형태 분석 말뭉치 구축을 위한 합성어의 처리 방법
- 띄어쓰기를 고려하여 -

조진현 김일환 이현희 이영제 강범모
고려대학교 민족문화연구원 전자텍스트연구소
{belus, haigh}@ikc.korea.ac.kr, hhlee311@hanmail.net,
kukl@netian.com, bmkang@korea.ac.kr

Dealing with Compouds in the Construction of a POS Tagged Korean Corpus

Jin Hyun Cho, IlHwan Kim, HyunHee Lee, YoungJe Lee, Beom-Mo Kang
Center for Electronic Texts, Korea University

요 약

이 연구는 형태 분석 정보가 부착된 말뭉치를 구축할 때 합성어를 처리하기 위한 방법론을 제시하고, 그 타당성을 검증해 보는 데 있다. 그동안 합성어 처리를 위해서 합성어 선정 기준을 이용하거나 목록을 이용하는 방법이 이용되었는데, 본고에서는 「표준국어대사전」의 합성어 목록을 참조하는 것이 적절한 방법이 될 수 있음을 보이고자 한다. 또한 이 방법을 실제 말뭉치 구축에 활용할 경우, 원문의 띄어쓰기 정보가 합성어 처리에서 중요한 요인이 될 수 있다는 점을 지적하고, 이러한 처리가 가지는 한계와 의의에 대해서도 논의하고자 한다.

1. 서론

이 연구는 형태 분석 말뭉치를 구축할 때 합성어를 처리하기 위한 방법론을 개발하고, 그 타당성을 검증하는 데 목적이 있다.

합성어에 대한 논의는 이미 국어학에서 풍부하게 다루어진 주제이다. 그럼에도 불구하고 형태 분석 말뭉치 구축과 같은 자연언어처리 분야에서 합성어가 다시 쟁점이 되는 데에는 몇 가지 이유가 있다. 첫째, 합성어와 구와의 구별이 여전히 애매하다는 점이다. 특히 국어학에서 논의된 합성어 판별 기준만으로는 국어의 다양한 유형의 합성어를 모두 포괄하기 어렵다. 둘째, 합성어 목록을 설정하기 어렵다는 점을 들 수 있다. 합성어의 기준에 대한 정의가 학자들마다 조금씩 다를 뿐 아니라, 합성어는 현재에도 계속 생산적으로 형성되기 때문에 이들의 확정된 목록을 확보하는 것은 단순한 일이 아니다.

2. 선행 연구 검토와 문제 제기

지금까지 구축된 국어의 tagged corpus로는 다음의 세 가지를 꼽을 수 있다. 이들에서는 합성어가 어떻게 처리되었는지 먼저 살펴보자.

2.1. KAIST tagged corpus (1997)

합성어를 분리해서 분석하였다. 즉, 합성어의 내부까지 분석하는 입장인데, 이때 분석의 한계가 문제가 될 수 있다.

(1) 가. 요한복음 요한/nq+복음/ncn
요한계시록 요한계시록/nq
나. 호흡곤란 호흡곤란/ncn
문명사회 문명/ncn+사회/ncn
다. 구조활동 구조활동/ncp

필요조건　　필요/ncps+조건/ncn

즉, (1가)에서는 유사한 어절을 분리해서 분석하거나 합성어로 통합해서 분석하는 비일관성이 나타나며, (1나, 다)에서처럼 합성어와 구를 처리하는 방식이 명확하지 않은 경우도 있다.

2.2. ETRI tagged corpus (1999)

표준국어대사전의 표제어 선정 기준을 이용하여 합성어를 처리하였다(한국전자통신연구원 1999). 즉, 표준국어대사전의 합성어 판정 기준을 이용한 것인데, 실제로 표준국어대사전의 합성어 목록을 이용한 것인지 아니면 단순히 판정 기준만을 이용한 것인지는 확인하기 어렵다. 단, 실제 형태 분석된 말뭉치를 검토해 본 결과 합성어 처리에서 일관성에 오류가 있다는 점이 확인되었다.

