형태소 분석을 위한 한국어 어절의 구성 양상 연구

황화상[U] 시정곤
[O]고려대 민족문화연구원 기계번역연구실
한국과학기술원 인문사회과학부
hhs87@ikc.korea.ac.kr, shikaist@cais.kaist.ac.kr

A Study on the Construction Pattern of Korean Syntactic Word for Morphological Analysis

Hwa-Sang Hwang[U] Chung-Kon Shi
[O]Institute of Korean Culture Center for Machine Translation, Korea Univ.
School of Humanities and Social Science, KAIST

요 약

한국어 자연언어처리에서 부딪치는 첫 번째 어려움은 형태소 분석 대상으로서의 어절(통사적 단어)이 형태론적으로 다양한 유형을 갖는다는 데 있다. 따라서 정확하고 효율적인 형태소 분석기를 설계하고 구현하는 데 있어서 우선적으로 요구되는 것은 다양한 유형의 어절을 형태론적으로 분석하여 체계화하는 것이다. 이러한 문제 인식에 따라 본 연구에서는 형태소 결합 관계를 중심으로 체언 어절과 용언 어절의 구성 양상에 대해 살펴보았다.

1. 서론

언어 유형론적으로 교착어에 속하는 한국어는 어간(stem, 어휘 형태소)에 다양한 유형의 어미(ending, 문법 형태소)가 결합함으로써 해당 어절(혹은 통사적 단어)의 문법적 쓰임이 결정된다. 이는 국어의 어절은 형태론적으로 복합적(polymorphemic)이며, 의미 면에서나 기능 면에서나 복합체적인 성격을 가짐을 의미한다.

순수 언어학적 관점에서 형태소 분석은 이렇듯 형태론적으로 복합적인 어절을 그것을 구성하고 있는 형태소로 분리하고, 각 형태소의 의미 혹은 기능을 해석하는 형태론적 절차(morphological process)로 이해된다. 따라서 형태소 분석은 실질적으로는 형태소 분리와 형태소 해석의 두 과정을 포함한다고 볼 수 있다(황화상 1998). 이러한 순수 언어학적 개념은 자연언어처리를 목적으로 하는 전산 언어학의 관점에서도 그대로 적용된다.[1] 그러나 순수 언어학에서의 형태소 분석과 전산 언어학에서의 형태소 분석 사이에는 두 가지 근본적인 차이가 존재한다.

첫 번째 차이는 분리 단위의 차이이다. 순수 언어학에서 형태소 분리의 단위는 최소 유의미 단위(minimal meaningful units)로서의 형태소가 되지만, 전산 언어학에서 형태소 분리의 단위는 전자 사전의 표제어 혹은 등재소(listeme)가 된다.[2] 따라서 언어학적 관점에서는 특정 어절에 대한 형태소 분석의 결과에 큰 차이가 없지만,[3] 전산 언어학적 관점에서는 전자 사전의 내용에 따

*본 연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음.

1) 김영택(1994:56)에서는 형태소 분석을 포함하여 자연 언어의 분석을 '분석 후보의 생성과 그 후보들로부터 옳은 분석 결과를 선택하는 과정'으로 기술했다.

2) 따라서 전산 언어학적 관점에서는 본질적으로 '형태소 분석'이라는 용어는 적절치 못하다고 볼 수 있다. 오히려 분석 대상을 어절로 한다는 점에서 '어절 분석'이나, 분리 단위를 등재소로 한다는 점에서 '등재소 분석'이 적절할 것으로 보인다. 그러나 본 연구에서는 이미 일반화되어 쓰이고 있는 '형태소 분석'이라는 용어를 그대로 사용하기로 한다.

3) 물론 순수 언어학적 관점에서도 형태소 분석의 결과가 항상 일치하는 것은 아니다. 예를 들어 '구두닦이, 해돋이, 돈벌이' 등 이른바 동사성 합성어(verbal compound)의 경우 직소 분석(immediate constituent analysis)에 관련된 논란이 끊임없이 이어지고 있다. 그러나 직소 분석의 다름에 관계없이 각 단어를 세 개의 형태소(예를 들

라 형태소 분석의 결과가 크게 다를 수 있다. 예를 들어 '어른스러웠다'의 분석에서 보이는 차이에 대해 살펴보자.

(1) ㄱ. 순수 언어학적 분석
[[[[어른]+스럽]+었]+다]
ㄴ. 전산 언어학적 분석
① [어른스러웠다]
② [[어른스러웠]+다]
③ [[[어른스럽]+었]+다]
[[[어른스러우]+었]+다]
[[[어른스러]+웠]+다]
④ [[[[어른]+스럽]+었]+다]
[[[[어른]+스러우]+었]+다]
[[[[어른]+스러]+웠]+다]

순수 언어학적 관점에서 '어른스러웠다'는 (1ㄱ)에서와 같이 명사 어근 '어른'에 형용사 파생 접사 '-스럽-'이 결합하여 형용사 어간 '어른스럽'이 형성되고, 여기에 다시 과거 시제 형태소 '-었-'과 문장 종결 형태소 '-다'가 순차적으로 결합하여 형성된 것으로 설명된다. 그러나 전산 언어학적 관점에서는 (1ㄴ)에서와 같이 논리적으로 다양한 분석이 가능하다. 먼저 (1ㄴ-①)은 어절 전체를, (1ㄴ-②)는 형용사 어간과 과거 시제 형태소의 결합형을, (1ㄴ-③)은 형용사 어간의 이형태(혹은 이형태의 일부)를, (1ㄴ-④)는 명사 어근과 형용사 파생 접사의 이형태(혹은 이형태의 일부)를 전사 사전의 표제어로 했을 때 가능한 분석이다.

