한국어 어절의 철자변화 현상 분류와 인식 방법

(A Method to Classify and Recognize Spelling Changes between Morphemes of a Korean Word)

김 덕 봉 †

(Deok-Bong Kim)

요 약 현재의 한국어 형태소 태그 부착 말뭉치에는 형태소 결합 경계의 철자변화 정보가 명시적으로 표시되어 있지 않다. 이로 인해 태그 부착 말뭉치로부터 형태소 분석에 필요한 사전을 자동으로 구축하거나 형태소 결합 경계의 철자변화 현상에 대한 체계적 예제 수집 등과 같은 한국어 형태론 연구에 필요한 자료 획득이 어렵다. 이 문제를 해결하기 위하여 본 논문은 사전과 음운 규칙을 이용하지 않고, 태그 부착 말뭉치의 어절 문자열과 형태소 문자열만을 비교하여 어절을 구성하는 형태소의 철자변화 현상을 인식하는 간단한 방법을 제안한다. 이 방법은 규칙을 사용하지 않기 때문에 두 형태소 결합으로 나타나는 모든 철자변화 현상을 유연하게 인식할 수 있고, 알고리즘 구현만으로 문제를 해결할 수 있기 때문에 비용이 싸다는 특징이 있다. 한 태그 부착 말뭉치에 대한 실험에서 본 방법은 실험 말뭉치 어절에 나타나는 철자변화를 100% 인식하는 것으로 나타났다.

키워드 : 철자변화현상, 형태음운규칙, 형태소분석, 단어인식

Abstract There is no explicit spelling change information in part-of-speech tagged corpora of Korean. It causes some difficulties in acquiring the data to study Korean morphology, i.e. automatically in constructing a dictionary for morphological analysis and systematically in collecting the phenomena of the spelling changes from the corpora. To solve this problem, this paper presents a method to recognize spelling changes between morphemes of a Korean word in tagged corpora, only using a string matching, without using a dictionary and phonological rules. This method not only has an ability to robustly recognize the spelling changes because it doesn't use any phonological rules, but also can be implemented with few cost. This method has been experimented with a large tagged corpus of Korean, and recognized the 100% of spelling changes in the corpus with accuracy.

Key words : Spelling Changes, Morphophonological Rules, Morphological Analysis

1. 서 론

지난 수년간 대량의 한국어 말뭉치가 여러 형태로 구축되어 자연언어처리 연구에 사용되고 있다. 특히 어절의 구성 형태소를 분리하여 품사 태그를 부착한 말뭉치(이하, 형태소 태그 부착 말뭉치)가 형태소 태그 부착기에 필요한 정보원으로 인식되어 많은 자연언어처리 연구자들로부터 큰 관심을 받고 있다. 이를 반영하여 국가에서도 지난 수년간 연구비를 지원하여 수천만 어절의 형태소 태그 부착 말뭉치를 구축하였다. 가장 대표적인 국가 말뭉치로는 국내 최초로 구축한 1,500만 어절의 KAIST 형태소 태그 부착 말뭉치[1]와 최근 국어학자 중심으로 새로 구축하고 있는 555만 어절의 세종 형태소 태그 부착 말뭉치[2]가 있다. 지금까지 이러한 한국어 형태소 태그 부착 말뭉치는 주어진 문장을 구성하는 각 형태소에 가장 적합한 태그를 결정하는 형태소 태그 부착기의 성능 향상을 위한 용도로 주로 이용되어 왔다[3].

그러나, 현재의 한국어 형태소 태그 부착 말뭉치는 기본적으로 문장을 구성하는 어절과 그 어절의 구성 형태소와 태그 정보만을 제공하기 때문에 그러한 정보를 이

† 종신회원 : 성공회대학교 컴퓨터정보공학부 교수
dbkim@mail.skhu.ac.kr
논문접수 : 2002년 9월 9일
심사완료 : 2003년 2월 18일

용한 자연언어처리의 성능 향상 쪽에만 활용되는 한계를 보이고 있다. 예를 들어, 가장 많이 활용하고 있는 형태소 태그 부착기에서는 주어진 문맥에서의 형태소 태그 확률을 구하는 데 사용하고 있다. 형태소 분석기에서는 형태소 분석 과정에서 발생하는 오분석 후보를 제거하기 위한 포함관계(subsumption relation) 지식 획득[4]과 형태소 분석의 고속화를 위한 기분석 어절 사전 구축[5]에 활용하고 있는 것으로 알려져 있다. 하지만, 기본적으로 태그 부착 말뭉치의 어절과 구성 형태소와 태그 정보만으로는 형태소 분석에 필요한 사전을 쉽게 구축할 수 없는 문제를 가지고 있다. 이는 어절을 구성하는 형태소의 결합 경계에서 나타나는 철자변화 정보를 명시적으로 표시하지 않아 생기는 문제라 할 수 있다. 즉, 현재의 한국어 형태소 태그 부착 말뭉치에서 형태소와 태그 분리와 같은 단순 처리만으로는 한국어 단어(어절)를 인식할 때 필요한 철자변화 복원 정보를 갖는 형태소 분석 사전을 구축할 수 없는 문제가 있다.

따라서 현재의 한국어 형태소 태그 부착 말뭉치의 한계를 극복하고 활용을 넓히기 위해서는 형태소 결합 경계의 철자변화에 대한 정보를 현재의 말뭉치에 추가하거나 현재의 말뭉치에서 자동으로 획득하는 방법이 필요하다. 먼저, 철자변화 정보를 현재의 말뭉치에 추가하는 것은 단순 처리로 원하는 결과를 쉽게 얻을 수 있는 잇점이 있지만, 형태소의 철자변화 복원 정보를 현재의 말뭉치에 추가하는 비용이 상당히 클 것으로 예상될 뿐 아니라, 그러한 정보의 표현 자체도 복잡하거나 어려워 이것을 효과적으로 표현하기 위한 연구가 선행되어야 하는 문제가 있다. 예를 들어, 현재의 말뭉치에서 어절의 표층형 '도왔다'는 어휘형으로 '돕/동사+었/선어말어미+다/어말어미'로 표현되고 있는데, 여기에는 사실 'ㅂ'-불규칙 어간 '돕-'이 '-어'로 시작하는 어미와 결합할 때 모음조화로 어미의 첫 모음 '어'가 어간의 끝 모음 특성(양성모음)에 따라 '아'로 교체되고, 어간의 끝음절 받침 'ㅂ'과 결합하여 '와'로 바뀌어 표층형으로 나타난다는 정보가 숨어 있다. 또 어절의 표층형 '됐다'는 어휘형으로 '되/동사+었/선어말어미+다/어말어미'인데, 여기에는 어간의 끝 모음 'ㅚ'가 어미의 첫 모음 '어'와 결합하여 'ㅙ'로 축약되고, 어간의 초성음 'ㄷ'과 어미의 종성음 'ㅆ'과 함께 '됐'으로 합성됐다는 정보가 감추어져 있다. 결과적으로 이렇게 숨어 있는 정보를 명시적으로 표현하는 작업은 철자변화 결합 형태소들의 어떤 문자(음)들이 결합하여 표층 어절의 어떤 문자(음)들로 변했는지를 나타낼 수 있어야 하고, 또 어절의 표층형과 어휘형에서의 철자변화 구간을 표시할 수 있어야 한다.