(2) 가. 살펴보다 - 살펴보/pv, 살피/pv+어/ec+보/px
　　옮겨지다 - 옮겨지/pv, 옮기/pv+어/ec+지/px
　　돌아보다 - 돌아보/pv, 돌/pc+아/ec+보/px

　　나. 말뜻 - 말뜻/nc, 말/nc+뜻/nc
　　막내딸 - 막내딸/nc, 막내/nc+딸/nc
　　신문기자 - 신문기자/nc, 신문/nc+기자/nc

2.3. 세종 형태 분석 말뭉치 (1999-2001)

21세기 세종계획에서 구축한 형태 분석 말뭉치의 합성어 처리의 기준을 다음과 같이 제시되어 있다(김흥규·강범모, 2000).

(3) 가. 합성명사: 세분해서 분석했을 때 전체 의미를 합성적으로 나타낼 수 없는 경우만 통합형으로 분석
　　예) 개미핥기, 책상다리, 큰아버지 등
　　나. 합성용언: 어느 한 쪽이 화석화되어 의미 합성성이 준수되지 않고 공시적으로 더 이상 세분하여 분석하기 어려운 경우만 통합형으로 분석
　　예) 따라잡다, 꺼내다, 다가오다, 떨어지다 등

또한 위의 기준과 별도로 합성어 목록을 활용하였는데, 그 목록은 국어학적으로 널리 용인되는 고유어 합성어에 국한하였다. 이러한 기준과 목록에 의한 처리에도 불구하고 세종 형태 분석 말뭉치에서도 여러 가지 일관성의 오류가 발견되었다.

위의 세 형태 분석 말뭉치를 검토한 결과, 합성어를 일관성 있고, 적절하게 처리하는 것이 매우 어려운 일임을 확인할 수 있다. 합성어의 내부를 분리해서 분석할 경우에는 분리하기 용이한 것들만 분석하는 오류를 범하기 쉽고, 통합할 경우에는 통합의 기준을 명확히 설정하기 어렵다는 문제가 있다. 이들이 야기하는 문제는 결국 분석 말뭉치의 일관성에 영향을 끼치게 되는데, 본 연구에서는 합성어 목록을 설정하고 그에 따라 합성어를 처리하는 방법이 대량으로 말뭉치를 구축하는데 있어 보다 합리적이라는 입장을 채택하였다. 이를 위해 이미 구축된 기존의 종이사전으로부터 합성어 목록을 추출하고 이를 엄밀히 적용하는 방식을 적용하였다.

또한 본 연구에서는 특히 '띄어쓰기'를 중시하고자 한다. 일단 합성어는 띄어쓰지 않는 것을 원칙으로 한다는 가정을 수용하고 실제 말뭉치에서 사전 표제어로 등재된 합성어가 사용되는 양상을 실험을 통해 살펴봄으로써 합성어 처리의 단서를 제시하고자 한다.

3. 종이사전을 이용한 합성어 처리 방법

3.1. 합성어 처리를 위한 원칙과 환경

본 연구에서 사용한 종이사전은 국립국어연구원에서 2000년에 출판한 「표준국어대사전」이다. 표제어로부터 추출된 목록에 의해 합성어를 처리하는 원칙을 다음과 같이 설정하였다.

(4) 합성어 처리 원칙

자동 태깅 결과로 추출된 목록과 「표준국어대사전」의 표제어 목록을 비교하여 표준국어대사전의 단일 표제어로 등재된 합성어는 통합형으로 처리한다. 단, 등재되어 있지 않은 것은 분석하여 처리한다.

이와 같은 원칙에 의해 실제 말뭉치를 대상으로 실험을 실시하였다. 실험을 위해 「표준국어대사전」로부터 추출된 총 152,480개의 파생어가 포함된 복합어[1] 목록을 이용하여 복합어 사전을 만들었다.[2] 또한 이 사전에는 53,292개의 띄어쓰기가 허용된 합성어 목록이 따로 제시되어 있는데, 이는 실제 생활에서 '붙여쓰되 띄어쓰는 것도 가능한 합성어'를 따로 표시한 것이다. 이러한 유형을 본 연구에서는 '준합성어'라고 부르기로 하고 이것도 역시 준합성어 사전을 따로 만들어 실험에 이용하였다.[3]