두 번째 차이는 분석 목적의 차이이다. 순수 언어학에서는 해당 어절이 갖는 의미와 기능을 밝히는 것뿐만 아니라, 그것이 어떻게 그러한 의미와 기능을 갖게 되었는지를 규명하는 것을 목적으로 한다. 따라서 순수 언어학에서의 형태소 분석은 '발견'과 '설명'을 위한 방법론이라고 할 수 있다. 이와 달리 전산 언어학에서의 형태소 분석은 단지 그 어절이 갖는 의미와 기능을 밝히는 것을 목표로 할 뿐이다. 따라서 전산 언어학에서의 형태소 분석은 '발견'을 위한 방법론이라고 볼 수 있다(황화상 1998).

요컨대 순수 언어학적 관점에서나 전산 언어학적 관점에서나 형태소 분석은 개념적으로 동질적이지만, 형태소 분리의 단위나 형태소 분석의 목적 등에서 두 접근 방식 사이에는 일정한 차이가 존재한다. 그런데 형태소 분석기의 구현이라는 실질적인 목표 아래에서는 결과적으로 전산 언어학적 관점에서의 접근이 요구되며, 순수 언어학적 접근은 이에 대한 이론적 토대를 제공할 수 있다는 점에서 일정한 역할을 할 것으로 기대된다. 이에 따라 본 연구에서는 전산 언어학적 관점에서의 형태소 분석을 가정하고, 형태소 분석기의 효율적 설계와 구현을 위한 기초 연구로서 한국어 어절의 구성 양상에 대해 체언 어절과 용언 어절을 중심으로 살펴보기로 한다.

어 '구두+닦+이')로 분리한다는 점에서는 대체로 의견이 통일된다. 한국어 복합어의 직소 분석에 대해서는 이익섭(1965), 허웅(1966/1975), 김창섭(1994/1996), 시정곤(1994), 황화상(2001) 등을 참고할 수 있다.

2. 한국어 어절의 내적 구성과 형태소의 속성

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 형태소와 어휘 형태소의 결합형은 그 결합형 전체를 대체로 사전에 등재하는 것이 효율적이므로, 하나의 어절은 하나의 어휘 형태소만을 포함한다. 또한 어휘 형태소와 문법 형태소를 모두 포함하는 어절의 경우 어휘 형태소는 모든 문법 형태소를 선행한다. 이를 토대로 한국어 어절의 내적 구성을 다음과 같이 요약할 수 있다.[4]

(2) 한국어 어절의 내적 구성
ㄱ. 한국어 어절은 하나 이상의 형태소로 구성되는데, 하나의 형태소만을 포함할 경우 그것은 어휘 형태소이다.
ㄴ. 한국어 어절은 어휘 형태소를 반드시 하나만 포함하며, 문법 형태소는 둘 이상 포함할 수 있다.
ㄷ. 한국어 어절에서 어휘 형태소는 모든 문법 형태소를 선행한다.

위 정보는 한국어 형태소 분석기의 설계 및 구현에 있어서 형태소 분리, 그리고 전자사전 탐색의 효율성 문제와 관련하여 중요성을 갖는다. 먼저 (2ㄱ)은 '(전체 어절에 대한) 어간 사전 탐색→형태소 분리'의 순서로 형태소 분석을 진행할 수 있음을 의미한다.[5] 다음으로 (2ㄴ)과 (2ㄷ)은 형태소를 분리할 경우 끝 음절부터 하는 것이, 그리고 '끝 음절 분리→(끝 음절에 대한) 어미 사전 탐색→(선행 음절 전체에 대한) 어간 사전 탐색'의 순서로 형태소 분석을 하는 것이 효율적임을 의미한다. 이는 형태소 분리는 최초 형태소(여러 형태소 가운데 처음으로 분석되는 형태소) 분석을 가장 빨리 할 수 있는 방향으로 하는 것이 효율적인데,[6] 한국어의 경우 문법 형태소의

4) 위에 제시한 한국어 어절의 내적 구성과 아래 제시할 한국어 형태소의 속성은 황화상(1998)에서 제시한 것을 일부 내용을 추가하여 재구성한 것이다.

5) 본 연구에서는 전사사전을 크게 어간 사전과 어미 사전의 두 가지로 나누어 구축하는 것으로 가정한다.

