한국어 형태소 분석에서 형태소 분리의 방법에 대하여

-다음절 접사 정보와 음절 사전의 활용-

(On the Division of Syntactic Word to Morpheme in the Process of Korean Morphological Analysis)

황화상*

Abstract

The purpose of this study is to investigate the division of Eojeol(=Syntactic Word) to morpheme in the process of Korean morphological analysis. In Korean Language the target of morphological analysis is Eojeol, which consists of one lexical morpheme(=root) and more than one grammatical morpheme(=affix). Each morpheme in Korean Language is multi-syllabic, and two morphemes(or a part of morpheme) can be fused in one syllable. Therefore it is important to divide Eojeol into morphemes on the design of Korean Morphological Analyzer. In this study we will investigate the division of Eojeol using multi-syllable information of an affix and morphemic fusion information of an syllable.

1. 서론

형태소 분석은 형태론적으로 복합적인 어절(혹은 통사적 단어synta-

* 고려대학교 민족문화연구원

ctic word)을[2] 그것을 구성하고 있는 형태소로 분리하고, 각 형태소의 의미 혹은 기능을 해석하는 형태론적 절차이다. 따라서 형태소 분석은 실질적으로는 형태소 분리와 형태소 해석의 두 과정을 포함한다고 볼 수 있다(황화상 1998). 이 가운데에서 본 연구에서는 다음절 접사 정보와 음절 사전을 활용한 형태소 분리의 문제에 대해 살펴보기로 한다.

형태소 분석 대상으로서의 어절은 하나의 어휘 형태소(lexical morpheme)에 다양한 유형의 문법 형태소(grammatical morpheme)가 결합하여 형성되는데, 어절 구성 요소로서 국어의 형태소는 다음과 같은 속성을 갖는다.(황화상 · 시정곤 2001)

(1) 한국어 형태소의 속성
   ㄱ. 형태소의 융합성 : 둘 이상의 형태소(혹은 형태소의 일부)가 하나의 음절에 융합될 수 있다. 따라서 어떤 형태소는 음소 단위로 실현될 수 있다.
   ㄴ. 형태소의 다음절성 : 하나의 형태소가 둘 이상의 음절로 실현될 수 있다.
   ㄷ. 형태소 결합의 서열성 : 형태소 결합에는 일정한 순서가 있다.
   ㄹ. 형태소 결합의 제약성 : 형태소 결합에는 일정한 제약이 있다.
   ㅁ. 형태소의 생략 가능성 : 어떤 형태소는 특정한 환경에서 생략될 수 있다.

---

* 이 논문은 2002 한국어학회 국제학술대회에서 '한국어 형태소 분석기의 설계와 구현'이라는 제목으로 발표한 원고를 수정하고 보완한 것이다. 이 자리를 빌어 지정 토론을 해 주셨던 강범모 선생님과 여러 회원 선생님들, 그리고 심사 과정에서 부족한 부분을 보충할 수 있도록 도움을 주신 익명의 심사 위원 선생님께도 감사드린다.

2) 통사적 단어의 설정은 굴절과 파생의 구분에 대한 문제 인식에서 출발하는데, 이에 대해서는 김창섭(1984), 고창수(1986, 1992), 임홍빈(1989), 김원경(1993), 시정곤(1994), 김양진(1995), 황화상(1996, 1997, 2001), 유혜원(1997) 등을 참조할 수 있다.

국어의 형태소가 갖는 위와 같은 속성을 적절하게 활용함으로써 형태소 분석의 효율성을 높일 수 있는데, 이 가운데에서 형태소 분리 문제와 관련하여 특히 주목할 수 있는 것은 형태소의 융합성과 다음절성이다.[3)]

둘 이상의 형태소(혹은 형태소의 일부)가 하나의 음절에 융합될 수 있으므로,[4)] 형태소 분리는 음절 단위에서 그치는 것이 아니라 때에 따라서는 음소 단위로까지 확대해야 한다. 예를 들어 '먹었다'는 음절 단위의 분리만으로도 '먹+었+다'의 분석이 가능하지만, '갔다'의 경우에는 형태소 융합 음절 '갔'을 포함하므로 음소 단위로 분리해야만 '가+(어)ㅆ+다'의 분석이 가능하다. 이렇듯 특정 어절의 분석 과정에서 형태소 분리의 단위를 음소까지 확대할 경우 형태소 분리 절차는 그만큼 복잡해질 수밖에 없다.

또한 하나의 형태소가 둘 이상의 음절로 실현될 수 있으므로, 어절 길이가 길어지면 길어질수록 형태소 분리의 복잡도는 현저하게 증가한

3) 형태소 결합의 서열성과 제약성은 형태소 분리 과정에서보다는 형태소 해석 과정에서 유용하게 쓰일 수 있다. 이는 형태소 해석을 효율적으로 하기 위해서는 형태소들 사이의 결합 관계를 토대로 한국어 어절의 구성 양상을 유형화하여 살펴볼 필요가 있는데, 한국어 형태소가 갖는 서열성과 제약성이 형태소들 사이의 결합 관계를 체계화하는 데 기초가 되기 때문이다. 예를 들어 조사 '는'은 모음으로 끝나는 명사에만 결합한다는 제약을 갖는데, 이를 토대로 어절 '먹는'을 '명사+조사'가 아닌 '동사+어미'로 옳게 분석할 수 있다. 대규모 말뭉치 자료에서 추출한, 형태소 결합 관계와 각종 통계 정보에 대해서는 임홍빈 외(2000), 홍종선 외(2001) 등을, 형태소 결합 조건 혹은 제약에 따른, 체언 어절과 용언 어절의 구성 양상에 대해서는 황화상・시정곤(2001)을, 조사와 어미 등 문법 형태소의 배열 순위와 제약에 대해서는 고영근(1967, 1974, 1976), 고창수(1997) 등을, 순수 언어학적 관점에서 한국어 형태소가 갖는 각종 정보에 대해서는 김양진(1999)를 참조할 수 있다.

4) 한국어 어절 가운데에는 둘 이상의 형태론적 분석이 가능한 중의(ambiguous) 어절이 많은데, 이 가운데에는 형태소의 융합으로 인해 생성되는 중의 어절이 상당수 포함되어 있다. 예를 들어 '가는'의 경우 형태소 융합 음절을 포함하지 않는 '가+는', '갈+는'의 분석도 가능하지만, 형태소 융합 음절 '는'을 포함하는 '가늘+은'의 분석도 가능하다.