반면에, 철자변화 정보를 현재의 말뭉치에서 자동으로 획득하는 방법은 추가 비용 없이 현재의 말뭉치를 그대로 이용할 수 있다는 잇점이 있다. 그러나, 문제는 이것을 어떻게 실현하느냐에 달려 있다. 여기에는 두 가지 극단적인 방법이 있을 수 있다. 그 중 하나는 기존의 형태소 분석/생성 시스템을 이용하는 방법이다. 이 방법은 미리 생성한 사전과 음운 규칙을 기본적으로 사용하기 때문에 규칙적인 철자변화에 대해서는 비교적 안정적인 인식률을 확보할 수 있을 것으로 기대되나, 시스템 사전에 없거나 음운 규칙으로 설명되지 않는 철자변화 현상에 대해서는 유연한 인식을 기대할 수 없고, 또한 구축 비용도 비싼 편이라 이러한 시스템이 일반에 공개되지 않는 한 이를 이용하는 연구자는 매우 제한적일 수밖에 없다는 문제가 있다. 또 다른 하나의 방법은 주어진 형태소 태그 부착 말뭉치의 어절 문자열과 형태소 문자열만을 비교하여 어절을 구성하는 형태소의 철자변화 현상을 인식하는 것이다. 이 방법은 어떤 철자변화에 대해서도 유연하게 인식할 수 있고 구축 비용도 거의 들지 않아 가장 이상적인 해결책으로 보이나, 안정적인 인식률 확보가 전제되어야 할 것으로 보인다.

본 논문은 사전과 음운 규칙을 이용하지 않고, 한국어 형태소 태그 부착 말뭉치에 있는 어절 문자열과 형태소 문자열만을 비교하여 어절을 구성하는 형태소의 철자변화 현상을 안정적으로 인식하는 방법을 제안한다. 본 논문에서 제안 방법은 다음과 같은 순서로 기술한다. 2장에서는 한글맞춤법과 현대 국어 문법에 바탕을 두어 한국어 어절 형성에서 나타나는 형태소 결합 경계의 철자변화 현상을 분류하고, 3장에서는 한국어 형태소 태그 부착 말뭉치에서 어절의 철자변화 현상을 자동 인식하는 방법을 기술한다. 4장에서는 한국어 형태소 태그 부착 말뭉치를 가지고 어절의 철자변화 현상 인식에 대한 실험과 분석을 수행하여 제안 방법의 안정성을 보인다. 5장은 논문의 결론으로, 제안 방법의 특징과 연구 결과에 대한 기대효과에 대하여 기술한다.

2. 어절의 철자변화 현상 분류 방법

교착어인 한국어는 두 형태소가 결합할 때 철자변화 규칙이 적용되어 두 형태소의 형태음운론적 환경에 따라 형태소 결합 경계의 철자가 변한다. 이러한 철자 변화는 특히 한국어 어절 인식을 어렵게 하는 원인이 되어 왔다. 철자변화 어절의 경우 단순 사전 탐색만으로는 구성 형태소의 분리가 어려워 형태소 분석을 할 수 없기 때문에 이를 극복하기 위한 방법으로 테이블 파싱법

[6], 최장일치법[7], 음절 기반 모델[8], 어절 사전 기반 모델[5], 두단계 모델[9], 예측 기반 모델[10] 등이 연구되어 왔지만, 아직도 문제의 대상이 되고 있다. 또한, 어절 인식 과정에서 철자변화 복원 규칙이 무리하게 적용되어 과분석 문제가 일어나 자연언어처리의 정확율을 떨어뜨리는 원인이 되기도 한다. 그렇다고 해서, 형태소 결합 경계의 철자변화를 피하기 위하여 기본형 이외에 철자변화된 다른 이형태들을 사전에 두는 경우에는 사전 크기 뿐만 아니라 형태소 결합 검사의 부담이 늘어나거나, 과분석 문제가 새로 나타날 수 있는 문제를 제공하고 있다. 더구나, 철자변화 처리 방법에 따라 사전 어휘에 대한 등록 기준이 제각각 달라 관련 연구를 생산적으로 지원하기 위한 자원 공유도 쉽지 않게 한다. 따라서 효율적인 한국어 정보처리를 연구하기 위해서는 우선 형태소 결합 경계의 철자변화 현상에 대한 조사와 체계적 분류가 요구된다. 철자변화 현상의 체계적 분류는 형태소 분석과 단어 형성 연구에 필요한 말뭉치의 실제 예제들을 현상별로 자동으로 모아, 추후 철자변화 현상을 규칙화하거나 효율적으로 다루게 하는 데 유용하게 쓰일 수 있다.

한글 맞춤법[11]과 현대 국어 문법[12]에 기술된 한국어 어절 형성에서 사용되는 철자변화 규칙에는 용언에 적용되는 규칙 활용(한글 맞춤법 제16항의 모음조화와 제32, 34~40항의 축약, 지정사 '이' 탈락, 매개모음 '으' 첨가)과 불규칙 활용(한글 맞춤법 제18항의 거라/너라/ㄷ/러/르/ㅂ/ㅅ/여/우/ㅎ-불규칙), 체언에 주로 적용되는 첨가(한글 맞춤법 제30항의 사이시옷 첨가와 매개모음 '이' 첨가), 축약(한글 맞춤법 제33항의 체언+조사 축약) 규칙 등이 있다. 이러한 전통적인 한국어의 철자변화 현상들은 형태소가 결합할 때 두 형태소 중 어느 쪽의 문자(음)가 바뀌느냐에 따라 무변화(no alternation), 좌변화(left alternation), 우변화(right alternation), 좌우변화(left and right alternation) 등 4가지 유형의 형태소 결합 형태로 분류할 수 있다[10]. 용언의 경우 좌변화는 어간이 바뀐 현상이고, 우변화는 어미가, 좌우변화는 어간과 어미가 함께 바뀐 현상이라 할 수 있다. 또한, 형태소의 철자변화가 어절의 표층형에 나타난 상태에 따라 음운첨가, 음운탈락, 음운대체로 분류할 수 있다. 다음은 이러한 분류를 이론적으로 체계화하고, 형태소 결합으로 생기는 철자변화 현상의 예들을 제안 방법에 따라 분류해 보인다.

정의 1: 한글 자모 문자집합 Σ는 $\Sigma_{초성음} + \Sigma_{중성음} + \Sigma_{종성음}$이다. 참조로 나중에 어절의 철자변화 현상 인식 알고리즘 이해와 구현을 돕기 위하여 아래에서는 시스템 효율과 가독성을 고려하여 초성과 종성을 구분한 시스템 내부 문자도 괄호 속에 표기하였다.

$\Sigma_{초성음}$ = {ㄱ(g), ㄲ(gg), ㄴ(n), ㄷ(d), ㄸ(dd), ㄹ(l), ㅁ(m), ㅂ(b), ㅃ(bb), ㅅ(s), ㅆ(ss), ㅈ(j), ㅉ(jj), ㅊ(c), ㅋ(k), ㅌ(t), ㅍ(p), ㅎ(h)}

$\Sigma_{중성음}$ = {ㅏ(a), ㅑ(ya), ㅓ(A), ㅕ(yA), ㅗ(o), ㅛ(yo), ㅜ(u), ㅠ(yu), ㅡ(U), ㅣ(i), ㅐ(E), ㅒ(yE), ㅔ(e), ㅖ(ye), ㅘ(wa), ㅙ(wE), ㅚ(wi), ㅝ(wA), ㅞ(we), ㅟ(yi), ㅢ(yU)}

$\Sigma_{종성음}$ = {ㄱ(G), ㄲ(GG), ㄳ(GS), ㄴ(N), ㄵ(NJ), ㄶ(NH), ㄷ(D), ㄹ(L), ㄺ(LG), ㄻ(LM), ㄼ(LB), ㄽ(LS), ㄾ(LT), ㄿ(LP), ㅀ(LH), ㅁ(M), ㅂ(B), ㅄ(BS), ㅅ(S), ㅆ(SS), ㅇ(Q), ㅈ(J), ㅊ(C), ㅋ(K), ㅌ(T), ㅍ(P), ㅎ(H)}

정의 2: 두 형태소가 결합할 때 표층형 어절에 철자변화 없이 그대로 나타난 형태소의 부분문자열을 형태소의 특성문자열(MCS)이라 하고, 왼쪽 형태소의 특성문자열을 MCS_l, 오른쪽 형태소의 특성문자열을 MCS_r이라고 하자. 또 AS를 두 형태소의 결합으로 표층형 어절에 나타난 철자변화 문자열이라 하고, LAS와 RAS를 각각 표층형 어절에 나타나지 않은 왼쪽 형태소와 오른쪽 형태소의 변화문자열이라고 하자. 그러면 MCS_l, MCS_r, LAS, RAS, $AS \in \Sigma^*$라 할 때, 두 형태소 결합의 **어휘형**은 $MCS_l \cdot LAS + RAS \cdot MCS_r$이고, 그것의 **표층형**은 $MCS_l \cdot AS \cdot MCS_r$이다.