6) 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. 그러나 조사 '만'의 경우 다른 조사가 선행할 수 있으므로(예:여기에서만), 선행 음절에 대해서도 음절 분리의 절차를 필요로 한다.

음절 길이가 어휘 형태소의 음절 길이보다 짧으며(따라서 음절 분리의 횟수가 적다), 문법 형태소가 어휘 형태소보다 중의성(ambiguity)을 덜 가지며(따라서 형태소 결합 정보를 참조하기 쉽다), 문법 형태소의 경우에는 미등록어가 없기 때문이다.

한편 한국어 형태소는 다른 형태소와의 관계에서 일정한 속성을 갖는데, 이를 적절히 활용함으로써 형태소 분석의 효율성을 높일 수 있다. 먼저 한국어 형태소는 둘 이상의 형태소(혹은 형태소의 일부)가 하나의 음절에 융합될 수 있으며, 하나의 형태소가 둘 이상의 음절로 실현될 수 있다는 속성을 갖는다. 또한 형태소 결합에는 일정한 순서와 제약이 있으며, 어떤 형태소는 특정 환경에서 생략될 수 있다는 속성도 갖는다. 이를 토대로 한국어 형태소의 속성을 다음과 같이 요약할 수 있다.

(3) 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.

한국어 어절(혹은 형태소)은 대체로 음절 단위로 구성되므로, 형태소 분리는 음절 단위로 하는 것이 바람직하다. 그러나 이상의 형태소가 하나의 음절로 융합되는 것도 가능하므로(3ㄱ), 때에 따라서는 다음과 같이 음소 단위의 분리도 고려해야 한다.

(4) '단'의 형태소 분리 가능성
- ㄱ. 음절 단위 분리
  - ① 명사
  - ② 관형사
  - 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 닿(형용사) + 은(어미)

(4ㄴ)은 종성이 'ㄴ'인 음절일 경우 논리적으로 분리 가능한 예를 보인 것인데, 각각 '난(나+는), 간(가+은), 운(울+은), 단(달+은), 까만(까맣+은)'의 예를 참조할 수 있다. 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.

(5) '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*나눟(동사) + ㄴ(어미)

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. 하나의 방법은 형태소 융합 가능성이 있는 음절을 모아 음절 사전을 구축하고,[7] 음절 정보를 토대로 형태소 분석을 하는 것이다(황화상 1998). 예를 들어 '는'의 경우 국어 동사 어간에는 '눟'으로 끝나는 것이 없으므로 '나는'의 경우 '나눟+ㄴ'의 분리 가능성은 미리 배제할 수 있다.[8]

3. 체언 어절의 구성 양상

체언 어절은 해당 어절에 포함된 어휘 형태소가 체언(명사, 대명사, 수사)인 어절을 말한다. 그런데 체언의 경우 조사의 결합 없이 단독으로 쓰일 수 있다는 속성을 가지므로, 체언 어절은 형태소 결합 유무에 따라 크게 단일 형태소 어절과 복합 형태소 어절, 그리고 이 두 가지 구성이 모두 가능한 단일-복합 형태소 어절로 나뉜다.

(6) 체언 어절의 구성 양상 1
- ㄱ. 단일 형태소 어절
- ㄴ. 복합 형태소 어절
- ㄷ. 단일-복합 형태소 어절

---

7) <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.

8) 황화상(1998)에 따르면 음절 정보를 토대로 한 형태소 분석이 오히려 분석 후보를 과생성하는 경우도 있으나, 형태소 결합 정보를 적절히 활용할 경우 큰 문제가 되지는 않는다.

(6ㄱ)의 단일 형태소 어절은 다시 두 가지 유형으로 구분되는데, 하나는 결합해야 할 조사가 생략된 경우이며, 다른 하나는 그것이 어근적 단어인[9] 경우이다.

(7) 체언 어절의 구성 양상 2 : 단일 형태소 어절
ㄱ. 조사 생략 어절
어제 산 책 읽었니?
ㄴ. 어근적 단어 어절
그는 국제 무역으로 큰돈을 벌었다.

어근적 단어 어절의 경우에는 통사적으로 후행 서술어의 논항(argument)으로서 기능하는 것이 아니고 뒤따르는 명사의 수식어로서 기능을 한다. 그러나 조사 생략 어절의 경우에는 문장에서 논항으로서 기능하므로, 생략된 조사를 복원하거나 통사적 기능을 결정해 줄 필요가 있다. 이는 다음과 같이 동일한 형태의 어절이 다른 문법적 기능을 할 수 있기 때문이다.

(8) ㄱ. 주어 : 어제 산 책 재미있니?
ㄴ. 목적어 : 어제 산 책 읽었니?

