체의 교정률은 높아진다.

본 논문에서 제안한 중의 오류에 대한 전체적인 교정률은 아직까지 실용적으로 쓰이기에는 미흡한 단계이다. 그러나 복잡한 구문 분석 단계 없이 교정되므로 전체 교정 시스템의 속도가 다소 저하되나 교정률을 더 높일 수 있다. 또한 기존의 교정 시스템에서 틀린 단어를 맞는 것으로 간과하게 됨으로써 신뢰도가 떨어지는 단점을 보완할 수 있다. 향후 동사의 의미 태깅과 구문 분석 및 의미 분석 등의 방법을 추가하면 교정률을 더 높일 수 있을 것이며, 이와 비슷한 유형의 다른 어휘 교정에 대한 기초 자료로 사용될 수 있다.

참 고 문 헌

[1] 심철민 외, "연어 정보에 기반한 한국어 철자 검사와 교정기의 구현", 한국정보과학회지 제 23권, 제 7호, 7월, 1996.

[2] 심철민 및 권혁철, "단어 간 지배 관계 및 연관 관계를 이용한 한국어 교열 시스템", 제5회 한글 및 한국어정보처리 학술발표논문집, 한국인지과학회, pp. 303 -316, 1993.

[3] 미승우, "새맞춤법과 교정의 실제", 어문각, 1993.

[4] 최재혁, "양방향 최장일치법을 이용한 한국어 띄어쓰기 자동 교정 시스템", 제 9 회 한글 및 한국어정보처리 학술발표논문집, pp. 145-151 1997.

[5] 이병모, "의존명사의 형태론적연구", 학문사, 1995.

[6] 최재혁, "양방향 최장일치법에 의한 한국어 형태소 분석기의 구현", 경북대학교 컴퓨터공학과 박사학위논문, 1993.
[7] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 한국정보과학회지 제 23 권, 제 9 호, 9월, 1996.
[8] 강승식외, "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기", 한국정보과학회지 제 23 권, 제 5 호, 5월, 1996.
[9] 이기문, "동아 새국어 사전", 동아출판사, 1996.
[10]임한규 및 김웅모, "철자오류의 통계자료에 근거한 철자오류 교정시스템", 한국정보처리학회지 제 2 권 제6호, 11월, 1995.
[11]이병훈외, "말뭉치를 기반으로한 한국어 철자 교정기의 구현", 한국어정보처리 학술발표논문집, pp. 285-294, 1993.
[12]정한민외, "자판 특성을 이용한 Nero -Fuzzy 한국어 철자 교정기의 구현", 한글 및 한국어정보처리 학술발표논문집, pp.317 -328, 1993.