다.[5] 예를 들어 'abcd' 네 음절로 구성된 어절의 경우, 형태소 융합 음절을 포함하지 않는다고 가정하더라도, 형태소 분리는 논리적으로 다음과 같은 일곱 가지가 가능하다.

(2) ㄱ. abc+d : 형태소는⇒ 형태소+는
ㄴ. ab+cd : 음소까지⇒ 음소+까지
ㄷ. ab+c+d : 예쁘시다⇒ 예쁘+시+다
ㄹ. a+bcd : 벽에다가⇒ 벽+에다가
ㅁ. a+bc+d : 좋으시다⇒ 좋+으시+다
ㅂ. a+b+cd : 먹었는데⇒ 먹+었+는데
ㅅ. a+b+c+d : 먹었겠다⇒ 먹+었+겠+다

이렇듯 한국어 형태소의 융합성과 다음절성으로 인해 형태소 분리의 복잡도는 현저하게 증가한다. 따라서 분석 어절로부터 정확하고 효율적으로 형태소를 분리하기 위해서는 형태소 분석기의 설계 과정에서부터 한국어 형태소가 갖는 융합성과 다음절성을 고려할 필요가 있다. 이와 관련하여 본 연구에서는 형태소 융합의 가능성이 있는 음절을 모아 음절 사전을 구축·활용하는,[6] 그리고 한국어 접사가 갖는 다음절성을 자질화하여 활용하는 방안을 검토할 것이다.

한편 형태소 분석은 그 목적에 따라 분석의 결과를 달리할 수 있다. 순수 언어학적 관점에서 한국어 형태소의 결합 양상을 살펴보는 것이

5) 참고로 강승식(1993:56)에 따르면 한국어 어휘 형태소는 2음절어가 가장 많고(44.238%), 그 뒤로 3음절어(36.850%), 4음절어(13.968%), 5음절어(2.857%), 6음절어(0.936%), 1음절어(0.981%)의 순서를 보이는데, 문법 형태소의 결합까지 고려할 경우 평균 어절 길이는 훨씬 더 길어질 것이다.

6) 강승식(1993)에서는 조사의 첫음절로 사용되는 음절, 1음절 체언에 사용되는 음절, '용언+ㄴ'에 의해 생성되는 음절 등 한국어 음절이 갖는 정보를 다양하게 검토하고, 이를 형태소 분석 과정에서 활용하는 방안을 제시했다. 본 연구는 둘 이상의 형태소(혹은 형태소의 일부)가 융합된 음절만을 대상으로 하며, 형태소 분리를 위해 음절 정보를 활용한다는 점에서 강승식(1993)과는 일정한 차이를 갖는다.

목적이라면 분석 대상 어절을 가장 작은 의미 요소(minimal meaningful element)로서의 형태소(Bloomfield 1933) 단위로 분석해야 한다. 그러나 응용 언어학적 관점에서 형태소 분석기를 설계하고 구현하는 데 있어서는 형태소 단위로의 분석이 꼭 유지되어야 할 필요는 없다.

(3) ㄱ. 서울에서부터 부산까지 다섯 시간이 걸린다.
ㄴ. It takes five hours from Seoul to Busan.

(3ㄱ)의 밑줄 친 어절 '서울에서부터'는 한 개의 어휘 형태소 '서울'과 두 개의 문법 형태소 '에서', '부터'로 구성되어 있다. 따라서 형태소 단위로 분리할 경우 어절 '서울에서부터'는 '서울+에서+부터'로 분석되어야 한다. 그런데 영어 대응 문장 (3ㄴ)을 고려할 경우 접사 '에서'와 '부터'가 각각 일정한 기능을 갖는다고 보는 것보다는 '에서부터' 전체가 영어의 전치사 'from'에 대응한다고 보는 것이 훨씬 더 효율적이므로, 해당 어절을 '서울+에서부터'로 분석할 수도 있다.[7] 본 연구에서는 이와 같이 접사의 결합형을 최대한 전자 사전에 등재하는 것으로 가정하기로 한다.[8]

---

7) 이 경우 접사의 결합형 '에서부터'는 전자 사전에 등재되어 있어야 한다. 그런데 '에서부터'를 하나의 형태소라고 볼 수는 없으므로, 형태소 분석이라는 용어는 적절치 못하다. 오히려 분석 대상을 어절로 한다는 점에서 '어절 분석'이나, 분리 단위를 등재소(listeme)로 한다는 점에서 '등재소 분석'이 적절할 것으로 보인다. 그러나 본 연구에서는 이미 일반화되어 쓰이고 있는 '형태소 분석'이라는 용어를 그대로 사용하기로 한다.

8) 일반적으로 전자 사전의 크기가 크면 분석 규칙은 간단해지며, 전자 사전의 크기가 작으면 분석 규칙은 복잡해진다. 따라서 해당 언어의 분석 과정에서 광범위하게 작용하는 규칙의 경우에는 해당 규칙을 설정하고 전자 사전의 크기를 줄이는 것이 효율적이지만, 작용의 범위가 작은 규칙의 경우에는 설정하지 않고 전자 사전의 크기를 늘리는 것이 훨씬 더 효율적이다. 그런데 조사의 경우 그 결합형의 수가 많지 않으므로(강승식 · 김영택(1992)에 따르면 조사 결합형의 수는 500여개에 불과하다.), 조사 결합형을 최대한 전자 사전에 등재함으로써 형태소 분석 규칙을 간단하게 하는 것이 훨씬 더 효율적이다. 참고로 본 연구는 고려대학교 민족문

2. 다음절 접사 정보를 활용한 형태소 분리

한국어 어절은 하나 이상의 형태소로 구성된다. 그런데 모든 어절은 그것이 단어인 이상 적어도 하나의 어휘 형태소를 포함해야 하므로, 하나의 형태소로 구성된 어절은 그 자체가 어휘 형태소이다. 그리고 어휘 형태소의 결합형은 그 결합형 전체를 사전에 등재하는 것이 효율적이므로, 하나의 어절은 하나의 어휘 형태소만을 포함한다.[9] 그러나 문법 형태소의 경우에는 그 수가 많아 모든 결합형을 사전에 등재할 수는 없으므로, 하나의 어절이 둘 이상의 문법 형태소를 포함할 수는 있다. 또한 어휘 형태소(혹은 '어근root')와 문법 형태소(혹은 '접사affix')를 모두 포함하는 어절의 경우 어휘 형태소는 모든 문법 형태소를 선행한다. 이에 따라 한국어 어절의 구성 양상을 다음과 같이 요약할 수 있다.[10]

화연구원 기계번역연구실에서 개발 중인 한영 기계 번역기(이하 <TransMaster>)에서 실제로 사용한 한국어 형태소 분석기와 전자 사전을 모델로 했음을 밝힌다.