또한 조사 생략 어절 가운데에는 다음과 같이 동음이의 어절과 단의 어절이 구분되는데, 동음이의 어절의 경우 다시 품사 중의성(ambiguity)을 갖는 항과 그렇지 않은 항이 구분된다.[10]

(9) 체언 어절의 구성 양상 3 : 조사 생략 어절
ㄱ. 단의 어절
의자 밑에 고양이가 있다.
ㄴ. 동음이의 어절
① 품사 중의성 어절
저 들 건너편에 우리집이 있다.
내가 들 가방이 어느 것이니?
먹 가는 사람이 누구니?
밥 먹고 가거라.
② 단순 동음이의 어절
배(梨) 주고 속 빌어먹는다.
철수는 배(腹) 아프다고 일찍 갔다.
나는 배(船) 타고 가는 여행을 좋아한다.

품사 중의성 어절 가운데에서 '먹'의 경우에는 동사로서는 단독으로 쓰일 수 없다는 속성을 가지므로, 중의성 해소에 특별한 어려움이 없다. 그러나 품사 중의성 어절 가운데에서 '들'이나 단순 동음이의 어절 '배'의 경우에는 형태적으로는 차이가 드러나지 않는다.

(6ㄴ)의 복합 형태소 어절은 크게 체언에 조사가 결합한 것, 체언에 동사화 접사 '-이-'(혹은 '-이' 결합 어미)와 '-답-', '-같-'이 결합하고 여기에 다시 어미가 결합한 것, 동사화 접사 '-이-'가 생략되어 어미가 직접 결합한 것의 세 유형으로 나뉜다.[11]

(10) 체언 어절의 구성 양상 4 : 복합 형태소 어절
ㄱ. 조사 결합 어절
책이 두껍다.
이 공책은 잘 찢어진다.
ㄴ. 동사화 접사 결합 어절
그는 훌륭한 학생이다.
그는 용감한 군인답다.
그는 너그러운 사람같다.
ㄷ. 어미 결합 어절
그는 친절한 의사다.
저것은 게으른 소다.

(10ㄱ)의 조사 결합 어절의 경우에는 중의항을 포함하는 어절과 그렇지 않은 어절로 구분되는데, 중의항을 포함하는 어절의 경우에는 다시 어간이 중의항인 어절, 조사가 중의항인 어절, 어간과 조사가 모두 중의항인 어절로 나뉜다.

(11) 체언 어절의 구성 양상 5 : 조사 결합 어절
ㄱ. 중의항 비포함 어절
의자에 앉았다.
ㄴ. 중의항 포함 어절
① 어간 중의항 어절
들에 꽃이 활짝 피었다.
그는 열흘 동안 물만 먹었다.
② 조사 중의항 어절
이 사람은 의리가 있는 사람이다.
이 책도 다 읽었다.
③ 어간-조사 중의항 어절
이 먹은 아주 비싸다.
밥 대신 떡을 먹는 사람이 아주 많다.

어간 중의항 어절이나 조사 중의항 어절의 경우에는 각각 후행 조사나 선행 어간의 품사 정보를 참조할 수 있으므로 처리에 어려움이 없다. 어간-조사 중의항 어절 가운데에서 '먹는'의 경우에는 '는'이 조사일 경우 끝 음절에 종성이 없는 명사만을 어간으로 취한다('의자는')는 형태소 결합 제약을 가지므로(예를 들어 '의자는'), '먹는'의 경우 '먹'을 명사가 아니라 동사로 분석하는 데 큰 어

9) 어근적 단어 혹은 어근은 "하나의 단어에서 의미적으로 가장 중심이 되는 형태소로 접사를 제외한 핵심부분이면서, 통사적 기능 요소(격조사, 어미 등)와의 직접적 결합이 불가능한 것"(남기심·고영근 1998)을 말한다.

10) 이밖에 다의어도 있으나 한 단어로 보는 것이 일반적이므로, 본 연구에서는 이에 대해서는 고려하지 않기로 한다.

11) (10ㄴ)과 (10ㄷ)의 경우 예에서 알 수 있듯이 부사의 수식을 받는 것이 아니고 관형어의 수식을 받는다. 따라서 본 연구에서는 이들이 용언인 동시에 체언이라고 가정한다. 이러한 가정의 근거와 언어학적 설명에 대해서는 황화상(1996, 2001)을 참고할 수 있다.

려움이 없다. 그러나 어간-조사 중의항 어절 가운데에서도 '먹은'의 경우에는 '은'이 '는'과 같은 형태소 결합 제약을 갖지 않으므로, '명사+조사', '동사+어미'의 분석이 모두 가능하다.

(10ㄴ)의 동사화 접사 결합 어절이나 (10ㄷ)의 어미 결합 어절의 경우에도 다음과 같이 선행 어간이 중의항인 어절과 그렇지 않은 어절이 구분된다.

(12) 체언 어절의 구성 양상 6
:동사화 접사 결합 어절, 어미 결합 어절
ㄱ. 어간 비중의항 어절
이것은 책상이다.
이것은 의사다.
ㄴ. 어간 중의항 어절
이것은 물이다.
이것은 개다.

어간 중의항 어절 가운데에서 '물이다'의 경우에는 '-이-'가 동사에는 결합하지 못한다는 형태소 결합 제약을 가지므로, '물'을 동사가 아닌 명사로 분석할 수 있다. 그러나 '개다'의 경우에는 '개'와 '다'가 모두 중의항이므로, '명사+어미'와 '동사+어미'의 분석이 모두 가능하다.