정의 3: 철자변화 현상은 표층형과 어휘형의 변화부분 대응 '$AS : LAS + RAS$'으로 인식된다. 여기에서 ':'는 표층형과 어휘형의 대응 구분자이고, '+'는 왼쪽 형태소와 오른쪽 형태소의 구분자이다.

1. $LAS = \lambda$, $RAS = \lambda$, $AS = \lambda$이면 무변화 결합이다. 즉, 두 형태소가 결합할 때 어떤 철자변화도 일어나지 않고 두 형태소의 어휘형이 모두 표층형에 연접하여 나타나면 **무변화** 결합이다.
2. $LAS \in \Sigma^*$, $RAS = \lambda$, $AS \in \Sigma^*$이거나 $LAS = \lambda$, $RAS = \lambda$, $AS \in \Sigma_{종성음}$이면 좌변화 결합이다. 즉, 두 형태소가 결합할 때 왼쪽(첫번째) 형태소의 경계만이 철자변화가 일어나 표층형에 나타나거나 종성음이 삽입되면 **좌변화** 결합이다.
3. $LAS = \lambda$, $RAS \in \Sigma^*$, $AS \in \Sigma^*$이거나 $LAS = \lambda$, $RAS = \lambda$, $AS \in \{\Sigma_{초성음}, \Sigma_{중성음}\}^+$이면 우변화 결합이다. 즉, 두 형태소가 결합할 때 오른쪽(두번째) 형태소의 경계만이 철자변화가 일어나 표층형에 나타나거나 초성음이나 중성음이 삽입되면 **우변화** 결합이다.

4. $LAS \in \Sigma^+$, $RAS \in \Sigma^+$, $AS \in \Sigma^*$이면 좌우변화 결합이다. 즉, 두 형태소 결합할 때 두 경계 모두에서 철자변화가 일어나면 **좌우변화** 결합이다.
5. $LAS = \lambda$, $RAS = \lambda$, $AS \in \Sigma^+$이면 **음운첨가형** 변화다.
6. $LAS \in \Sigma^+$, $RAS = \lambda$, $AS = \lambda$이거나 $LAS = \lambda$, $RAS \in \Sigma^+$, $AS = \lambda$이면 **음운탈락형** 변화다.
7. $LAS \in \Sigma^+$, $RAS \in \Sigma^*$, $AS \in \Sigma^+$이거나 $LAS \in \Sigma^*$, $RAS \in \Sigma^+$, $AS \in \Sigma^+$이면 **음운대체형** 변화다

이제 정의 3에서의 철자변화 현상 인식법에 따라 예제에 나타난 철자변화 현상들을 체계적으로 분류할 수 있다. 이를 확인하기 위해 예를 들어 살펴 보자.

(1) 부어 붓/동사+어/어말어미

용언 '부어(buA)'는 동사어간 '붓(buS)'과 어말어미 '어(A)'가 결합된 어절로 다음과 같은 대응을 이룬다.

표층형: **b**(ㅂ) **u**(ㅜ) **A**(어)
어휘형: **b**(ㅂ) **u**(ㅜ) **S**(ㅅ) + **A**(어)

두 형태소의 결합으로 나타난 왼쪽 형태소의 특성 문자열 MCS_l은 'bu(부)'이고 철자변화 문자열 LAS는 'S(ㅅ)'이다. 그리고 오른쪽 형태소의 특성 문자열 MCS_r은 형태소 문자열 'A(어)'와 같기 때문에 철자변화 문자열 RAS는 공백(λ)이다. 또 표층에 나타난 철자변화 문자열 AS도 공백이기 때문에 예제 (1)은 'λ : S(ㅅ) + λ'($AS = \lambda$, LAS=S(ㅅ), $RAS = \lambda$)의 철자변화 특성을 갖고 있어, 결과적으로 두 형태소 결합에서 왼쪽 형태소의 끝문자 'ㅅ(S)'이 탈락된 좌변화와 음운탈락형 변화 현상의 예로 분류할 수 있다.

(2) 이르러 이르/동사+어/어말어미

또 용언 '이르러(ilUlA)'는 동사어간 '이르(ilU)'와 어말어미 '어(A)'가 결합된 어절로 다음과 같은 대응을 이룬다.

표층형: i(이) l(ㄹ) **U**(ㅡ) l(ㄹ) **A**(ㅓ)
어휘형: i(이) l(ㄹ) **U**(ㅡ) + **A**(어)

위와 같은 대응 결과로 예제 (2)는 'l(ㄹ) : λ + λ'(AS=l(ㄹ), $LAS = \lambda$, $RAS = \lambda$)의 철자변화 특성을 갖고 있는 것으로 인식되어, 두 형태소 결합으로 오른쪽 형태소 앞에 초성음 'ㄹ(l)'이 첨가된 우변화와 음운첨가형 변화 현상의 예로 분류할 수 있다.

(3) 도와 돕/동사+어/어말어미

또 다른 예로 용언 '도와(dowa)'는 동사어간 '돕(doB)'과 어말어미 '어(A)'가 결합된 어절로 다음과 같은 대응을 이룬다.

표층형: **d**(ㄷ) **o**(ㅗ) **w**(오) **a**(ㅏ)
어휘형: **d**(ㄷ) **o**(ㅗ) **B**(ㅂ) + **A**(어)

대응 결과로 볼 때 예제 (3)은 'wa(와) : B(ㅂ) + A(어)'(AS=wa(와), LAS=B(ㅂ), RAS=A(어))의 철자변화 특성을 보여, 두 형태소 결합으로 왼쪽 형태소의 끝문자 'ㅂ(B)'과 오른쪽 형태소의 첫문자 '어(A)'가 '와(wa)'로 대체된 좌우변화와 음운대체형 변화 현상의 예로 분류할 수 있다.

(4) 그건 그것/대명사+은/조사

마지막 예로 어절 '그건(gUgAN)'은 대명사 '그것(gUgAS)'과 조사 '은(UN)'이 결합된 어절로 다음과 같은 대응을 이룬다.

표층형: **g**(ㄱ) **U**(ㅡ) **g**(ㄱ) **A**(ㅓ) **N**(ㄴ)
어휘형: **g**(ㄱ) **U**(ㅡ) **g**(ㄱ) **A**(ㅓ) **S**(ㅅ) + **U**(ㅡ) **N**(ㄴ)

대응 결과 예제 (4)는 'λ : S(ㅅ) + U(ㅡ)'($AS = \lambda$, LAS=S(ㅅ), RAS=U(ㅡ))의 철자변화 특성을 갖고 있어, 두 형태소 결합으로 왼쪽 형태소의 끝문자 'ㅅ(S)'과 오른쪽 형태소의 첫문자 '으(U)'가 탈락된 좌우변화와 음운탈락형 변화 현상의 예로 분류할 수 있다.

3. 어절의 철자변화 현상 인식 알고리즘

앞에서 논의한 철자변화 현상은 하나 이상의 형태소로 이루어진 어절과 그 어절의 형태소 태그 부착 자료로 이루어진 말뭉치에서 자동으로 인식하여 분류할 수 있다. 먼저 본 알고리즘을 기술하기에 앞서 어절의 철자변화 현상 인식 문제와 알고리즘 이해에 필요한 몇 가지 용어와 개념들을 정의하겠다.