9) 물론 '육층' 등과 같이 띄어 쓰는 것이 원칙이지만 붙여 쓰는 것을 허용함으로써 만들어진 어절, '육천오백구십칠만' 등과 같이 아리비아 숫자를 한글로 풀어 쓴 형태의 어절, '홍길동군, 홍길동박사' 등과 같이 띄어 쓰는 것이 맞지만 습관적으로 붙여 씀으로써 만들어진 어절 등의 경우에는 하나의 어절이 둘 이상의 어휘 형태소를 포함한다. 그러나 하나의 어절에 형태소의 수가 많으면 많을수록 형태소 분석은 그만큼 어려워진다. 예를 들어 'abcd' 네 개의 음절로 구성된 어절은 (2)에서 살펴보았듯이 논리적으로 일곱 가지로 달리 분리될 수 있는데, 해당 어절이 최대 세 개의 형태소로 구성되었다고 가정할 경우 'a+b+c+d'의 분리 가능성을 미리 배제할 수 있으며, 해당 어절이 두 개의 형태소로 구성되었다고 가정할 경우 'ab+c+d, a+b+cd, a+bc+d, a+b+c+d'의 분리 가능성을 미리 배제할 수 있다. 이에 따라 <TransMaster>에서는 하나의 어절은 하나의 어휘 형태소만을 포함한다고 가정하고, 대신 둘 이상의 어휘 형태소를 포함하는 어절은 별도 처리하는 분석 방식을 선택했다.

10) 한국어의 접두사는 파생의 기능만 가지며, 모든 파생어는 전자 사전에 등재될 것이므로, '접사+어근+…'의 내적 구성을 갖는 어절은 형태소 분석의 대상이 되지 않는다.

(4) 한국어 어절의 내적 구성
ㄱ. 어근
ㄴ. 어근+접사
ㄷ. 어근+접사+접사
⋮

형태소 분석은 (4)의 각 어절에서 형태소를 분리함으로써 시작되는데, 형태소의 분리는 일반적으로 음절을 단위로 한다. 형태소 분리 방법에는 우측에서 좌측으로 분리하는 우좌 분석법(right-to-left analysis), 좌측에서 우측으로 분리하는 좌우 분석법(left-to-right analysis), 그리고 양방향 분석법(bi-directional analysis)이 있다.[11] 그런데 한국어 어절의 경우 (4)에서 볼 수 있듯이 어근이 접사에 선행하므로, 우좌 분석법을 이용할 경우 접사를 먼저 분석하게 되며, 좌우 분석법을 이용할 경우 어근을 먼저 분석하게 된다.

본 연구에서는 우좌 분석법이 한국어 형태소 분석에 더 효율적이라고 가정한다. 이는 형태소 분리는 최초 형태소(여러 형태소 가운데에서 처음으로 분석되는 형태소) 분석을 가장 빨리 할 수 있는 방향으로 하는 것이 효율적인데,[12] 한국어의 경우 접사의 수가 어근의 수보다 훨씬 적으며, 접사가 어근보다 중의성(ambiguity)을 덜 가지며(따라서 형태소 결합 정보를 참조하기 쉽다), 어근의 경우에는 미등록어가 있을 수 있지만 접사의 경우에는 미등록어가 없기 때문이다.

우좌 분석법을 통해 접사를 먼저 분석할 경우 해당 어절의 어떤 음절

11) 한국어 형태소의 분석 방법에 대해서는 강승식(1993), 김영택 외(2001) 등을 참조할 수 있다.

12) 이는 어느 한 형태소가 분석되면 그것의 결합 정보를 참조함으로써 다른 형태소의 분석을 위한 사전 탐색의 횟수를 획기적으로 줄일 수 있기 때문이다. 예를 들어 조사 '에서'가 분석된 경우 '에서'는 명사와 직접 결합하므로, 선행 음절 전체가 명사인지를 확인한 후 아닌 경우 곧바로 미등록어 추정을 할 수 있다.

을 기준으로 분리할 것인가 하는 문제가 남는다. 한국어의 경우 하나의 접사가 둘 이상의 음절로 실현될 수 있으며, 특히 본 연구에서는 접사의 결합형을 최대한 전자 사전의 표제어로 등재하는 것을 가정하므로, 정확하고 효율적인 형태소 분석기를 설계하기 위해서는 최초 분리의 기준이 되는 음절을 적절하게 설정하는 것이 무엇보다도 중요하다.

본 연구에서는 한국어의 각 접사가 갖는 음절 수 정보(이하 '다음절 접사 정보')를 활용함으로써 최초 분리의 기준 음절을 설정하는 방안에 대해 검토하기로 한다. 여기에서 다음절 접사 정보는 한국어 접사의 끝에 쓰이는 음절이 갖는 것으로서, 해당 음절이 끝 음절로 쓰이는 모든 접사 가운데에서 최대 길이의 접사가 갖는 음절 수를 수치화한 것이다. 예를 들어 끝 음절이 Y인 접사들의 집합이 다음과 같을 때, 음절 Y는 최대 길이의 접사인 (5ㄹ)의 음절 수 정보 '4'를 다음절 접사 정보로 갖는다.[13]

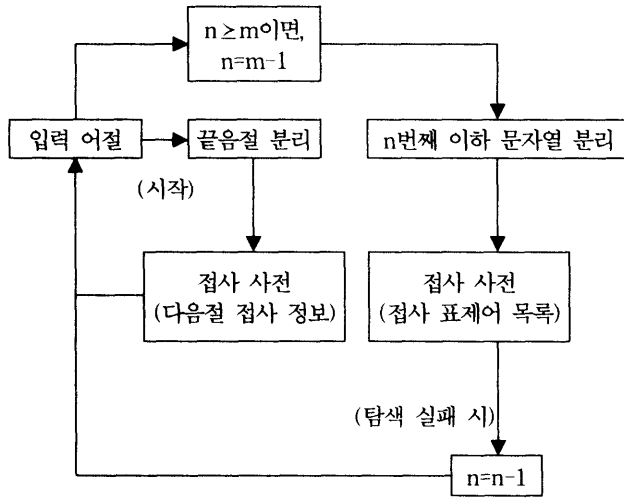
(5) 다음절 접사 정보 : Y=4
ㄱ. Y
ㄴ. XY
ㄷ. WXY
ㄹ. VWXY