(6ㄷ)의 단일-복합 형태소 어절은 복합 형태소 어절일 경우의 형태소 결합 유형에 따라 크게 세 가지로 구분되는데, 첫 번째는 끝 형태소가 조사인 경우이며, 두 번째는 끝 형태소가 어미인 경우이며, 세 번째는 끝 형태소가 조사일 수도 있고 어미일 수도 있는 경우이다.

(13) 체언 어절의 구성 양상 7 : 단일-복합 형태소 어절
ㄱ. 끝 형태소 조사 어절
①단독 : 경춘 가도 가는 길이 복잡하다.
②복합 : 이제 가도 소용이 없다.
ㄴ. 끝 형태소 어미 어절
① 단독 : 문 닫고 다녀라.
② 복합 : 저 개가 내 다리를 문 개다.
ㄷ. 끝 형태소 조사-어미 어절
①단독 : 나는 요즘 낚 키우는 재미에 푹 빠졌다.
②복합 : (조사) 낚 요즘 아주 바쁘다.
(어미) 이가 낚 후에 곧바로 씩었다.

위에 제시한 예 가운데에서 끝 형태소가 어미로 분석될 수 있는 경우에는 해당 어미의 형태소 결합 제약을 토대로 어느 정도 형태소 분석을 옳게 할 수 있다. 예를 들어 '문'의 경우 뒤따르는 형태소가 명사가 아닌 경우에는 '문'을 단일 명사 어절로 분석할 수 있다. 그러나 뒤따르는 형태소가 명사인 경우, 그리고 해당 어절의 끝 형태소가 조사로 분석될 수 있는 경우에는 형태소 결합 정보나 인접 형태소의 품사 정보 등 형식적 표지만으로는 옳은 분석 결과만을 도출하기 어렵다.

이상에서 형태소 결합 양상에 따라 한국어 체언의 어절 구성 양상에 대해 살펴보았다. 이를 요약하면 다음과 같다.

(14) 체언 어절의 구성 양상
ㄱ. 단일 형태소 어절
① 조사 생략 어절
-단의 어절
-동음이의 어절 : 단순 동음이의 어절, 품사 중의성 어절
② 어근적 단어 어절
ㄴ. 복합 형태소 어절
① 조사 결합 어절
-중의항 비포함 어절
-중의항 포함 어절 : 어간 중의항 어절, 조사 중의항 어절, 어간-조사 중의항 어절
② 동사화 접사 결합 어절 : 어간 비중의항 어절, 어간 중의항 어절
③ 어미 결합 어절 : 어간 비중의항 어절, 어간 중의항 어절
ㄷ. 단일-복합 형태소 어절
① 끝 형태소 조사 어절
② 끝 형태소 어미 어절
③ 끝 형태소 조사-어미 어절

한편 체언 어절 가운데에서 복합 형태소 어절이나 단일-복합 형태소 어절 가운데에는 많은 예가 형태소 융합 음절을 포함한다. 그런데 앞에서 살펴보았듯이 음절 정보를 활용할 경우 분석의 효율성을 얼마간 높일 수 있다. 예를 들어 '문'의 경우 한국어에서 '무'로 끝나는 동사 어간이 없으므로, '물+은'의 가능성만을 고려함으로써('문, 깨문, 다문, 드문' 등) 사전 탐색의 횟수를 줄일 수 있다.

4. 용언 어절의 구성 양상

용언 어절은 해당 어절에 포함된 어휘 형태소가 용언(동사, 형용사)인 어절을 말한다. 그런데 용언의 경우 대체로 어미의 결합 없이 쓰일 수 없으나, 어미가 생략되어 쓰이기도 하며, 한 어절이 단일 형태소 어절과 복합 형태소 어절의 둘로 분석될 수도 있다. 이에 따라 용언 어절을 크게 단일 형태소 어절(=어미 생략 어절), 복합 형태소 어절, 단일-복합 형태소 어절로 구분할 수 있다.

(15) 용언 어절의 구성 양상 1
ㄱ. 단일 형태소 어절
ㄴ. 복합 형태소 어절
ㄷ. 단일-복합 형태소 어절

(15ㄱ)의 단일 형태소 어절은 다시 어떤 어미가 생략되었는지에 따라 크게 관형사형 어미 생략 어절, 부사형 어미 생략 어절, 종결 어미 생략 어절로 구분된다.

(16) 용언 어절의 구성 양상 2 : 단일 형태소 어절
ㄱ. 관형사형 어미 생략 어절
내 힘으로는 도저히 문을 열 수 없었다.
ㄴ. 부사형 어미 생략 어절
열심히 뛰어가 보아라.
ㄷ. 종결 어미 생략 어절
지금 학교에 걸어가.