정의 4: 주어진 한국어 형태소 태그 부착 말뭉치에서 k개의 형태소로 구성된 어절은 $k-1$번의 형태소 결합을 갖기 때문에 이 어절의 어휘형 WPL과 표층형 WPS는 정의 2에서 기술한 두 형태소 결합의 어휘형과 표층형의 반복으로 표현되고, 어휘형 형태소마다 부착되어 있는 형태소 태그 정보 TAG는 설명의 편의를 위해 형태소 결합과 같은 순서로 된 형태소 태그 결합으로 따로 표현된다.

$$WPL = MORPH_1 + MORPH_2 + MORPH_3 + \cdots + MORPHK$$
$$= MCS_1 \cdot LAS_1 + RAS_2 \cdot MCS_2 \cdot LAS_2 + RAS_3 \cdot MCS_3 \cdot LAS_3 + \cdots + RAS_k \cdot MCS_k$$
$$TAG = TAG_1 + TAG_2 + TAG_3 + \cdots + TAG_k$$
$$WPS = MCS_1 \cdot AS_2 \cdot MCS_2 \cdot AS_3 \cdot MCS_3 \cdot \cdots \cdot AS_k MCS_k$$

여기에서 MCS_i, LAS_i, RAS_i, $AS_i \in \Sigma^*$이고, $RAS_1 = \lambda$, $LAS_k = \lambda$라 할 때 어절을 구성하는 한 형태소의

어휘형 $MORPH_i$는 $RAS_i \cdot MCS_i \cdot LAS_i$이다. 그리고 알고리즘 설명의 편의를 위해 AS의 첨자는 2부터 시작한다.

결과적으로 주어진 한국어 형태소 태그 부착 말뭉치에서 어절의 철자변화 현상 인식 문제는 어절의 형태소 태그 정보 TAG를 참조하면서 원형 형태소 열로 구성된 어휘형 WPL의 각 형태소 $MORPH_i$와 표층형 어절 WPS의 문자열 비교를 통해 형태소 특성문자열 MCS_i를 찾고 두 형태소 결합 경계에서 일어난 철자변화 정보, 즉 LAS_i, RAS_{i+1}, AS_{i+1}를 얻는 것이라 할 수 있다. 일반적으로 주어진 텍스트 문자열에서 패턴(문자열)을 찾는 단순 문자열 비교는 간단한 문제로 볼 수 있지만, 어절의 철자변화 현상 인식을 위한 문자열 비교 문제는 그렇게 단순하지 않다. 이 문제의 어려움은 형태소 결합에 의한 철자변화로 주어진 형태소 문자열(패턴)이 표층형 어절(텍스트)에 부분만 나타나거나 전혀 나타나지 않을 수 있고, 형태소 경계(위치정보)가 표층형 어절에 명시적으로 표시되어 있지 않을 뿐 아니라 한 형태소가 다른 형태소의 부분문자열일 수 있기 때문에 발생한다. 이로 인해 단순 문자열 비교에만 의존할 경우 어절의 철자변화 현상 인식에 많은 오류가 생길 수 있다. 이러한 오류를 방지하기 위하여 본 알고리즘에서는 표층형 태소공간(정의 5), 부분문자열 매칭 전략(정의 6), 착시 매칭오류 조건(정의 7), 근접매칭전처리(정의 8), 완전대체문자열(정의 9), 단순결합경계(정의 10) 등과 같은 개념들을 사용하였다.

정의 5: 표층형태소공간 SMS_i는 한 형태소 $MORPH_i$가 어절의 표층형 WPS에서 인식될 수 있는 WPS의 부분문자열로, WPS의 현재 탐색 문자 위치부터 시작하여 마지막 문자 위치까지의 범위 내에서 $|AS_i| + |MORPH_i|$의 크기를 갖는다. $|AS_i|$는 $MORPH_i$가 인식되어야 알 수 있기 때문에 본 연구에서는 AS_i의 의미가 두 형태소 $MORPH_{i-1}$과 $MORPH_i$의 결합에 의해 생긴 표층형의 철자변화 문자열이고, 많은 예제에서 철자변화로 한 형태소 앞에서 첨가되거나 앞 형태소가 변화하여 인식되지 않은 문자가 4자 이상 나타나지 않는다는 점을 감안하여 $|AS_i|$를 4로 가정한다. 이는 $MORPH_i$가 WPS의 현재 탐색 위치에서 $|MORPH_i|$+4개의 표층형태소공간 문자열 안에서 인식될 수 있다는 것을 의미한다.

정의 6: 한 형태소의 철자변화 정보 RAS_i, MCS_i, LAS_i는 $MORPH_i$와 표층형태소공간 SMS_i와의 **부분문자열 매칭(Substring Matching)**에 의해 인식된다. 부분문자열 매칭의 모호성을 줄이기 위하여 본 논문에서는 표층형태소공간의 첫문자부터 끝문자 방향으로 비교하는 **좌우매칭**과, 부분문자열 중 가장 긴 것을 우선하는 **최장일치 우선**, 같은 부분문자열이 표층형태소공간에 1번 이상 나타날 때 먼저 나타난 것을 우선하는 **근접매칭** 전략을 사용한 부분문자열 매칭을 기본으로 사용하여 매칭 조건에 따라 다음과 같은 결과를 얻는다.

1. $MORPH_i$의 문자열이 SMS_i에 그대로 나타나면 MCS_i=$MORPH_i$, RAS_i= λ, LAS_i= λ이다.
2. $MORPH_i$의 부분문자열이 SMS_i에 나타나면 MCS_i는 SMS_i에 나타난 $MORPH_i$의 부분문자열이고, RAS_i는 MCS_i 앞에 있는 $MORPH_i$의 부분문자열, LAS_i는 MCS_i 다음에 있는 $MORPH_i$의 부분문자열이다.
3. SMS_i에서 $MORPH_i$ 문자들을 전혀 찾을 수 없다면 MCS_i= λ이다. MCS_i= λ이면 $MORPH_i$가 표층형에서 **완전 탈락**되거나 다른 문자열로 **대체**되어 표층형으로 나타난 것이다.

- $MORPH_i$가 어간인 경우 이 현상은 다음 형태소인 $MORPH_{i+1}$과의 결합에 의한 철자변화로 해석되어 RAS_i= λ, LAS_i=$MORPH_i$로 인식한다.
- $MORPH_i$가 어미인 경우에는 앞 형태소인 $MORPH_{i-1}$과의 결합에 의한 철자변화로 RAS_i=$MORPH_i$, LAS_i= λ로 인식한다.

정의 7: 표층형태소공간(SMS_i)과의 부분문자열 매칭을 통해 한 형태소($MORPH_i$)의 철자변화 정보를 인식할 때 다음의 조건을 하나라도 만족하면 **착시매칭오류**이다.

1. **완전착시조건**: $MORPH_i$가 $MORPH_{i+1}$의 부분문자열이고 $MORPH_{i+1}$의 첫문자부터 문제의 부분문자열을 포함한 문자열이 SMS_i에 나타난다. 예를 들어, (5a)에서 두번째 형태소 어말어미 '어(A)'는 두번째 표층형태소공간 'ㅐㅂㅓㄹㅣ(EbAli)'에 나타나고, 또 세번째 형태소인 보조동사 '버리(bAli)'의 부분문자열이면서 세번째 형태소에서 문제가 되는 부분문자열 'ㅂㅓ(bA)'가 역시 두번째 표층형태소공간 'ㅐㅂㅓㄹㅣ(EbAli)'에 나타난다. 따라서, 이 경우에 두번째 형태소가 표층에 나타난 것으로 인식하면 완전착시조건을 만족하는 착시매칭오류이다.
2. **부분착시조건**: SMS_i에서 부분 인식된 $MORPH_i$의 부분문자열과 완전 인식된 $MORPH_{i+1}$의 문자열이 겹치는 부분이 있다. 예를 들어, (5b)는 첫번째 형태소 대명사 '누구(nugu)'가 첫번째 표층형태소공간 'ㄴㅜㄱㅏ(nuga)'에서 부분적으로 나타나 'ㄴㅜㄱ(nug)'이 매칭되고, 또한 거기에서 두번째

형태소 조사 '가(ga)'가 완전 매칭되어 첫번째 매칭 문자열과 겹치는 문자 'ㄱ(g)'가 나타난다. 이 경우에 첫번째 형태소의 특성문자열을 'ㄴㅜㄱ(nug)'로 인식하면 부분착시조건을 만족하는 착시매칭오류이다.