각 접사가 갖는 다음절 접사 정보가 주어졌을 때 형태소 분석의 과정을 대략 다음과 같이 형식화할 수 있다. 여기에서 n은 해당 접사의

13) 다음절 접사 정보는 본질적으로 단음절 접사가 갖게 되는 정보이다. 따라서 조사 '부터'와 같이 단독으로 접사로 쓰이지 못하는 음절('터')이 끝 음절인 접사의 경우에는, 그 끝 음절이 접사 사전에 없을 것이므로, 다음절 접사 정보를 표시할 방법이 없다. <TransMaster>에서는 이를 처리하기 위해 '터'와 같은 음절을 접사 사전에 등재하여 다음절 접사 정보를 표시하되, 이것이 단독으로는 접사로 쓰일 수 없다는 자질 정보를 추가했다.

끝 음절이 갖는 다음절 접사 정보이며, m은 입력 어절의 음절 수를 수치화한 것이다. 분석에 앞서 입력 어절의 각 음절에는 뒤에서부터 앞으로 1부터 m까지의 지표가 차례대로 부여되는데, 최초 분리의 기준이 되는 음절은 해당 접사가 갖는 다음절 접사 정보와 같은 지표를 갖는 음절이 된다. 그리고 최초 분리의 기준이 되는 음절이 결정되면 해당 음절 이하 전체를 분석 어절에서 분리하여 접사 사전을 탐색한다. 탐색에 실패할 경우 분리 기준 음절을 하나씩 뒤로 이동하여 동일한 절차를 반복하는데, 이러한 과정은 접사 사전의 탐색에 성공할 때까지 계속된다.

(6) 다음절 접사 정보를 활용한 형태소 분리



이제 실례를 통해 형태소 분리 과정에서 다음절 접사 정보를 어떻게 활용할 수 있을지에 대해 살펴보자. 분석 과정에서 다음과 같은 어절이 주어졌다고 가정해 보자.

(7) ㄱ. 여기까지밖에는
ㄴ. 영희에게서는
ㄷ. 집에서는
ㄹ. 고려대학교에는
ㅁ. 먹는

접사 '는'이 끝 음절로 쓰이는 접사의 집합에 '는, 에는, 에서는, 에게서는, 까지밖에는' 등이 포함된다고 가정하면, 접사 '는'은 최대 길이 접사인 '까지밖에는'의 음절 수 '5'를 다음절 접사 정보로 갖는다. 따라서 예 (7)에 제시한 각 어절은 다음과 같은 과정을 거쳐 분리된다.

(8) ㄱ. '여기까지밖에는'(m=7)
끝 음절 '는' 분리→접사 사전 탐색→다음절 접사 정보 확인(n=5)→'까지밖에는'을 분리하여 접사 사전 탐색→(성공)→ …
ㄴ. '영희에게서는'(m=6)
끝 음절 '는' 분리→접사 사전 탐색→다음절 접사 정보 확인(n=5)→'희에게서는'을 분리하여 접사 사전 탐색→(실패)→n=n-1=4→'에게서는'을 분리하여 접사 사전 탐색→(성공)→ …
ㄷ. '집에서는'(m=4)
끝 음절 '는' 분리→접사 사전 탐색→다음절 접사 정보 확인(n=5)→n=m-1=3→'에서는'을 분리하여 접사 사전 탐색→(성공)→ …
ㄹ. '고려대학교에는'(m=7)
끝 음절 '는' 분리→접사 사전 탐색→다음절 접사 정보 확인(n=5)→'대학교에는'을 분리하여 접사 사전 탐색→(실패)→n=n-1=4→'학교에는'을 분리하여 접사 사전 탐색→(실패)→n=n-1=3→'교에는'을 분리하여 접사 사전 탐색→(실패)→n=n-1=2→'에는'을 분리하여 접사 사전에서 탐색→(성공)→ …
ㅁ. '먹는'(m=2)
끝 음절 '는' 분리→접사 사전 탐색→다음절 접사 정보 확인

(n=5)→n=m-1=1→'는'을 분리하여 접사 사전 탐색→(성공)→
…

이상에서와 같이 형태소 분리 과정에서 다음절 접사 정보를 활용할 경우, 접사 분석의 가능 범위를 줄임으로써, 사전 탐색의 횟수를 얼마간 줄일 수 있다.[14] 특히 이와 같은 분석 방식은 어근이 짧고, 접사가 길고, 다음절 접사 정보가 큰 '방에서만큼은'과 같은 어절을 분석할 경우, 혹은 어근이 길고, 접사가 짧고, 다음절 접사 정보가 작은 '대입수능시험과'와 같은 어절을 분석할 경우 사전 탐색의 횟수를 획기적으로 줄일 수 있다.

그런데 형태소 분리의 과정이 (8)에서와 같이 항상 간단한 것만은 아니다. 예를 들어 '어근+접사+접사' 등 둘 이상의 접사가 연속한 어절의 경우에는 접사의 수만큼 접사 분리의 과정을 되풀이해야 한다. 용언 어절[15] '잡아먹으시었다'의 형태소 분리 과정을 살펴보자. 여기에서 '잡아먹'은 어근 사전에 등재되어 있고, 접사 '시'는 2, '었'은 1, '다'는 3을 각각 다음절 접사 정보로 갖는다고 가정한다.

(9) '잡아먹으시었다'의 형태소 분리[16]
끝음절 '다' 분리→접사 사전 탐색→다음절 접사 정보 확인(n=3)
→'시었다'를 분리하여 접사 사전 탐색→(실패)→'었다'를 분리하여 접사 사전 탐색→(실패)→'다'를 분리하여 접사 사전 탐색→

14) 형태소 융합 음절을 포함하지 않는다고 하더라도, 특별한 장치 없이 어절 길이가 m인 '어근+접사' 유형의 어절을 음절 단위로 분석할 경우 접사 분석의 가능 범위는 끝에서 m-1번째 어절까지가 되며, 따라서 접사 사전의 탐색 횟수는 최대 m-1회가 된다. 그러나 다음절 접사 정보를 활용할 경우 위에서 알 수 있듯이 (8ㄱ), (8ㄷ), (8ㅁ)은 각각 1회, (8ㄴ-)은 2회, (8ㄹ)은 4회 탐색으로 접사를 분석할 수 있다.