(16ㄱ)의 관형사형 어미 생략 어절은 종성이 'ㄹ'인 동사 어간이 관형사형 어미 '을'과 결합할 때 생성된다. 그리고 (16ㄴ)의 부사형 어미 생략 어절과 (16ㄷ)의 부사형 어미 생략 어절은 종성이 없는 동사 어간 가운데에서 중성이 'ㅏ, ㅓ, ㅐ, ㅔ'인 것이 '어, 아'와 결합할 때 생성된다. 따라서 어절 전체가 동사 어간으로 사전에 등재되어 있는 경우 끝 음소 정보를 활용하여 적절한 형태의 어미를 생성할 필요가 있다. 다만 탐색된 어절 전체가 동사 이외의 다른 품사 정보를 가질 경우에는 품사 중의성을 해소한 이후에 어미를 생성해야 한다.[12]

(15ㄴ)의 복합 형태소 어절은 몇 가지 기준에 따라 하위 유형화가 가능하다. 먼저 용언은 체언과는 달리 불규칙 활용을 하기도 하며, 어간의 일부가 탈락하기도 하며, 어미와의 융합 가능성도 많은데, 불규칙 형태(불규칙 활용 형태, 탈락 형태, 음절 융합 형태)를 포함한 어절의 경우 형태소 분석의 복잡도가 현저히 증가한다. 이에 따라 용언 어절을 불규칙 형태를 포함하는 어절과 그렇지 않은 어절로 크게 구분할 수 있다.

(17) 용언 어절의 구성 양상 3 : 복합 형태소 어절
ㄱ. 규칙 형태 포함 어절
밥을 맛있게 먹었다.
신문을 읽은 후에 학교에 갔다.
ㄴ. 불규칙 형태 포함 어절
① 불규칙 활용 형태 포함 어절
영희는 아름다운 소녀다.
너는 아름다우니 소녀다.
철수는 열심히 동생을 도왔다.
그녀에게는 빨간 옷이 잘 어울린다.
② 탈락 형태 포함 어절
집을 지은 후에 그곳으로 이사를 갔다.
칼로 자를 부분에 금을 그었다.
입을 다문 상태로 마주 보고 있었다.
편지를 써 보아라.
③ 음절 융합 형태 포함 어절
그녀는 반가운 얼굴로 그에게 뛰어갔다.
뛰어간 사람도 있고 걸어간 사람도 있다.

(17ㄴ-①)의 불규칙 형태 포함 어절은 다시 어미와 음절이 융합되는 것('아름다운, 도왔다')과 그렇지 않고 어간만 교체되는 것('아름다우니')이 구분된다. 음절이 융합되는 경우에는 음절 정보를 활용함으로써, 그리고 어간만 교체되는 경우에는 불규칙 형태를 규칙 형태와 함께 사전의 표제어로 등재함으로써 분석의 효율성을 높일 수 있다.[13]다만 '빨간'에서 '간'의 경우에는 (17ㄴ-③)의 '뛰어간, 걸어간'의 예에서 알 수 있듯이 '가+은'과 '갈+은'의 분리가 가능하므로, 이에 대한 처리가 요구된다. 음절 '랐'이나 '렀'의 경우에는 그 양상이 좀 더 복잡하다.

(18) '랐/렀'의 형태소 분리 가능성
ㄱ. 형과 동생이 아주 달랐다.
ㄴ. 출발한 지 30분만에 도착지에 이르렀다.
ㄷ. 철수는 나를 잘 따랐다.

위 예에서 알 수 있듯이 '랐/렀'을 포함하는 어절의 경우 세 가지로 다르게 분석된다. 먼저 '달랐다'와 같이 선행 형태의 종성이 'ㄹ'인 경우 'ㅡ+었'으로 분리되며, 선행 형태의 종성이 없는 경우에는 '이르렀다'와 같이 단순히 과거 시제 형태소 '었'이 '렀'으로 교체되는 것과 '따랐다'와 같이 '르+었'으로 분리되는 것이 구분된다.

(17ㄴ-②)의 탈락 형태 포함 어절은 어간이 어미와 융합되는 것('써')과 단순히 어간의 일부가 탈락되는 것('지은, 그었다, 다문')이 구분된다. 어간이 어미와 융합되는 경우에는 음절 정보를 활용함으로써 어렵지 않게 형태소 분석을 할 수 있다. 어간의 일부가 단순히 탈락하는 어절은 다시 탈락한 형태가 사전 표제어로 있는 것('지은, 그었다')과 그렇지 않은 것('다문')이 구분된다. 탈락 형태가 사전 표제어로 없는 경우에는 음절 정보를 활용하여 형태소 분석을 할 수 있다. 탈락 형태가 사전 표제어로 있는 경우에는 분석의 복잡도가 훨씬 증가하는데, 분석 과정에 후행 어미의 형태소 결합 제약을 일부 활용할 수 있다. 예를 들어 '지은, 그었다'에서 '지'는 동사 어간으로, '그'는 명사 어간으로 사전에 등재되어 있을 것이지만, 후행 어미 '은, 었'은 종성이 있는 동사 어간과만 결합한다는 제약을 갖는다. 따라서 선행 어간 '지'와 '그'가 어간의 일부가 탈락한 형태임을 어렵지 않게 예측할 수 있다.