3. **교차착시조건**: SMS_i에서 $MORPH_i$가 발견되고, 또한 $MORPH_i$ 대응 문자열 앞에 있는 SMS_i 부분 문자열에 $MORPH_{i+1}$가 있다. 예를 들어, (5c)는 두번째 형태소 '하(ha)'가 두번째 표층형태소공간 'ㅋㅐㅎㅏㄷㅏ(kehada)'에서 발견되고, 또한 첫번째 형태소 대응 문자열 앞에 있는 표층형태소공간의 부분문자열 'ㅋㅐ(ke)'에 세번째 형태소가 변형되어 나타난다. 이 경우에 두번째 형태소가 표층에 그대로 나타난 것으로 인식하면 교차착시조건을 만족하는 착시매칭오류이다.
4. **근접착시조건**: $MORPH_i$의 전체 혹은 부분 문자열이 SMS_i에 1번 이상 나타나고, 맨 앞에 매칭된 것이 아니라 그 뒤에 나타나는 것이 실제 $MORPH_i$의 구성문자열이다. 예를 들어, (5d)는 두번째 형태소 '어라(Ala)'의 부분 문자열 'ㄹㅏ(la)'는 두번째 표층형태소공간 'ㄹㄹㅏㄹㅏ(Llala)'에 두 번 나타나고, 첫번째 매칭된 것이 아니라 나중에 나타나는 것이 실제 구성 문자열이다. 따라서, 이 경우에 맨처음 매칭된 문자열을 첫번째 형태소로 인식한다면 근접착시조건을 만족하는 착시매칭오류이다.

(5) a. 해버려 하/동사+어/어말어미+버리/보조동사+어/어말어미
b. 누가 누구/대명사+가/조사
c. 가능케하다 가능/명사+하/동사파생접미사+게/어말어미+하/보조용언+다/어말어미
d. 올라라 오르/동사+어라/어말어미

정의 8: 앞 형태소의 철자변화 정보 LAS_{i-1}와 거기에 대응하는 표층의 문자열이 항상 명확하여 현재 형태소 $MORPH_i$의 표층형태소공간 SMS_i를 조정할 때 이를 **근접매칭전처리**라 한다. 이는 특히 착시매칭오류 중에서 근접매칭 전략 때문에 생기는 근접착시조건을 만족하는 착시매칭오류를 방지하기 위하여 사용한다. 예를 들어, '르'-불규칙 활용의 경우 어간 형태소의 철자변화 부분 'ㄹㅡ(lU)'와 그것에 대응하는 표층 문자열 'ㄹㄹ(Ll)'가 항상 명확하여 이를 토대로 표층형태소공간을 조정하여 착시매칭오류를 방지할 수 있다. 예제 (5d)에서 첫번째 형태소인 동사어간 '오르(olU)'를 매칭한 후 LAS1는 'ㄹㅡ(lU)'이고, 두번째 형태소 인식을 위한 표층형태소공간 SMS_2(ㄹㄹㅏㄹㅏ(Llala))의 맨앞 부분에서 'ㄹㄹ(Ll)'를 찾을 수 있기 때문에, 'ㄹㄹ(Ll)'를 제외한 나머지 문자열 'ala'를 SMS_2로 조정할 수 있다. 이 경우에 두번째 형태소는 조정된 SMS_2에서 인식된다. 여기에서 원래 표층형태소공간에서 제외한 표층의 문자열 'ㄹㄹ(Ll)'을 **예비철자변화문자열** PAS라 한다. 새로 조정된 표층형태소공간에서 인식한 표층의 철자변화 문자열을 AS라 할 때, 첫번째 형태소와 두번째 형태소의 결합으로 생기는 표층의 철자변화 문자열 AS_2는 PAS와 AS의 결합 문자열로 한다.

정의 9: 완전대체문자열은 한 형태소가 완전 대체되어 표층형으로 나타난 문자열이다. 예를 들어, (6a)에서 '아'는 어말어미 '어'가 양성모음 동사어간과의 결합으로 나타난 '어'의 완전대체문자열이다. 이 이외에 자주 나타나는 형태소의 완전대체문자열은 (6b)와 (6c)에서와 같이 지정사 '이'가 어미의 첫모음 '어'와의 결합으로 대체하여 나타난 'y'와, 동사 어간 '오'가 어미의 첫모음 '어'와의 결합으로 나타난 'w'가 있다.

(6) a. 담아 담/동사+어/어말어미
b. 사고여서 사고/명사+이/지정사+어서/어말어미
c. 와 오/동사+어/어말어미

정의 10: 임의의 형태소 $MORPH_i$가 어미(조사)이고 그 다음 형태소 $MORPH_{i+1}$가 어간(단어)이면 두 형태소는 **단순결합경계**를 이룬다. 단순결합경계를 이루는 두 형태소의 결합으로 인한 철자변화는 없다. 즉 $LAS_i = \lambda$, $RAS_{i+1} = \lambda$, $AS_{i+1} = \lambda$이다. 표층형에서 단순결합경계는 형태소 특성문자열(MCS)이나 완전대체문자열에 의해 인식된다. 즉, 두 형태소 $MORPH_i$와 $MORPH_{i+1}$이 단순결합경계를 이루는 경우 **표층단순결합경계점**은 다음의 위치이다.

1. $MCS_i \in \Sigma^+$이라면 MCS_i 직후의 위치.
2. $MCS_i = \lambda$이라면 $MORPH_i$의 완전대체문자열 직후 위치
3. $MCS_{i+1} \in \Sigma^+$이라면 MCS_{i+1} 직전의 위치.
4. $MCS_{i+1} = \lambda$이라면 $MORPH_{i+1}$의 완전대체문자열 직전 위치.

(7) 도와줘 돕/동사+어/어말어미+주/보조동사+어/어말어미

어절 '도와줘(dowajwA)'는 동사 '돕(doB)', 어말어미 '어(A)', 보조용언 '주(ju)', 어말어미 '어(A)'가 결합된 어절로 두번째 형태소인 어미와 세번째 형태소인 어간 사이에 단순결합경계를 갖고 있다. 표층형에서 이 단순결합경계를 찾는 방법을 보기 위하여 다음과 같은 대응을 살펴보자.

표층형: **d**(ㄷ) **o**(ㅗ) **w**(오) **a**(ㅏ) j(ㅈ) w(ㅜ) A(ㅓ)
| | | | |
어휘형: d(ㄷ) o(ㅗ) B(ㅂ) + A(어) + j(ㅈ) u(ㅜ) + A(어)

위의 대응 결과로 보면 첫번째 형태소의 특성문자열 MCS_1는 'do(도)'이고, 두번째 형태소 'A(어)'는 완전 대체되어 'a(ㅏ)'로 나타나 MCS_2=이고, 세번째 형태소의 특성문자열 MCS_3는 'j(ㅈ)'이고, 마지막 형태소 'A(어)'는 변화 없이 표층에 그대로 나타나 있다. 결과적으로 이 어절의 표층단순결합경계점은 두번째 형태소의 완전대체문자열 'a(ㅏ)'의 직후 위치이거나 MCS_3 직전 위치로 명확히 인식될 수 있다.