15) 특히 어미 결합체의 경우 그 수가 많아 모두 표제어로 등재하기 어려우므로, 용언의 경우 둘 이상의 접사를 포함하는 어절을 흔히 찾아볼 수 있다.

16) 어느 하나의 형태소가 분석되면 분석되지 않고 남아 있는 각 음절의 지표는 우측 끝음절부터 차례대로 1부터 다시 지표가 부여된다.

(성공)→'잡아먹으시었'을 어근 사전에서 탐색→(실패)→'었'을 분리하여 접사 사전 탐색→다음절 접사 정보 확인(n=1)→'었'의 정보를 읽어 온 후 '잡아먹으시'를 어근 사전에서 탐색→(실패)→'시'를 분리하여 접사 사전 탐색→다음절 접사 정보 확인(n=2)→'으시'를 분리하여 접사 사전 탐색→(성공)→'잡아먹'을 어근 사전에서 탐색→(성공)

한편 '어근+접사'의 내적 구성을 갖는다고 하더라도 어근이 미등록어인 어절의 경우, '자라고'('자+라고' 또는 '자라+고')와 같이 형태론적 중의성을[17] 갖는 어절의 경우에는 별도의 처리 장치가 필요하다.

3. 음절 사전을 활용한 형태소 분리

한국어 형태소는 대체로 음절 단위로 실현되므로, 형태소의 분리는 음절 단위를 기본으로 한다. 그러나 다음과 같이 둘 이상의 형태소가 하나의 음절로 융합되는 것도 가능하므로, 때에 따라서는 음소 단위의 분리도 고려해야 한다.

(10) '단'의 형태소 분리 가능성
ㄱ. 음절 단위 분리
① 명사
② 관형사

17) 형태론적 중의성은 어절의 중의성 유형 정보를 활용함으로써 효율적으로 해소할 수 있는데, 이에 대해서는 황화상 · 최정혜(2003)을 참조할 수 있다. 황화상 · 최정혜(2003)에서는 중의성 해소의 과정을 유형적 해소와 어휘 개별적 해소의 두 가지로 구분하고, 유형적 해소 규칙이 어휘 개별적 해소 규칙에 우선하여 적용되는 것으로 보았는데, 이 가운데에서 유형적 해소 규칙은 어절의 중의성 유형 정보를 토대로 설정된 것이다.

③ 부사

ㄴ. 음소 단위 분리

① 다(명사) + 는(조사)

② 다(동사) + 은(어미)

③ 달(동사) +은(어미)

④ 달(형용사) + 은(어미)

⑤ 닳(형용사) + 은(어미)

(10ㄴ)은 종성이 'ㄴ'인 음절의 경우 논리적으로 분리 가능한 예를 보인 것인데, '다+는'의 분리는 '난(나+는)'을 분석하기 위해, '다+은'의 분리는 '간(가+은)'을 분석하기 위해, '달(동사)+은'의 분리는 '운(울+은)'을 분석하기 위해, '달(형용사)+은'의 분리는 '다디단(다디달+은)'의 끝 음절 '단'을 분석하기 위해, '닳+은'의 분리는 '까만(까맣+은)'의 끝 음절 '만'을 분석하기 위해 필요하다.

그런데 이러한 음소 단위의 분리는 불필요한 분리의 가능성을 내포한다는 단점을 갖는다(분석 후보의 과생성). 이기오 · 이근용 · 이용석(1996:256-257)에서 제시된 다음의 예를 보자.

(11) '나는'의 형태소 분리 가능성

ㄱ. 음절 단위 분리

① 나(명사) + 는(조사)

② 나(동사) + 는(어미)

③ 날(동사) + 는(어미)

ㄴ. 음소 단위 분리

①*나느(동사) + ㄴ(어미)

②*나늘(동사) + ㄴ(어미)

③*나늫(동사) + ㄴ(어미)

음절 '는'에 대한 'ㄴ+ㄴ'의 분리는 드물긴 하지만 '꼬느다, 그느다, 노느다' 등의 용언이 있고, 'ㄹ+ㄴ'의 분리도 드물긴 하지만 '가늘다, 넘늘다' 등의 용언이 있으므로 타당하다고 볼 수 있다. 그러나 한국어의 단어 가운데에는 'ㄶ'으로 끝나는 것은 없으므로 'ㄶ+ㄴ'의 분리는 어떤 경우에도 옳지 않다.

분석 후보의 과생성은 사전 탐색의 횟수를 늘림으로써 결국 형태소 분석의 효율성을 떨어뜨리는 결과를 낳는다. 따라서 분석 후보의 수를 줄여, 즉 불필요한 형태소 분리의 가능성을 줄여 형태소 분석의 효율성을 높일 필요가 있다. 본 연구에서는 형태소 융합 가능성이 있는 음절에 대해 가능한 모든 음절 분리 정보를 부여하여 음절 사전을 구축하고,[18] 이를 토대로 형태소 분석을 함으로써, 분석 후보의 과생성 문제를 극복할 수 있다고 본다.[19] 예를 들어 똑같이 종성이 'ㄴ'이지만, 음절 '는'과 음절 '간'은 다음과 같이 서로 다른 유형의 음절 분리 정보를 가질 것이므로, 분석 후보의 과생성 문제는 발생하지 않는다.