12) 단일 형태소 어절 가운데에서도 단의 어절과 동음이의 어절(단순 동음이의 어절, 품사 중의성 어절)이 구분된다. 그러나 대체로 체언 어절에서 기술한 내용과 겹칠 것이므로, 이에 대한 기술은 생략하고 요약 부분에서만 이를 반영하기로 한다.

13) 앞에서 살펴보았듯이 자연언어처리 시스템의 개발에서 우선적으로 고려되어야 할 것은 분석의 정확성과 효율성이다. 형태소 분석의 경우 전자 사전과 밀접한 관련을 갖는데, 전자 사전의 표제어를 증감하는 것과 형태소 분석의 복잡도를 높이고 낮추는 것을 적절하게 조정함으로써 분석의 효율성을 높일 수 있다. 즉 규칙으로 처리할 수 있는 분석 대상 어절이 상대적으로 많은 경우에는 사전의 표제어를 줄이고 규칙을 추가하는 것이 더 효율적이지만, 규칙으로 처리할 수 있는 분석 대상 어절이 상대적으로 적은 경우에는 규칙을 줄이고 사전 표제어를 늘리는 것이 더 효율적이다. 그런데 불규칙 활용 형태의 경우에는 그 수가 그리 많지 않으므로, 사전 표제어로 등재하는 것이 더 효율적일 것이다.

(17ㄴ-③)의 음절 융합 형태 포함 어절은 융합 음절의 음절 정보를 활용함으로써 쉽게 형태소 분석을 할 수 있다.

또한 (15ㄴ)의 복합 형태소 어절은 중의항을 포함하는 어절과 그렇지 않은 어절로도 구분할 수 있는데, 중의항을 포함하는 어절의 경우에는 다시 어간이 중의항인 어절, 어미가 중의항인 어절, 어간과 어미가 모두 중의항이 어절로 나뉜다.

(19) 용언 어절의 구성 양상 4 : 복합 형태소 어절
ㄱ. 중의항 비포함 어절
철수가 의자에 앉아서 신문을 읽었다.
ㄴ. 중의항 포함 어절
① 어간 중의항 어절
영수가 밥을 지었다.
철수가 밥을 먹었다.
나는 설탕은 달아서 싫다.
모르는 것이 있어 그에게 물어 보았다.
② 어미 중의항 어절
그녀의 팔을 잡은 손이 조용히 떨렸다.
③ 어간-어미 중의항 어절
철수가 영희에게 이제 그만 자라고 했다.

어간 중의항 어절 가운데에서 '먹었다'의 경우에는 후행 어미 '었'이 동사와만 결합한다는 형태소 결합 제약을 가지며, '지었다'의 경우에는 '었'이 'ㅅ' 탈락 어간과 결합할 수 있다는 속성을 가지므로, 이를 토대로 형태소 분석을 옳게 할 수 있다. 그러나 '달아서'의 경우에는 어간으로 동사 '달(다)'과 형용사 '달(다)'이 모두 가능하며, '물어'의 경우에는 어간으로 '물(다)'과 '묻(다)'이 모두 가능하므로, 어절 내부에서는 옳은 분석을 하기 어렵다.

어미 중의항 어절인 '잡은'의 경우에는 '은'이 동사와 명사에 모두 결합할 수 있지만, 선행 어간 '잡'이 동사 어간이므로, 어렵지 않게 형태소 분석을 할 수 있다.

어간-어미 중의항 어절의 경우에는 가능한 분석 후보의 수가 많아 처리에 어려움이 있다. 예를 들어 '자라고'의 경우 예에서 제시한 것 이외에도 다음과 같은 분석이 모두 가능하다.

(20) '자라고'의 분석
ㄱ. 이것은 자라고 저것은 거북이다.
철수는 연필을 보고 자라고 한다.
ㄴ. 이 꽃은 잘 자라고 저 꽃은 잘 자라지 않는다.

즉 '자라고'는 '자+라고'로 분리될 경우 '자(동사)+라고(어미)', '자(명사)+라고(어미)'의 분석이 가능하며, '자라+고'로 분리될 경우에도 '자라(동사)+고(어미)', '자라(명사)+고(어미)'의 분석이 가능하다.

(15ㄷ)의 단일-복합 형태소 어절은 다음과 같이 종성이 'ㄹ'인 단일 형태소 어절 가운데에서 'ㄹ'이 어미로 분리될 수 있는 경우에 생성된다.

(21) 용언 어절의 구성 양상 5
ㄱ. 철수는 학교 갈 것이다.
농부는 밭을 갈 것이다.
갈 봄 여름 없이 꽃이 피네.
ㄴ. 그는 아직 잘 것이다.
이 나무는 뿌리가 잘 것 같다.
그는 노래를 잘 부른다.