(8) 돌아와 돌/동사+어/어말어미+오/보조동사+어/어말어미

다른 예를 통해 표층단순결합경계점 찾는 방법을 하나 더 살펴보자. (8)의 어절 '돌아와(doLawa)'는 동사 '돌(doL)', 어말어미 '어(A)', 보조용언 '오(o)', 어말어미 '어(A)'가 결합된 형태로 두번째 형태소인 어미와 세번째 형태소인 어간 사이에 단순결합경계를 갖고 있다.

표층형: d(ㄷ) o(ㅗ) L(ㄹ) a(아) w(오) a(ㅏ)
| | | |
어휘형: d(ㄷ) o(ㅗ) L(ㄹ) + A(어) + o(오) + A(어)

이 어절에서 표층단순결합경계점을 찾기 위하여 위의 대응 결과를 살펴보면, 첫번째 형태소 'doL(돌)'은 어떤 변화 없이 표층에 그대로 나타나 MCS_1='doL(돌)'이나, 두번째 형태소 'A(어)'는 완전 대체되어 'a(아)'로 나타나 $MCS_2=\lambda$이고, 세번째 형태소 'o(오)' 역시 완전 대체되어 'w'로 나타나 $MCS_3=\lambda$일 뿐 아니라, 마지막 형태소 'A(어)'도 완전 대체되어 'a(ㅏ)'로 나타나 있다. 결과적으로 단순결합경계를 이루는 두 형태소의 특성문자열 MCS_2와 MCS_3가 공백(λ)이기 때문에 표층형에서 이 단순결합경계의 인식은 완전대체문자열을 이용하여야 가능하다. 즉, 두번째 형태소 'A(어)'의 완전대체문자열 'a(아)'가 MCS_1 바로 다음 위치에서 발견되기 때문에 이 완전대체문자열 다음 위치가 표층단순결합경계점이라는 것을 인식해야 한다.

지금까지 설명한 개념들을 사용하여 한국어 형태소 태그 부착 말뭉치에서 어절의 철자변화 현상을 인식하는 논리적인 절차는 알고리즘 1에 기술되어 있다.

알고리즘 1: **어절의 철자변화 현상 인식**

1. 입력: *WPS, WPL, TAG*
2. 출력: 철자변화 형태소와 철자변화 현상 인식 정보
3. $pos = 0$
4. $\omega = 4$
5. $ctype = 0$
6. $LAS_0 = \lambda$
7. For $i = 1$ to k Do
8. SMS = **substr**(WPS, pos, $|MORPH_i| + \omega$)
9. (RAS_i, MCS_i, LAS_i, AS_i) = **identifyMorph**(SMS, LAS_{i-1}, $MORPH_i$, TAG_i, $MORPH_{i+1}$, TAG_{i+1})
10. If $i > 1$ Then
11. $ctype$ = **printMorph**($ctype$, MCS_{i-1}, $MORPH_{i-1}$, RAS_{i-1}, AS_i, LAS_i, $MORPH_i$)
12. EndIf
13. $pos = pos + |MCS_i| + |AS_i|$
14. EndFor
15. If $RAS_{i-1} \equiv \lambda$ and **형태소결합유형**(AS_{i-1}, LAS_{i-2}, RAS_{i-1}) $\in$ {*무변화, 좌변화*} Then
16. If $pos < |WPS|$ Then
17. AS = **substr**(WPS, pos)
18. Else
19. AS = λ
20. EndIf
21. **printMorph**($ctype$, MCS_{i-1}, $MORPH_{i-1}$, RAS_{i-1}, AS, λ, λ)
22. EndIf

알고리즘 1은 정의 4에서 기술한 어절의 표층형 WPS와 어휘형 WPL와 태그 정보 TAG를 입력으로 받아, 우선 정의 5에서와 같이 어절 인덱스 pos의 초기값을 0으로 하여 어절 WPS를 구성하는 형태소마다 표층형태소공간 SMS를 구한다. 각 형태소 $MORPH$의 철자변화 정보 RAS, MCS, LAS, AS는 SMS와의 부분문자열 매칭을 하는 함수 identifyMorph(알고리즘 2)를 사용하여 구한다. 실제 형태소 결합이 일어나는 두번째 형태소부터($i > 1$) 함수 printMorph(알고리즘 4)를 사용하여 형태소 정보를 출력한다. 어절 인덱스 pos는 철자변화 정보로 인식한 MCS와 AS의 크기만큼 증가시켜 다음 형태소의 표층형태소공간을 구하는데 사용한다. 입력 어절의 구성 형태소에 대한 부분문자열 매칭이 끝나면($i > k$), 마지막 형태소의 철자변화 상태와 결합상태에 따라 함수 printMorph를 사용하여 마지막 형태소 정보를 출력한다.

알고리즘 2: **identifyMorph**(SMS, $BLAS$, $CMORPH$, $CTAG$, $NMORPH$, $NTAG$)

1. (PAS, SMS) = **근접매칭전처리**($BLAS$)
2. $m = |CMORPH|$
3. $p = 0$
4. For $i = m$ to $i > 0$ by -1 Do

5. For j = m to $j \geq i$ by -1 Do
6. MCS_p = **substr**(*CMORPH*, m-j, i)
7. RAS_p = **substr**(*CMORPH*, 0, m-j)
8. LAS_p = **substr**(*CMORPH*, m+i-j, j-i)
9. p = p+1
10. EndFor
11. EndFor
12. For i = 0 to $i < p$ Do
13. s = **stringMatching**(*SMS*, MCS_i)
14. **return**(RAS_i, MCS_i, LAS_i, *PAS*+**substr**(*SMS*, 0, s)) if $s \geq$ 0 and ¬ **착시매칭오류**
15. EndFor
16. If 단순결합경계(*CTAG*, *NTAG*) Then
17. *AS*=**substr**(*SMS*, 0, **표층단순결합경계점**)
18. Else
19. *AS* = λ
20. EndIf
21. If *CTAG* ∈ *어미류* Then
22. **return** (*CMORPH*, λ, λ, *PAS*+*AS*)
23. Else
24. **return** (λ, λ, *CMORPH*, *PAS*+*AS*)
25. EndIf

알고리즘 2는 좌우매칭, 최장일치 우선, 근접매칭 전략을 사용하는 부분문자열 매칭을 하여 표층형태소공간에서 형태소의 철자변화 정보를 구하는 함수를 기술한다. 먼저 근접착시조건을 만족하는 착시매칭오류를 방지하기 위하여 정의 8에서 설명한 근접매칭전처리를 한다. 길이가 긴 순서로 인식 대상 형태소 *CMORPH*의 부분문자열 배열 *MCS*을 구하고, 또한 *CMORPH*에서 *MCS*의 왼쪽 문자열과 오른쪽 문자열을 가지고 철자변화 문자열 배열 *RAS*와 *LAS*를 구한다. 문자열 매칭 함수 stringMatching(알고리즘 3)을 이용하여 각 *MCS*에 대하여 표층형태소공간 *SMS*와 매칭한다. 가장 먼저 매칭이 성공하고 착시매칭오류가 아닌 *MCS*에 대하여, 그것과 관련있는 *RAS*와 *LAS*를 찾고, *SMS*에서의 *MCS* 앞의 문자열로 *AS*를 구하여 함수 값으로 리턴한다. 만약 모든 *MCS*가 표층형태소공간에서 발견되지 않는 경우 현재 형태소와 다음 형태소가 단순결합경계를 이루는지 조사한다. 단순결합경계라면 정의 10에서 설명한 표층단순결합경계점을 찾아 *AS*를 구하고, 그렇지 않으면 *AS*를 공백으로 한다. 그리고 현재 형태소의 태그 *CTAG*에 따라 정의 6에서와 같이 *RAS*와 *LAS*를 구하여 함수 값으로 리턴한다.

