

의미역 태깅의 제문제 고찰

- 구축된 격틀사전의 문제점과 대상말뭉치의 문제점을 대상으로

김윤정* †
울산대학교

옥철영‡
울산대학교

Yun-Jeong Kim. 2016. Consideration of various problems of Sematic Roles Tagging. *Language and Information* 20.2, 01-26. This paper deals with two problems occurring in the process of labelling semantic role to Sejong Corpus. The first problem is some discrepancy among a case-frame dictionary used for labelling semantic role, a constructed case-frame dictionary and Sejong corpus. The typical examples of discrepancy between a case-frame dictionary and Sejong corpus are 'appearance of compound words' and semantic role tagging to the sentence with syntactic auxiliary verb constructions. To solve this problems, an extended case-frame dictionary should be constructed with supplement and modification of annotated semantic roles. The other problem is a verb problem not solvable with the semantic tagging program, 'UTagger-SR'. Since 'UTagger-SR' generates and presents verbs as their original forms of presented predicates, compound words with space or syntactic auxiliary verbs are not taken into consideration. The worker tagging semantic roles should tag semantic roles, treating a compound word with spacing as one word, and a sentence with an auxiliary verb construction as one verb consisting of a non-auxiliary verb and an auxiliary verb. Then, in the process of generating results, post-processing is needed. The post-processing includes treating a verb belonging to compound words as a compound word regardless of spacing between letters and combining the auxiliary verb and the non-auxiliary verb together in case of an auxiliary verb construction. (Ulsan University)

keyword: semantic role labelling, The case-frame dictionary, argument, Sejong Corpus, space, auxiliary verb construction, 'UTagger-SR', semiautomatic semantic role labelling.

* 귀중한 지적과 논평으로 심사를 위해 애써주신 두 분의 심사자께 감사를 드린다.

† 울산광역시 남구 대학로 93 울산대학교 국어국문학과 44610
E-mail: jungi0006@ulsan.ac.kr

‡ 울산광역시 남구 대학로 93 울산대학교 전기공학부 IT융합전공 44610
E-mail: okcy@ulsan.ac.kr

1. 서론

본고는 한국어 의미역 주석 말뭉치를 구축하는 데에 있어 고려해야 할 격틀 사전의 미흡한 정보량과 실제 주석 말뭉치의 비문법적인 문제점을 지적하고 이를 해결하기 위한 방안을 모색하는 데에 목적을 두고 있다. 김윤정 외(2014)¹를 통해 22개의 의미역을 상정하고, 이를 기반으로 세종균형말뭉치를 대상으로 의미역 태깅 작업을 시행한 것에 대해 논의한 바 있다. 그 후 의미역을 상정하고 의미역을 태깅하기 위한 프로그램의 반자동화를 위해 의미역이 주석된 사전(이후 격틀사전이라고 함)을 구축하여 탑재하였다. 격틀사전은 의미역 주석 말뭉치 구축에 있어서 작업자의 직관에만 의존하지 않고 객관적인 기준에 의한 태깅 작업이 이루어지도록 하기 위해 사전에 등재된 용언의 제시 문형을 토대로 의미역을 주석한 사전이다. 의미역태깅프로그램²에는 격조사별로 의미역이 주석된 조사별 의미역 주석 결과물이 탑재되어 있어서 격조사에 관련된 의미역을 제시해 주고 있었다. 그러나 격조사별 의미역은 일대일의 관계에 있는 것보다 일대 다의 관계에 있는 결과물이 많아 작업자의 직관에 의존하기는 마찬가지였다. 즉, 격조사만으로는 서술어의 필수논항이 명시적으로 처리되지 않는다는 점이 의미역을 반자동으로 태깅하는 데에 걸림돌이 되었다는 것이다. 그리하여 논항의 의미역 정보를 좀 더 명시적으로 제시하기 위한 방안으로 격틀사전을 구축하게 되었다. 이를 “UTagger-SR”에 탑재하였고, 현재 의미역 태깅 프로그램은 반자동으로 구동되고 있다. 반자동 “UTagger-SR”로 기존의 작업자의 1,2차 의미역 주석말뭉치 구축분과 일치율을 비교한 결과 약 70%의 일치율을 보였다³. 일치하지 않는 나머지 30%에 대한 결과물을 토대로 그 원인을 분석해 보았다. 크게 기계적 오류와 비기계적 오류로 나누어 볼 수 있다. 기계적 오류는 프로그램상의 오류를 찾아 수정하여 문제를 해결하였다⁴. 비기계적 오류는 말뭉치 자체의 문제와 구축된 격틀사전의 문제로 나뉘볼 수 있다. 말뭉치의 문제는 문장의 처리 기준이나 의미역주석을 위한 후처리의 필요성을 인식하게 했고, 격틀사전의 문제는 사전의 한정된 격틀사전만으로는 다양한 문장의 경우의 수를 모두 대비할 수 없다는 점을 인식하게 했다. 본고에서는 이 두 가지를 살펴보고 문제를 줄일 수 있는 방안을 모색해보겠다.