(12) ㄱ. '는'의 음절 분리 정보
① 'ㄴ+은' : 꼬느다, 노느다, …
② 'ㄹ+은' : 가늘다, 넘늘다, …
ㄴ. '간'의 음절 분리 정보
① '가+은' : 뛰어간, 달려간, …

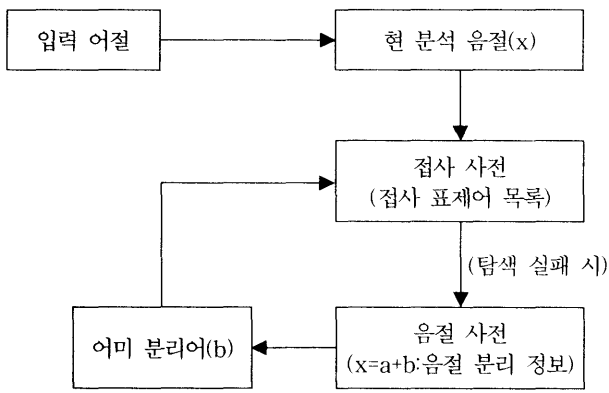
18) <고려대학교 한국어 말모듬 I >(1996)을 대상 자료로 하여 음절 수를 추출한 김홍규 · 강범모(1997)에서 2,305개의 음절 글자가 조사된 바 있다. 그런데 이들 음절 글자가 모두 형태소 융합의 가능성이 있는 것은 아니다. 형태소 융합은 주로 동사, 형용사의 어근이 어미와 결합할 때 발생하는데, 강승식(1993:154-155)에 따르면 용언이 어미와 결합할 때 생성 가능한 음절 글자의 수는 734개이다. 본 연구에서 모델로 삼은 한영 기계번역기 <TransMaster>에서는 700여개의 음절을 모아 음절 사전을 구축했다.

19) 음절 사전을 활용한 형태소 분석의 필요성은 황화상(1998), 황화상 · 시정곤(2001)에서 제시된 바 있는데, 본 연구는 이를 구체화하여 개발한 형태소 분석기를 모델로 한 것이다.

② '갈+은' : 되간, 삭간, …
③ '갛+은' : 빨간, 말간, …

각 음절이 갖는 음절 분리 정보가 주어졌을 때 형태소 분리의 과정을 대략 다음과 같이 형식화할 수 있다. 여기에서 현재 분석 중인 음절 x는 어근(혹은 어근의 일부) a와 어미 b의 융합 음절이라고 가정한다.[20]

(13) 음절 사전을 활용한 형태소 분리



음절 '간'을 포함하는 다음의 어절을 대상으로 하여 음절 사전을 활용한 형태소 분석의 과정을 살펴보자.[21] 여기에서 음절 '간'은 단독으로는 접사로 쓰이지 않으며(따라서 접사 사전에 등재되어 있지 않다.), (?ㄴ)에서와 같이 '가+은, 갈+은, 갛+은'의 음절 분리 정보를 차례대로 갖는다고 가정한다.

---

20) b가 단독 어미가 아니라 어미의 일부인 경우에는 접사 사전의 탐색에 실패할 수도 있는데, 이런 경우의 분석에 대해서는 4장에 제시한 모델 (15)를 참조할 수 있다.
21) 형태소 융합 음절을 포함하는 어절을 분석할 경우 어근의 분석도 고려해야 하므로, 어근의 분석 과정도 같이 보이기로 한다.

(14) ㄱ. 뛰어간

끝음절 '간' 분리→접사 사전 탐색→(실패)→음절 사전 탐색→음절 분리 정보1(가+은) 확인→'은'을 접사 사전에서 탐색→(성공)→'뛰어가'를 어근 사전에서 탐색→(성공)→분석 종료

ㄴ. 되간[22]

끝음절 '간' 분리→접사 사전 탐색→(실패)→음절 사전 탐색→음절 분리 정보1 확인→'은'을 접사 사전에서 탐색→(성공)→'되가'를 어근 사전에서 탐색→(실패)→음절 분리 정보2(갈+은) 확인→'되갈'을 어근 사전에서 탐색→(성공)→분석 종료

ㄷ. 빨간

끝음절 '간' 분리→접사 사전 탐색→(실패)→음절 사전 탐색→음절 분리 정보1 확인→'은'을 접사 사전에서 탐색→(성공)→'빨가'를 어근 사전에서 탐색→(실패)→음절 분리 정보2 확인→'빨갈'을 어근 사전에서 탐색→(실패)→음절 분리 정보3(갛+은) 확인→ '빨갛'을 어근 사전에서 탐색→(성공)→분석 종료

그런데 음절 사전을 활용하더라도 둘 이상의 분석 후보를 갖는 어절의 경우에는 별도의 처리 장치를 마련할 필요가 있다. 예를 들어 (12ㄱ)에서 제시한 단어들은 실제로 그 활용 형태가 '꼬는, 노는, 가는, 넘는'과 같이 될 것인데, '꼬+는, 노+는/놀+는, 가+는/갈+는, 넘+는'의 분석도 가능하므로, 형태소 분리의 복잡도는 훨씬 더 증가한다.[23]

---

22) 「금성판 국어대사전」(김민수외 3인 편, 1992)에 따르면 '되간'은 '①논밭을 다시 갈다. ②가루 등을 다시 갈다.'의 의미를 갖는 동사 '되갈다'의 활용형이다.

23) 이런 이유로 <TransMaster>에서는 '는'을 음절 사전에 포함하지 않았다. 참고로 '자라고(자명사+라고, 자동사+라고, 자라+고), 가는(가명사+는조사, 가동사+는어미, 갈+는, 가늘+은)' 등과 같이 둘 이상의 분석이 가능한 어절의 경우 형태소 분석 과정에서 가능한 모든 후보를 생성하고, 구문 분석 과정에서 가능한 후보 가운데 어느 하나를 선택하는 것이 일반적이다. 그러나 <TransMaster>에서는 구문 분석 과정에서 이들 어절에 대해 형태소 분석을 함으로써 언제나 하나의 분석 후보만을 생성하는 방식을 택한다.

한편 둘 이상의 형태소가 융합되어 있는 음절은 대체로 두 개의 음절로 분리되지만, 두 개의 분리 요소 가운데 어느 하나가 음소 단위가 되는 경우도 있다. 예를 들어 음절 '라, 랐'은 '으' 탈락 용언의 활용형인 '따라, 따랐다' 등에서는 각각 '르+어, 르+었'으로 분리되지만, '르' 불규칙 용언의 활용형인 '달라, 달랐다'에서는 각각 'ㄹ+어, ㄹ+었'으로 분리된다.[24]

4. 형태소 분리의 방법: 종합

이상에서 제안한, 접사가 갖는 다음절 정보와 형태소 융합 음절이 갖는 음절 분리 정보를 활용한 형태소 분리 방법을 간략하게 도식화하면 다음과 같다.[25]