실험 대상 말뭉치는 1999년부터 2001년까지 구축된 세종 형태 분석 말뭉치로, 전체 어절은 총 5,576,034 어절이다. 이 형태 분석 말뭉치로부터 한 어절 내에서 연속되는 태그열을 선정하여 합성어 처리 대상 후보 어절을 결정하였는데, 이는 연속되는 형태소를 결합할 경우 합성어가 될 수 있는 형태소 결합을 말하는 것이다. 이에 따라 명사류 154,168어절, 동사류(형용사 포함) 136,829어절이 선정되었다. 다음은 선정된 형태소 결합 유형을 나타낸 것이다.[4]

(5) 가. 명사류(154,168어절)

명사+명사, 어기+명사(+어기), 관형사+명사, 동사(형용사)+관형형 어미(ㄴ/ㄹ)+명사

나. 동사류(136,829어절)

동사+연결어미+동사, 동사+연결어미+보조용언, 형용사+연결어미+형용사, 형용사+연결어미+보조용언, 명사+동사

위와 같은 형태소 결합을 지닌 어절은 모두 실험 후보가 되는데, 3개 이상의 형태소 결합을 허용한다. 따라서 명사류에서 '명사+명사'의 경우 '국제(명사)+관계(명사)'뿐 아니라 '국제+관계+개선', '국제+관계+개선+도모' 등의 결합도 처리 대상으로 하였다.

3.2 합성어 처리를 위한 실험과 해석

3.2.1 합성어 처리를 위한 실험 (I)

첫 번째로 세종 형태 분석 말뭉치에서 분리하여 태깅한 단어 중 「표준국어대사전」에 붙여쓰는 합성어로 등재된 단어의 비율을 조사한 결과, 53,385개의 단어가 추출되었다. 이 중 명사류는 총 4,628개의 타입에 36,921개의 토큰이었으며, 동사류의 경우 779개 타입에 16,464개 토큰이었다. 다음은 실제 합성어로 처리될 수 있는 단어를 나타낸 것이다.

(6)

그+때	그때	2707
이+날	이날	1476
지난+해	지난해	965
그+날	그날	953
올+해	올해	870
이+때	이때	836
한+마디	한마디	562
오늘+날	오늘날	515
다음+날	다음날	423
눈+앞	눈앞	398
불+빛	불빛	291
눈+빛	눈빛	261
전+날	전날	255
대중+문화	대중문화	247
마음+속	마음속	247
지난+달	지난달	236
몸+짓	몸짓	191
그+해	그해	160
밥+상	밥상	140
가슴+속	가슴속	131
지난+날	지난날	131
온+몸	온몸	124

1) 복합어 목록은 국립국어연구원으로부터 직접 전달받은 것이다. 연구에 도움을 준 국립국어연구원측에 감사한다. 이 목록에는 파생어가 포함되어 있어 합성어와 파생어가 동일하게 '-'표시를 이용하여 맨 마지막으로 이루어진 합성과 파생을 나타내고 있다.
2) 접사 목록을 통해 합성어와 파생어를 구분하고자 했으나 처리 방식이 사전을 '단어' 단위로 참조하는 방식이기 때문에 굳이 둘을 구분할 필요가 없을 것으로 판단되어 그대로 사전화하였다.
3) 준합성어 목록은 복합어와는 다르게 '^'로 표시되어 있다.
4) 이러한 결합 유형은 본 연구를 위해 임시로 선정한 것이다. 따라서 필요에 의해 추가되거나 삭제될 수 있다.

앞+뒤	앞뒤	123
한+잔	한잔	123
어리+ㄴ+애	어린애	122
뜻+밖	뜻밖	121
술+잔	술잔	119
달+빛	달빛	118
꿈+속	꿈속	114
사고+방식	사고방식	113

이와 같은 단어들은 수작업을 거치지 않아도 합성어로 자동으로 처리할 수 있다. 단 동사의 경우 음절 융합이 발생되므로 합성어로 통합시킬 때 별도의 처리가 필요하다.

두 번째는 사전에 띄어쓰기가 허용된 준합성어의 처리 문제인데, 준합성어의 사용 양상을 살펴보기 위해 한 어절 내에서와 두 어절에서의 사용 양상을 조사하였다. 즉, '환경+문제'라는 형태소 결합열이 실제 말뭉치에서 한 어절 내에서 사용된 경우(예: 환경문제가)와 띄어쓰기가 된 경우(예: 환경 문제가)를 추출하여 결합 여부를 살펴보았다. 이에 따라 나타난 형태소 결합은 다음과 같다.