위 예에서 알 수 있듯이 '갈'은 '가+을', '갈(동사)', '갈(명사)'의 분석이 가능하며, '잘'은 '자+을', '잘(동사)', '잘(부사)'의 분석이 가능하다.

이상에서 형태소 결합 양상에 따라 한국어 용언 어절의 구성 양상에 대해 살펴보았다. 이를 요약하면 다음과 같다.

(22) 용언 어절의 구성 양상
ㄱ. 단일 형태소 어절(=어미 생략 어절)
<생략 어미의 유형에 따라>
① 관형사형 어미 생략 어절
② 부사형 어미 생략 어절
③ 종결 어미 생략 어절
<중의항 포함 여부에 따라>
① 단의 어절
② 동음이의 어절(단순 동음이의 어절, 품사 중의성 어절)
ㄴ. 복합 형태소 어절
<불규칙 형태 포함 여부에 따라>
① 규칙 형태 포함 어절
② 불규칙 형태 포함 어절
-불규칙 활용 형태 포함 어절
-탈락 형태 포함 어절
-음절 융합 형태 포함 어절
<중의항 포함 여부에 따라>
① 중의항 비포함 어절
② 중의항 포함 어절
-어간 중의항 어절
-어미 중의항 어절
-어간-어미 중의항 어절
ㄷ. 단일-복합 형태소 어절

5. 결론

한국어는 언어 유형론적으로 교착어로서 어간에 어미가 결합하여 어절을 구성하는데, 이 어절이 형태소 분석의 대상이 된다. 따라서 정확하고 효율적인 형태소 분석기를 설계하고 구현하기 위해서는 우선적으로 한국어 어절의 구성 양상을 살펴보는 것이 필요하다. 이러한 연구의 필요성에 따라 본 연구에서는 형태소 결합 양상을 토대로 체언 어절과 용언 어절의 구성 양상에 대해 살펴보았다.

체언 어절은 먼저 형태소 결합 유무에 따라 단일 형태소 어절, 복합 형태소 어절, 단일-복합 형태소 어절의 세 가지 유형으로 구분할 수 있다. 단일 형태소 어절은 다시 조사 생략 어절(단의 어절, 동음이의 어절)과 어근적 단어 어절로 구분된다. 복합 형태소 어절은 후행 요소의 형태 범주에 따라 다시 조사 결합 어절(중의항 비포함 어절, 중의항 포함 어절), 동사화 접사 결합 어절(어간 비중의항 어절, 어간 중의항 어절), 어미 결합 어절(어간 비중의항 어절, 어간 중의항 어절)로 나뉜다. 그리고 단일-복합 형태소 어절은 복합 형태소일 경우 끝 형태소의 형태 범주에 따라 끝 형태소 조사 어절, 끝 형태소 어미 어절, 끝 형태소 조사-어미 어절로 구분된다.

용언 어절도 먼저 형태소 결합 유무에 따라 단일 형태소 어절, 복합 형태소 어절, 단일-복합 형태소 어절의 세 가지 유형으로 구분할 수 있다. 단일 형태소 어절은 특정 환경에서 어미가 생략된 것으로서 생략 어미의 범주에 따라 관형사형 어미 생략 어절, 부사형 어미 생략 어절, 종결 어미 생략 어절로 구분된다. 복합 형태소 어절은 먼저 불규칙 형태의 포함 여부에 따라 규칙 형태 포함 어절, 불규칙 형태 포함 어절(불규칙 활용 형태 포함 어절, 탈락 형태 포함 어절, 음절 융합 형태 포함 어절)로 구분된다. 또한 복합 형태소 어절은 중의항 포함 여부에 따라 중의항 비포함 어절과 중의항 포함 어절(어간 중의항 어절, 어미 중의항 어절, 어간-어미 중의항 어절)로 나뉜다. 그리고 단일-복합 형태소 어절은 종성이 'ㄹ'인 단일 형태소 어절 가운데에서 'ㄹ'이 어미로 분리될 수 있는 경우에 생성된다.

6. 참고 문헌

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

김영택(1994). 「자연 언어 처리」 교학사.

김창섭(1994/1996). 「국어의 단어형성과 단어구조 연구」 태학사.

김홍규·강범모(1997). 「한글 사용빈도의 분석」 고려대 민족문화연구소.

남기심·고영근(1998). 「표준국어문법론(개정판)」 탑출판사.

시정곤(1994). 「국어의 단어형성 원리」 국학자료원.

이기오·이근용·이용석(1996). "효율적인 한국어 분석을 위한 확장된 최장일치법," 「제8회 한글 및 한국어 정보처리 학술대회 발표 논문집」, 255-261

이익섭(1965). "국어 복합명사의 IC 분석," 「국어국문학」 30, 121-129.

허웅(1966/1975). 「우리 옛말본」 샘문화사.

황화상(1996). "국어 체언서술어의 연구," 고려대 석사학위논문.

황화상(1998). "자연언어처리를 위한 형태소 분석 방법론," 「어문논집」(고려대) 37, 439-458.

황화상(2001). 「국어 형태 단위의 의미와 단어 형성」 월인.