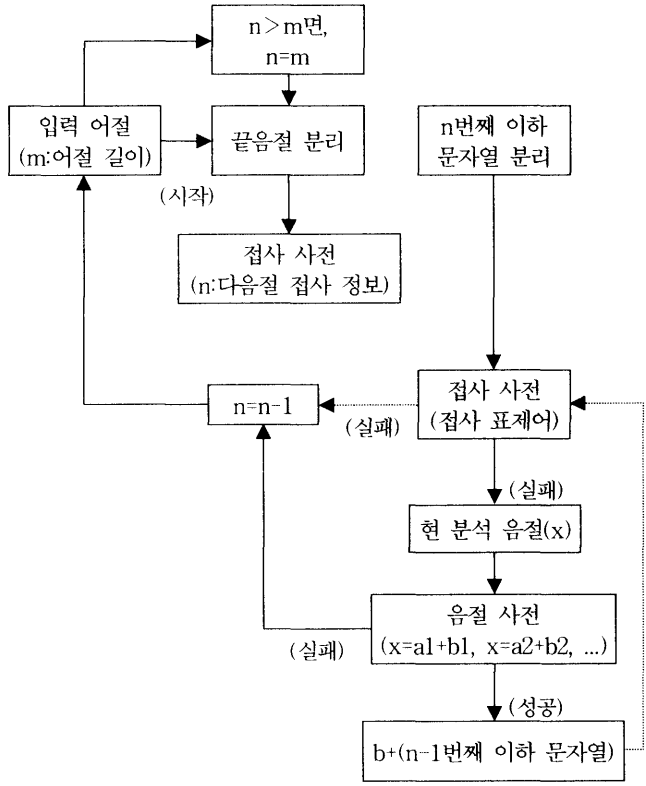
(15) 다음절 접사 정보와 음절 사전을 활용한 형태소 분리[26]

---

24) '르' 불규칙 용언의 활용으로 만들어지는 음절 '라, 러, 랐, 렀' 등을 제외하면 한국어 융합 음절을 모두 두 개의 음절로 분리할 수 있다. 이렇게 음소 단위가 아닌 음절 단위로 분리할 경우 접사 사전의 크기는 훨씬 작아진다. 이에 따라 <TransMaster>에서는 음절 '라, 러, 랐, 렀' 등을 포함하는 어절은 별도의 분석 규칙을 적용하여 분석하고, 'ㄴ, ㄹ, ㅁ, ㅂ, ㅆ' 등 음소 단위의 접사 혹은 이들로 시작하는 'ㄴ다, ㄹ까, ㅂ니다' 등의 접사는 설정하지 않았다.

25) 형태소 분석 과정에서 어근 사전 또한 중요한 역할을 하지만, 설명의 필요에 따라 표시하지 않고 간략하게 제시했다. 어근 사전의 구축에 대해서는 최호철 · 이정식(1998)을 참조할 수 있다.

26) (6)에서 'n≥m이면, n=m-1'로 설정했던 것과는 달리 'n>m이면, n=m'으로 설정한 것은 '간데다가'와 같이 분석 어절의 첫 음절이 형태소 융합 음절일 가능성이 있기 때문이다. 다만 n=m인 경우 전체가 접사인 어절은 없을 것이므로 접사 사전을 탐색하지 않고 곧바로 음절 사전을 탐색하게 된다. 한편 표시하지는 않았지만 현 분석 음절(n)이 x개의 음절 분리 정보를 가질 경우 '접사 사전→음절 사전→접사 사전'의 탐색 과정은 최대 x회 되풀이된다.



다음에 제시한 몇 개 어절의 예를 가지고 위에 제시한 모델을 토대로 형태소 분리의 과정이 어떻게 진행되는지에 대해 살펴보자. 설명을 위해 '은데다가'는 접사 사전에 등재되어 있으며, '가'는 접사 사전에 등재되어 있고 다음절 접사 정보 '5'를 가지며, '픈, 간, 곤, 한'은 음절 사전에 등재되어 있고 각각 '프+은, 가+은, 고+은, 하+은'의 음절 분리 정보를 갖는다고 가정한다.

(16) ㄱ. 먹은데다가

끝 음절 '가' 분리→접사 사전 탐색→다음절 접사 정보 확인

(n=5)→'먹'을 음절 사전 탐색→(실패)→n=n-1=4→'은데다가'를 접사 사전에서 탐색→(성공)→…

ㄴ. 슬픈데다가

끝 음절 '가' 분리→접사 사전 탐색→다음절 접사 정보 확인 (n=5)→'슬'을 음절 사전에서 탐색→(실패)→n=n-1=4→'픈데다가'를 접사 사전에서 탐색→(실패)→'픈'을 음절 사전에 탐색→'은+데다가'를 접사 사전에서 탐색→(성공)→…

ㄷ. 잡아간데다가

끝 음절 '가' 분리→접사 사전 탐색→다음절 접사 정보 확인 (n=5)→'아간데다가'를 접사 사전에서 탐색→(실패)→'아'를 음절 사전에 탐색→(실패)→n=n-1=4→'간데다가'를 접사 사전에서 탐색→(실패)→'간'을 음절 사전에서 탐색→'은+데다가'를 접사 사전에서 탐색→(성공)→…

ㄹ. 피곤한데다가

끝 음절 '가' 분리→접사 사전 탐색→다음절 접사 정보 확인 (n=5)→'곤한데다가'를 접사 사전에서 탐색→(실패)→'곤'를 음절 사전에 탐색→'은한데다가'를 접사 사전에서 탐색→(실패)→n=n-1=4→'한데다가'를 접사 사전에서 탐색→(실패)→'한'을 음절 사전에서 탐색→'은+데다가'를 접사 사전에서 탐색→(성공)→…

'어근+접사'의 내적 구조를 갖는 어절에 대해 접사 부분의 분석 과정만을 간략하게 살펴보았는데, 분석되지 않고 남아 있는 나머지 요소가 이후의 분석 과정에서 어근으로 분석될 경우 형태소 분석은 그대로 종료된다. 그러나 나머지 요소가 어근으로 분석되지 않을 경우에는 때에 따라 이미 분석한 접사 부분의 분석을 다시 해야 할 수도 있으며, 나머지 요소가 미등록어일 가능성도 고려해야 하는 등 분석의 복잡도는 훨씬 더 증가한다.