「표준국어대사전」에 등재된 준합성어	1 어절	2 어절
환경문제	37	164
정상회담	65	136
경제성장	29	129
지방자치	126	129
자본주의사회	33	123
선거운동	78	121
금융기관	30	113
문학작품	60	109
지방의회	16	101
세계경제	27	93

[표-1] 준합성어의 사용 양상 (상위 빈도 10개)

위의 표에서 준합성어의 경우 대부분이 실제 말뭉치에서 붙여쓰는 것보다는 띄어쓰는 것이 더 높은 빈도를 나타내고 있음을 알 수 있다. 이러한 결과는 한 어절 내에서 사용된 준합성어는 형태소별로 분리하여 분석하는 것이 타당함을 시사한다. 그리고 이러한 분석 방법이 어절을 기본 단위로 하는 태깅 방식에서는 합성어를 일관성있게 처리할 수 있는 방법이 될 것이다.

3.2.2 합성어 처리를 위한 실험 (II)

다음은 말뭉치에서 띄어쓰기가 된 형태소 결합열을 표준국어대사전」의 합성어, 준합성어 목록과 비교하였는데[5], 이에 대한 결과를 나타내면 다음과 같다.

(7) 가. 띄어쓴 형태소 결합열 중 「표준국어대사전」에 합성어로 등재된 단어
- 명사류 : 타입 3,171개, 토큰 13,595개
- 동사류 : 타입 3,390개, 토큰 6,390개

나. 띄어쓴 형태소 결합열 중 「표준국어대사전」에 준합성어로 등재된 단어
- 명사류 : 타입 5,184개, 토큰 24,114개
- 동사류 : 없음

우선 (7나)의 경우 이미 사전에서부터 띄어쓰기를 허용하고 있으므로 어절이 분리된 채로 분석을 시행하는 데 문제가 되지 않음을 이미 3.2.1절에서 언급하였다. (7가)의 경우는 합성어 처리에 있어 문제가 되는 부분인데, 이를 자세히 살펴보도록 하자. 다음은 띄어쓰기가 되어 있는 형태소 결합열을 합성어 목록과 비교한 결과를 빈도순으로 나타낸 표이다.

「표준국어대사전」에 등재된 합성어 (명사류)	1 어절	2 어절
중소기업	225	214
마음속	264	189
창밖	87	165
국회의원	282	156
물속	73	154
환경오염	48	124
일상생활	119	102
고속도로	311	100
이해관계	79	91
사고방식	114	90

[표-2] 합성어의 사용 양상 - 명사류 (상위 빈도 10개)

5) 어절 경계를 넘어선 형태소 결합열은 일단 두 어절만을 대상으로 하였고 명사류의 경우 결합유형을 일단 '명사+명사', '어기+명사', '형용사+명사' 등 3개로 한정하였다. 동사류의 경우 연결어미로 끝나는 어절 다음에 동사나 형용사, 보조용언으로 시작되는 어절을 대상으로 하여 추출하였다.

「표준국어대사전」에 등재된 합성어 (동사류)	1 어절	2 어절
들려오다	363	149
놀고먹다	5	80
찾아보다	485	116
먹고살다	35	192
물어보다	276	198
살아오다	377	88
받아들이다	1128	119
돌아오다	2269	80
사들이다	111	37
살아가다	804	35

[표-3] 합성어의 사용 양상 - 동사류 (상위 빈도 10개)

위의 표에서, 실제로 띄어쓰는 것보다는 붙여쓰는 경향이 많은 단어들-'마음속, 국회의원, 고속도로' 등-은 그대로 합성어로 인정하되 띄어쓰는 경우에는 띄어쓴 그대로 두 어절로 분석하는 것이 바람직할 것이다. 단, 앞에서 언급한 '개미핥기, 책상다리, 큰아버지', '떨어지다, 헤어지다'와 같이 떨어뜨려 사용하는 빈도가 거의 낮으면서 분석하면 의미적인 합성성이 준수되지 않거나 공시적인 해석이 어려운 것은 원어절을 수정하여 띄어쓰기를 제거한 후 합성어로 처리해도 무방하다.