알고리즘 3: **stringMatching**(*T*, *P*)

1. m = | P |
2. n = | *T* |
3. w = 4
4. For s = 0 to $s <$ **min**{n, w} Do
5. **return** s if $P \cong$ **substr**(*T*, s, m)
6. EndFor
7. **return** -1

알고리즘 3은 주어진 문자열 *T*에서 패턴 문자열 *P*를 처음 위치부터 끝 혹은 최대 네번째 위치까지 위치를 하나씩 옮겨가면서 찾는 간단한 문자열 매칭 함수을 기술한다. 패턴을 찾으면 위치를 옮긴 횟수를 함수 값으로 리턴하고, 찾지 못하면 -1을 리턴한다.

알고리즘 4: **printMorph**(*ctype*, *LMCS*, *LMORPH*, *LAS*, *AS*, *RAS*, *RMORPH*)

1. *ConType* = **형태소결합유형**(*AS*, *LAS*, *RAS*)
2. If *ConType* ∈ {*좌변화*, *좌우변화*} Then
3. **출력**(*LMORPH*, *AS*, *LAS*, *RAS*)
4. ElseIf *ConType* ∈ {*우변화*} Then
5. **출력**(*LMORPH*, λ, λ, λ) if *ctype* ∈ {*무변화*, *좌변화*}
6. **출력**(*RMORPH*, *AS*, *LAS*, *RAS*) if *RMORPH* ≠ λ
7. Else
8. **출력**(*LMORPH*, λ, λ, λ) if *ctype* ∈ {*무변화*, *좌변화*}
9. EndIf

알고리즘 4는 단지 앞에서 인식한 형태소의 철자변화 정보를 출력하는 함수를 기술한다. 철자변화 정보 *AS*, *LAS*, *RAS*로 정의 3에서 설명한 두 형태소의 결합 유형을 구한다. 결합 유형이 우변화가 아닌 한 두 형태소 중 왼쪽 형태소 *LMORPH*만을 출력한다. 우변화 결합이면 왼쪽 형태소와 함께 오른쪽 형태소도 출력한다.

4. 실험 결과

본 논문에서 제안한 방법을 구현하여 두 가지 자료에 대해 실험하였다. 실험 자료 1은 제안 방법이 한국어 어절 구성 형태소 사이에서 일어나는 잘 알려진 철자변화 현상들을 제대로 인식하는가를 실험하기 위한 것으로, 한글맞춤법[11](제16항의 모음조화, 제18항의 불규칙활용, 제32항부터 제40항까지의 축약 등)에서 설명한 형태소 결합에 의해 생기는 철자변화 관련 기술 조항의 예제들을 형태소 태그를 부착하여 구성하였다. 실험 자료 2는 실제 한국어 어절에 나타나는 철자변화 현상들을

제안 방법으로 조사해 보기 위한 것으로, 한국어 처리 연구를 위해 일반적인 형태로 구축된 343,886어절 규모의 한국어 형태소 태그 부착 말뭉치(표 1)에서 어절의 표층형 문자열과 형태소의 단순 결합 문자열이 차이가 나는 잠재적인 철자변화를 보이는 9,391개의 다중 형태소 결합 고유 순 한글 어절들을 추출하여 실험 자료로 구성하였다. 실험 말뭉치에서 자료 2는 전체 어절의 9.9%를 차지하고 있으며, 어절 당 형태소 수와 형태소 결합 수는 각각 3.24개와 2.24개로 철자변화가 없는 어절들보다 다소 높은 특징을 가지고 있다.

문법적인 단순 어절들로 구성된 자료 1에 대한 실험 결과, 제안 방법은 잘 알려진 한국어 형태소 경계의 철자변화 현상들에 대하여 100% 정확히 인식하여 최소한 문법적인 단순 어절에 대해서는 안정적인 인식률을 보이는 것으로 나타났다. 따라서 이에 대한 분석은 불필요해 보여, 아래에서는 실제 언어 사용에서 나타나는 말뭉치 어절들로 구성된 자료 2에 대한 실험 결과를 중심으로 기술한다.

우선, 제안 방법에 의해 인식한 실험 자료 2에 나타난 어절을 구성하는 두 형태소 결합 경계의 철자변화에 대한 통계를 살펴보았다. 표 2는 앞에서 설명한 두 형태소 결합으로 나타난 철자변화 유형(좌변화, 우변화, 좌우변화, 탈락, 대체, 첨가 유형)별로 실험 말뭉치 내의 철자변화 현상 수와 빈도를 조사하여 정리한 것이다. 표 2에

표 1 실험 자료 2에 사용한 한국어 형태소 태그 부착 말뭉치

	무변화	잠재적 철자변화	합 계
어절 수	309,769 (90.1%)	34,177 (9.9%)	343,886 (100.0%)
고유 어절 수	65,300	9,411	74,711
다중 형태소 결합 어절 수	207,908 (60.5%)	34,072 (9.9%)	241,980
다중 형태소 결합 고유 어절 수	57,874	9,391	67,265
형태소 수/어절	2.30	3.24	2.42
형태소 결합 수/어절	1.30	2.24	1.42

표 2 실험 말뭉치에 나타난 형태소 결합 철자변화 통계

형태소 결합 유형		탈락	대체	첨가	합계
좌변화	철자변화 현상 수	12	22	1	35
	빈도	3,221	8,342	32	11,595
우변화	철자변화 현상 수	7	10	6	23
	빈도	3,901	3,712	5,215	12,828
좌우변화	철자변화 현상 수	4	30		34
	빈도	179	10,111		10,290
합 계	철자변화 현상 수	23	62	7	92
	빈도	7,301	22,165	5,247	34,713

표 3 실험 말뭉치에서 철자변화가 나타나는 형태소 결합: 유형(빈도)

좌 \ 우		용언어간	접미사	용언어미		지정사	체언	수식언	조사	접사	어절경계
				선어말	어말						
용언어간	동사			31 (4,786)	40 (13,478)	1 (2)			1 (1)		
	형용사			19 (564)	26 (2,681)	1 (1)					
	보조			14 (2,177)	17 (845)						
용언파생 접미사				7 (2,530)	16 (4,225)						
용언어미	선어말				2 (85)						
	어말	1 (28)			2 (2)	2 (26)			4 (22)		2 (11)
지정사				2 (8)	2 (37)						
체언		1 (1)				8 (1.650)	3 (25)		17 (1,414)	3 (16)	
수식언		1 (1)				2 (12)			2 (5)		
조사						2 (30)					1 (12)
접사				1 (2)	2 (2)	2 (34)					

서 알 수 있는 것은 실험 자료에 나타난 형태소 경계의 철자변화의 유형이 92개이고, 그 중에서 좌변화 유형이 35개(38%), 우변화가 23개(25%), 좌우변화가 34개(37%)이고, 또 탈락 유형이 23개(25%), 대체 유형이 62개(67%), 첨가 유형이 7개(8%)라는 것이다. 또한 어절당 1.02개('형태소 결합 철자변화 빈도(34,713)' / '잠재적 철자변화 다중 형태소 결합 어절 빈도(34,072)')의 형태소 경계 철자변화가 나타남을 유추할 수 있다.

이와 같은 철자변화가 어떤 형태소 태그 결합에서 나타나는가를 보기 위하여 실험 말뭉치에서 철자변화가 나타날 때의 형태소 태그 결합 관계에 대하여 조사하였다. 조사 결과는 표3에 요약되어 있다. 표3에서 알 수 있는 사실은 실험 말뭉치에 나타나는 형태소 결합 철자변화의 70.7%는 용언 어간과 용언 어미의 결합에서 발생하고, 19.5%는 용언파생접미사와 용언 어미의 결합에서 일어나, 사실상 '어간+어미'의 철자변화가 90%이상 차지한다는 것이다. 나머지는 거의 모두가 체언과 조사, 체언과 지정사 등과 같은 '단어+단어'의 철자변화이다. 참고로 용언어간과 지정사 혹은 조사의 결합과 같은 비문법적 형태소 결합에서 나타나는 철자변화는 실험 말뭉치에 태깅 오류가 있음을 암시한다.