5. 결론

형태소 분석은 형태소 분리와 형태소 해석의 두 과정을 포함한다. 이 가운데에서 본 연구에서는 형태소 분리의 문제를 중심으로 한국어 형태소 분석기의 설계 과정에서 한국어 형태소가 갖는 다음절성(하나의 형태소가 둘 이상의 음절로 실현될 수 있다.)과 융합성(둘 이상의 형태소, 혹은 형태소의 일부가 하나의 음절에 융합될 수 있다.)을 활용할 수 있는 방안에 대해 살펴보았다.

한국어의 경우 형태소 분리는 음절을 기준으로 하는데, 접사를 먼저 분석하는 우좌 분석법이 더 적절하다. 음절을 형태소 분리의 기준 단위로 할 때 형태소 분석기의 효율성을 높이기 위해서는 최초 분리의 기준 음절을 설정하는 것이 무엇보다도 중요하다. 본 연구에서는 최초 분리의 기준 음절 설정에서 한국어의 각 접사가 갖는 다음절 접사 정보를 활용할 수 있는 방안에 대해 검토했다. 다음절 접사 정보는 각 접사의 끝에 쓰이는 음절이 갖는 정보로서, 해당 음절이 끝 음절로 쓰이는 모든 접사 가운데에서 최대 길이의 접사가 갖는 음절 수 정보를 수치화한 것이다.

한편 한국어의 경우 둘 이상의 형태소가 하나의 음절에 융합될 수 있기 때문에, 때에 따라 형태소 분리의 단위를 음소 단위까지 확대해야 한다. 그러나 형태소 분리를 음소 단위까지 확대할 경우 형태소 분리의 복잡도가 현저하게 증가함은 물론, 분석 후보가 과생성된다는 단점을 갖는다. 본 연구에서는 불필요한 형태소 분리의 가능성을 줄여 형태소 분석의 효율성을 높이기 위해 음절 사전을 구축·활용하는 방안에 대해 살펴보았다. 음절 사전은 형태소 융합 가능성이 있는 음절을 모아 구축하는데, 각 음절은 가능한 모든 음절 분리 정보를 갖는다.

형태소 분석 대상으로서의 어절은 둘 이상의 형태소를 포함하는데,

각 형태소들 사이의 결합에는 일정한 조건 혹은 제약이 존재한다. 형태소 분석은 이러한 조건 혹은 제약에 대한 포괄적인 이해를 바탕으로 할 때 정확성과 효율성을 확보할 수 있다. 따라서 각 형태소들 사이에 존재하는 결합 조건 혹은 제약을 토대로 한국어 어절의 구성 양상을 유형화하고 체계화하는 일이 형태소 분석기를 설계하고 구현하는 데 있어서 무엇보다도 시급하다고 하겠다.

참고문헌

강승식. 1993. "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석." 서울대 박사학위논문.

강승식 · 김영택. 1992. "한국어 형태소 분석기에서 불규칙 용언의 분석 모형." 「한국정보과학회 논문지」 19-2.

고영근. 1967. "현대국어의 선어말어미에 대한 구조적 연구-특히 배열의 차례를 중심으로." 「어학연구」(서울대 어학연구소) 3-1.

고영근. 1974. "현대국어의 종결어미에 대한 구조적 연구." 「어학연구」(서울대 어학연구소) 10-1.

고영근. 1976. "특수조사의 의미분석." 「문법연구」.

고창수. 1986. "어간형성접미사의 설정에 대하여." 고려대 석사학위논문.

고창수. 1992. "국어의 통사적 어형성." 「국어학」 22.

고창수. 1997. "한국어 조사결합에 대한 연구." 「한국어학」 5.

김민수 외 3인 편. 1992. 「금성판 국어대사전」 서울: 금성출판사.

김양진. 1995. "국어 동사의 단어구성 연구." 고려대 석사학위논문.

김양진. 1999. "국어 형태 정보 연구." 고려대 박사학위논문.

김영택 외. 2001. 「자연언어처리」 서울: 생능출판사.

김원경. 1993. "국어 접사피동의 생성론적 연구." 고려대 석사학위논문.

김창섭. 1984. "형용사 파생 접미사들의 기능과 의미-'-답-, '-스럽-, '-롭-, 하-'와 '-적'의 경우-." 「진단학보」 58.

김홍규 · 강범모. 1997. 「한글 사용빈도의 분석」 서울: 고려대 민족문화연구소.

서정곤. 1994. 「국어의 단어형성 원리」 서울: 국학자료원.

유혜원. 1997. "'-시-'에 대한 형태통사적 고찰." 고려대 석사학위논문.

이기오 · 이근용 · 이용석. 1996. "효율적인 한국어 분석을 위한 확장된 최장일치법." 「제8회 한글 및 한국어 정보처리 학술대회 발표 논문집」.

임홍빈 외. 2000. "형태 분석 말뭉치를 이용한 국어 문법 현상의 계량적 연구." 「21세기 세종계획 국어 기초자료 구축」(문화관광부 · 국립국어연구원).

임홍빈. 1989. "통사적 파생에 대하여." 「어학연구」(서울대 어학연구소) 25-1.

최호철 · 이정식. 1998. "자연 언어 처리를 위한 전자 사전 구축 방안." 「어문논집」(안암어문학회) 37.

홍종선 외. 2001. "분석말뭉치를 이용한 한국어 형태소 연접 관계 연구." 「21세

기 세종계획 국어 기초자료 구축」(문화관광부 · 국립국어연구원).
황화상. 1996. "국어 체언서술어의 연구." 고려대 석사학위논문.
황화상. 1997. "국어의 접사 체계." 「한국어학」 5.
황화상. 1998. "자연언어처리를 위한 형태소 분석 방법론." 「어문논집」(안암어문학회) 37.
황화상. 2001. 「국어 형태 단위의 의미와 단어 형성」 서울: 월인.
황화상 · 시정곤. 2001. "형태소 분석을 위한 한국어 어절의 구성 양상 연구." 「제13회 한글 및 한국어 정보처리 학술대회 발표 논문집」.
황화상 · 최정혜. 2003. "한국어 어절의 형태론적 중의성 연구," 「제26차 한국어학회 전국학술대회 발표 논문집」.
Bloomfield, L.. 1933. *Language*. New York: Henry Holt and Co.

황화상(Hwa-Sang Hwang)
고려대학교 강사
주소 : (136-784) 서울시 성북구 상월곡동 101번지 동아아파트 112동 1305호
Tel : 02-918-1973
E-mail : hhs87@korea.ac.kr