그러나 띄어쓰는 경향이 많은데 사전에 합성어로 등재된 단어들의 처리는 문제가 된다. 일단 본 연구에서는 합성어를 띄어쓰지 않은 단위로 인정한다는 가정을 하였고 실제로도 띄어쓰지 않고 붙여쓰는 경향이 높은 단어들이 합성어가 될 확률이 높다. 그런데 위의 표에서 알 수 있듯이 명사류의 경우 「표준국어대사전」에서 합성어로 등재된 '창밖', '물속', '환경오염', '이해관계' 등은 띄어쓰는 빈도가 높게 나타났고, 동사류의 경우에도 '놀고먹다', '먹고살다'[6] 등이 사전 기술과는 다르게 띄어쓰는 경향이 높은 것으로 조사되었다[7]. 또한 붙여쓰는 경우보다 많지는 않지만 '중소기업', '일상생활' 등도 거의 대등하게 띄어쓰는 경향을 보인다.[8]

이와 같이 사전적으로는 합성어인데 띄어쓴 경우에는 다음과 같이 네 가지 방식 중에서 선택하여 처리할 수 있을 것이다.

(8) 가. 단일어절만 합성어로 분석하고 어절이 분리된 경우는 고려 안 함
나. 사전 자체의 합성어 처리에 문제가 있는지 고려하고 문제가 있는 경우에는 사전의 합성어 목록에서 삭제한 후 분리하여 처리
다. 다중어절을 하나의 합성어로 묶어 표기하는 방안
라. 원어절의 띄어쓰기를 수정한 후 합성어로 처리

이 중에서 (8다)와 (8라)는 형태 분석 말뭉치를 구축하는 방법론과 연관되는 것으로 신중을 기해야 한다. (8다)의 경우 어절 단위의 태깅 원칙을 지키는 형태 분석 말뭉치에는 적용하기가 어렵다. 왜냐하면 어절 경계를 넘어서 태깅하는 문제이므로 새로운 태그셋의 설정이 필요하고, 합성어 처리 뿐 아니라 어절 경계를 넘어서는 다른 문제들에 대한 새로운 논의가 이루어져야 하기 때문에 말뭉치를 처음부터 새로이 구축하는 것과 다를 바가 없기 때문이다. 또한 형태 분석 말뭉치를 (8라)와 같은 방식으로 수정할 경우, 원시 말뭉치와의 호환성에 문제가 생기므로 원시 말뭉치도 같은 방식으로 수정해 주어야 한다. 그러나 이러한 방식의 처리는 자연스런 언어 현상의 반영을 훼손하고 그에 대한 가치를 포기함으로써 말뭉치를 통해 실제적인 언어의 사용 양상을 살피고자 하는 말뭉치 구축의 목적을 왜곡할 우려가 있다.

6) 동사류에서 '놀고먹다'와 '먹고살다'의 경우 모두 연결어미 '-고'가 사용된 단어이다. 연결어미 '-고'가 주로 대등적 연결어미로 많이 사용되어 띄어쓰는 경향이 높은 어미이기 때문에 이와 같은 영향으로 합성어도 띄어쓴다는 해석도 가능하다.

7) 붙여써야만 하는 합성어를 띄어쓰는 경향은 동사류보다는 명사류에서 특히 많이 나타나는데, 이는 동사나 형용사가 명사에 비해 합성어를 만들 수 있는 가짓수가 많지 않고, '흘러가-'의 경우처럼 내부적으로 '흐르+어+가-'와 같이 분리해서 인식하여야만 하기 때문에 복잡도가 증가하므로 통합하여 인식하는 경향을 보이는 것으로 추정된다.

8) '돌아오다'의 경우 의미적 중의성이 있기 때문에 이것이 띄어쓰기에 영향을 미쳤을 가능성도 존재한다. 그러나 이러한 문제에 대한 엄밀한 검토가 필요하지만 본 논문에서는 이에 대한 논의는 일단 보류하기로 한다.

따라서 본 연구에서는 (8다)와 (8라)의 방식에 대한 고려는 배제하고 (8가)와 (8나)를 이용하여 두 단계에 걸쳐 합성어를 처리하는 방법을 제안하고자 한다. 이에 따라 3.1절에서 제시된 합성 처리 원칙을 다음과 같이 수정한다.