이제 실험 말뭉치에 나타난 92개의 형태소 결합 철자변화 현상들을 한글맞춤법과 현대 국어 문법서에 기술된 현상(이하, 문법적 현상)과 그렇지 않은 현상(이하, 비문법적 현상)으로 분류하여 평가해보았다. 먼저 평가 결과를 표 4에 요약하였다.

표 4에 나타난 바와 같이 실험 말뭉치에서 자동 추출한 92개의 철자변화 현상 중 67개(73%)는 문법적인 것이고, 25개(27%)는 비문법적인 현상이다. 빈도로 볼 때 실험 말뭉치에 나타나는 철자변화의 97.7%는 문법적 현상이고, 나머지 2.3%만이 비문법적 현상이다. 문법적 현상 중에서 실험 자료 1과 공통으로 나타난 현상은 유형 44개, 빈도 31,293으로 전체 90%를 차지하는 것으로 나타났다. 이는 한글맞춤법에서 기술된 기본적인 철자변화가 실험 말뭉치에 자주 나타난다는 것을 암시한다. 또 실험 말뭉치에 나타난 철자변화의 두드러진 특징으로는 모음조화, 불규칙 활용, 어간+어미 축약 등 어간+어미 결합에 의한 용언의 철자변화가 대부분(빈도 기준으로 88% 차지함)이라는 것이다. 그리고 비문법적 현상은 분석 결과 모두가 맞춤법 오류와 태그부착 오류 등으로 생긴 철자변화들이었다. 결과적으로 시스템에서 인식한 92개의 철자변화 현상은 모두가 주어진 자료에서 발견할 수 있는 철자변화들이었다. 즉, 시스템 오류로 인식한 철자변화 현상은 하나도 없었다.

5. 결 론

본 논문에서는 사전과 음운 규칙을 이용하지 않고, 주어진 한국어 형태소 태그 부착 말뭉치의 표층형 어절

표 4 실험 말뭉치에 나타난 형태소 결합 철자변화 현상 평가

평가		철자변화 현상			비 고
		설 명	유 형	빈 도	
문법적 현상	자료 1과 공통현상	모음조화	1	2,729	한글맞춤법 제16항
		불규칙 활용	23	11,772	한글맞춤법 제18항
		체언+조사 축약	7	643	한글맞춤법 제33항
		어간+어미 축약	13	16,149	한글맞춤법 제34~40항, 제32항
	기 타	지정사 '이' 탈락	1	1,171	차다(차+이+다)
		매개모임 '으' 첨가	1	53	먹으면(먹+면)
		사이시옷 첨가	1	32	한글맞춤법 제30항
		매개모음 '이' 첨가	1	2	영철이와(영철+와)
		체언+지정사 축약	6	584	무언가(무엇+이+ㄴ가)
		대명사+조사 축약	6	664	제가(저+가)
		어간+어미 축약	6	56	어쩌나(어쩌하+나)
		어미+어미 축약	1	59	가세요(가+시+어+요)
	소 계		67	33,914	
비문법적 현상	맞춤법 오류	타이핑	6	7	해치지(해지+지)
		띄어쓰기	1	1	뭐하러(무엇+하+러)
	태그부착 오류	이형태 기본형 선정	9	154	할만(하+으ㄹ 만)
		철자변화복합어미 사용	5	594	하셨다(하+시었+다)
		기타	4	43	칼이에요(칼+이+어+요) 아니에요(아니+어+요)
	소 계		25	799	
합 계			92	34,713	

문자열과 어휘형 형태소열을 비교하여 어절을 구성하는 형태소의 철자변화 현상을 인식하는 비교적 간단한 방법을 제안하였다. 이 방법은 교과서적인 철자변화 규칙을 사용하지 않기 때문에 두 형태소 결합으로 나타나는 모든 철자변화 현상을 유연하게 인식할 수 있고, 알고리즘 구현만으로 문제를 해결할 수 있기 때문에 비용이 싸다는 특징이 있다. 게다가 한 형태소 태그 부착 말뭉치에 대한 실험에서 실험 말뭉치 어절에 나타나는 철자변화를 안정적으로 인식하는 것으로 나타나, 결과적으로 이 방법은 한국어 단어 인식 연구와 관련이 있는 두 가지 응용에 최소한 활용될 수 있을 것으로 본다. 첫째는 한국어 형태소 태그 부착 말뭉치로부터 형태소 분석 사전을 자동 구축하는데 이용할 수 있다. 둘째는 한국어 형태소 태그 부착 말뭉치로부터 단어 인식의 장애가 되고 있는 형태소 결합 경계의 철자변화에 대한 문맥정보와 통계정보를 획득하는데 활용할 수 있다. 앞으로 이러한 응용 결과에 힘입어 국가에서 구축한 한국어 말뭉치의 폭넓은 활용과 한국어 단어 인식 연구에 대한 효율성과 다양성을 기대해 본다.

참 고 문 헌

[1] 윤준태, 최기선, KAIST 말뭉치에 대한 고찰, 한국과학기술원 기술보고서, CS-TR-99-02/KORTERM-TR-99-02, 1999.

[2] 21세기 세종계획 홈페이지, http://www.sejong.or.kr/sejong-kr/index.html, 2002.

[3] 윤준태, 최기선, 한국어 품사 부착 말뭉치에 대한 고찰, 한국과학기술원 기술보고서, CS-TR-99-138/KORTERM-TR-99-01, 1999.

[4] J. H. Kim and B. G. Jang, Acquiring Rules for Reducing Morphological Ambiguity from POS Tagged Corpus in Korean, Natural Language Engineering, pp. 1-15, 1997.

[5] 양승현, 김영섭, 부분 어절의 기분석에 기반한 고속 한국어 형태소 분석 방법, 정보과학회논문지: 소프트웨어 및 응용, 제27권 제3호, pp. 290-301, 2000.

[6] 김성용, Tabular Parsing 방법과 접속정보를 이용한 한국어 형태소 분석기, 한국과학기술원 석사논문, 1987.

[7] 최재혁, 이상조, 양방향 최장일치법을 이용한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안, 한국정보과학회 논문지, 제20권 제10호, pp. 1497-1507, 1993.

[8] S. S. Kang and Y. T. Kim, Syllable-based Model for the Korean Morphology, Proceedings of COLING 94, Kyoto, Japan, pp. 221-232, 1994.

[9] D. B. Kim, S. J. Lee, K. S. Choi and G. C. Kim. A Two-Level Morphological Analysis of Korean, Proceedings of COLING 94, Kyoto, Japan, pp. 535-539, 1994.

[10] D. B. Kim, K. S. Choi and K. H. Lee, A Computational Model of Korean Morphological Analysis: A Prediction-based Approach, Journal of East Asian Linguistics 5, Kluwer Academic Publishers, pp. 183-215, 1996.

[11] 한글 맞춤법 문교부 고시 제88-1호, 1988.

[12] 남기심, 고영근, 표준 국어문법론, 탑출판사, 1994.

김 덕 봉

1984년 숭실대학교 전자계산학과 졸업(학사). 1986년 한국과학기술원 전산학과 졸업(석사). 1996년 한국과학기술원 전산학과 졸업(박사). 1986년 1월~1989년 9월 산업기술연구원 연구원. 1994년 9월~1995년 3월 일본 ATR 음성번역통신연구소 초빙연구원. 1996년 9월~현재 성공회대학교 컴퓨터정보공학부 부교수. 관심분야는 한국어정보처리, 기계번역, 정보검색