(9) 수정된 합성어 처리 원칙

가. 자동 태깅 결과로 추출된 목록과 「표준국어대사전」의 표제어 목록을 비교하여 「표준국어대사전」의 단일 표제어로 등재된 합성어는 통합형으로 처리한다.

ㄱ. 단일 어절 내의 합성어만을 대상으로 처리하고 띄어쓴 합성어는 있는 그대로 분리하여 태깅한다.

ㄴ. 어절이 분리된 합성어의 경우, 다양한 통계적 방식을 통해 띄어쓰는 예가 많고 의미적으로도 의미 합성성에 영향을 주지 않는 단어들은 합성어 목록에서 삭제한 후 분리하여 태깅한다.

나. 등재되어 있지 않은 것은 분석하여 처리한다. 단, 조사를 통하여 붙여쓰는 예가 훨씬 많은 표현들(준합성어 포함)은 새로운 합성어의 후보로 고려하고, 경우에 따라 합성어 목록에 새롭게 추가한다.

이와 같은 원칙에 의해 '창밖', '물속', '환경오염', '놀고먹다' 등은 단일 어절 내에서 사용되었다 하더라도 분리해서 '창-밖', '물-속', '환경-오염', '놀-고-먹-다'로 분리하여 분석하게 된다. 또한 띄어쓰는 빈도가 높은 '중소기업', '일상생활' 등도 면밀한 검토를 거쳐 분리하여 분석할 수 있는 후보가 된다.

4. 결론 및 남은 문제

지금까지 기존의 합성어 처리 방법에 대한 고찰과 아울러 「표준국어대사전」의 합성어 목록을 채택하여 합성어 처리에 엄밀히 적용하는 방식에 대해 살펴보았다.

기존 형태 분석 말뭉치에 대한 합성어 처리의 고찰을 통해 'KAIST 형태 분석 말뭉치'는 분석 한계 설정의 문제가 있었고, 'ETRI 형태 분석 말뭉치'는 표제어 선정 기준을 제시하였지만 사전 참조에 대한 정보가 미약하고 합성어 처리에 대한 일관성에 문제가 있음을 제시하였다. 또한 세종 말뭉치의 경우에도 합성어 처리 원칙을 도입하였지만 원칙 자체가 구체적이지 못해 엄밀한 합성어 처리에 있어 부분적인 문제점이 있음을 지적을 지적하였다.

이와 같은 말뭉치 분석으로 합성어의 분석을 분석주의, 또는 통합주의 등 어느 한 쪽의 기준만을 채택할 경우 분석의 기준과 한계를 객관적이고 명시적으로 설정하기가 어려움을 확인할 수 있었다. 이와는 다르게 목록을 통한 처리 방법은 객관적이고 포괄적인 합성어 목록을 확보하는 것이 관건이지만, 일단 목록이 확보되면 보다 안정적인 합성어 처리가 가능하다는 장점이 있다. 본 연구에서는 합성어 목록을 확보하는 일차적인 방법으로 「표준국어대사전」의 합성어 목록을 참조하여 띄어쓰기를 고려한 합성어 사용 양상을 살펴보았다. 또한 이를 통해 새로운 합성어 처리 원칙을 제안하고자 하였다.

약 550만 어절의 세종 형태 분석 말뭉치로부터 형태소 결합열들을 추출하여 「표준국어대사전」의 합성어, 준합성어 목록과 비교한 결과 합성어 목록에 존재하는 단일 어절 내의 결합열들은 그대로 합성어로 인정하고, 준합성어 목록에 존재하는 결합열들은 분리하여 분석하는데 문제가 없었다. 다만 두 어절로 나누어진 형태소 결합열 중 사전의 합성어 목록에 등재되어 있는 단어들이 존재하였는데, 이와 같은 단어들에 대한 처리를 위해 두 단계로 처리하는 방안을 제시하였다.

즉 어절이 분리된 채로 그대로 태깅을 수행하고 다음 단계에서 이러한 유형의 단어들에 대한 목록을 추출하여 비교한다. 이렇게 함으로써 사전에 잘못 등재되어 있다고 판단된 단어는 합성어 목록에서 삭제하고 이것이 한 어절 내에서 사용되었다 하더라도 분리하여 태깅한다. 만일 띄어쓰는 경향이 많은데 합성성 등의 관점에서 합성어가 분명한 경우에는 한 어절 내에서는 통합하여

분석하고, 띄어쓰는 쓰는 경우에는 있는 그대로 분석하는 방식이다.

이와 같은 작업을 위해서 지속적인 사전의 합성어 목록에 대한 검증 작업이 필수적으로 요구된다. 왜냐하면 「표준국어대사전」의 경우 제시된 합성어와 준합성어의 구분기준이 명확하지 않기 때문에 어떠한 경우에 띄어쓰는 것을 허용하는가에 대한 정확한 근거가 마련되어야 띄어쓰기 정보에 의한 처리가 보다 객관적이고 합리적일 수 있다.[9] 또한 합성어를 구성하는 형태소들간의 계열 관계를 검토하여 합성어를 구성하는 빈도가 높은 형태소를 추출하고 이를 목록화하는 것도 합성어 처리에 있어 큰 도움이 될 것이다.

합성어에 대한 처리는 그 논의가 방대하고 다양해서 쉽게 처리될 수 있는 성질의 것이 아니다. 이에 대한 처리를 위해서는 우선 '단어'가 무엇인가에 대한 명확한 정의와 해석이 내려져 있어야 하는데 이 역시도 많은 이견이 있어 명확한 결론을 내지 못하고 있는 실정이다. 기존의 국어학이나 언어학에서 언급된 합성어에 대한 정의와 본 연구의 띄어쓰기를 고려한 합성어 처리는 다소 차이가 있을 수 있지만 합성어를 자동적, 계량적으로 처리하기 위한 실마리를 제공했다는 의의가 있다. 이상의 논의를 바탕으로 앞으로도 더욱 많은 자료에 대한 지속적인 실험과 검증을 통해 보다 안정되고 효율적인 합성어 분석이 이루어질 수 있기를 기대한다.

9) 「표준국어대사전」의 합성어 목록과 준합성어 목록을 살펴보면 다음과 같이 동일한 구조와 의미적 관계임에도 불구하고, 합성어 처리를 다르게 한 예가 발견된다.
국제결혼 vs 국제^ 경기, 대중문화 vs 대중^ 문학

[참고문헌]

[1]. 강승식(1998), "한국어 복합명사 분해 알고리즘", 정보과학회논문지(B), 25권 1호.
[2]. 국립국어연구원 편(2001), 표준국어대사전, 두산동아.
[3]. 김규선(1970), "국어의 복합어에 대한 연구", 어문학 23, 한국어문학회, 93-123.
[4]. 김일환(2002), "표지 부착된 말뭉치 구축에서의 합성어 처리 방법", 2002 한국어학회 국제학술대회 발표요지.
[5]. 김일병(2000), 국어 합성어 연구, 도서출판 역락.
[6]. 김흥규 외(2000), 「21세기 세종계획 국어 기초자료 구축 연구보고서」, 문화관광부.
[7]. 김흥규 외(2001) 「21세기 세종계획 국어 기초자료 구축 연구보고서」, 문화관광부.
[8]. 김흥규, 강범모(2000), 「한국어 형태소 및 어휘 사용 빈도의 분석 1」, 고려대 민족문화연구원.
[9]. 서정수(1981), "합성어에 관한 문제", 한글 173·174.
[10]. 조진현(2001), "형태소 분석 말뭉치 구축의 한 방법", 제13회 한글 및 한국어 정보처리 학술대회 발표요지.
[11]. 지식정보연구부(1999), 「품사 태그 부착 말뭉치 구축 지침서」, 한국전자통신연구원.
[12]. 한영균(1996), "전산기에 의한 형태분석과 사전정보", 국어학 27.
[13]. 한영균(1998), "문어 코퍼스의 형태 정보 주석에서 선결되어야 할 몇 문제", 한국어 전산학 2.
[14]. 홍종선(2001), "국어 말모둠의 문법 표지와 전처리", 계량언어학 1.
[15]. 황화상·시정곤(2001), "형태소 분석을 위한 한국어 어절의 구성 양상 연구", 제13회 한글 및 한국어 정보처리 학술대회 발표요지.