우선 격틀사전의 구축에 대한 과정을 간략히 설명하고자 한다.

격틀사전의 구축대상 사전은 국립국어원편 표준국어대사전⁵이다. 의미역을 주석하기 위한 첫째 기준이 서술어이므로 서술어의 위치에 올 수 있는 용언류를 《표준》에서 모두 추출하였다. 이 중에서 북한어, 고어, 방언 등을 대상에서 배제하고 나머지 용언류를 다의단계까지 세분화하여 자료화하였다. 이때 구축된 다의단계까지의 용언의 수는 9만 개정도가 된다.

이렇게 다의단계까지 용언의 정보를 세분화한 결과에 의미역을 주석하는 것은 개별의미

1 김윤정·김완수·육철영(2014), <전산언어학에서의 한국어 필수논항의 의미역 상정과 재고>, 언어와 정보 제18권 제2호, pp.169-199.

2 의미역태깅프로그램의 정식명칭은 “UTagger-SR”이다. 이후 “UTagger-SR”이라고 한다.

3 김완수, 육철영(2015), <한국어 격틀 사전과 의미역 빈도 정보를 사용한 한국어 의미역 결정>, 《한국정보과학회 학술발표논문집》, 한국정보과학회, pp.651-653.

4 전산적인 처리상에서의 실수나 오류는 프로그램 개발자에 의해 수정이 될 부분이라서 본고에서는 논의하지 않기로 한다.

5 이후 《표준》이라고 함.

에 각각의 의미역을 주석할 수 있다는 장점을 가지고 있다. 그러나 용언의 다의단계까지 의미역을 주석하기에는 적당 분량의 정보가 부족하다는 문제가 발견되었다. 그 문제는 사전에 등재된 모든 용언의 정보 즉 문형 정보, 의미 정보, 용례 정보가 다의단계까지 세분화되어 일목요연하게 제공되지 않고 있다는 점이다. 문형 정보, 의미 정보, 용례 정보가 다 없는 경우도 있었다. 이렇게 정보가 부족한 경우 최대한 유의어를 찾아 빠진 정보를 채워넣으려고 노력했다. 그럼에도 정보가 채워지지 않는 부분에 대해서는 미봉책으로 선택한 것이 국어의 가장 기본이 되는 문장 구조인 '주어 서술어'에 맞춰 주어항을 기본으로 하여 의미역을 상정하였다. 형용사의 경우에는 대상역이 주석되었고, 동사의 경우에는 대체로 자동사인 경우가 많아 대상역이 주석되었고, 간혹 타동사인 경우에 행위주가 주석되기도 하였다. 구체적으로 의미역이 태깅되는 기준은 실제 말뭉치에서 문장이 실현될 때 나타나는 의미역에 맞춰 태깅하도록 하였다.

반자동으로 의미역을 태깅하게 되면서 작업자에게는 두 가지 기준이 제시되게 된다. 먼저 격조사별로 가능한 모든 의미역이 주석된 결과물은 작업자가 명시적으로 태깅할 결과물이 떠오르지 않을 때 참고하기 위한 자료로서의 '거시적' 기준으로 사용이 되고 있다. 격틀사전에 주석된 의미역 정보는 해당 용언의 다의 정보에 의미역을 부착한 자료이므로 해당용언과 해당 문형이 제시되면 한정된 의미역이 태깅될 수 있도록 정보를 제공한 것이므로 '미시적' 기준으로 보기로 하였다. 거시적 기준의 경우는 문맥 안에서 상관성이 고려되지 않은 채 격조사별의 형태정보로서 가능한 모든 의미역을 제시하여 보여주는 것이고 미시적 기준의 경우는 거시적 기준의 광범위한 범위와는 달리 매우 한정적인 즉 용례에 제시된 정보에 의한 의미역을 제시하여 보여주는 것이다. 그러므로 미시적 기준인 격틀사전의 의미역 정보의 상세화가 의미역을 태깅하는 작업을 가장 원활하게 할 수 있게 해주는 준거들이 되어야 한다. 그러나 앞에서도 언급한 바와 같이 격틀사전에 주석된 의미역과 의미역을 태깅한 결과물의 일치율이 70% 수준에 머물고 있어 정확도를 높이기 위해 구축된 격틀사전의 수정보완 및 확장이 필요하다.

본 연구 과정에서 진행한 의미역 태깅 결과물은 1, 2차 과정에서 100만 개 정도가 구축되었다. 두 차례에 걸쳐 의미역 태깅을 하면서 격틀사전의 문형 정보 부족과 실제 문장의 비문법적인 특징이 문제점으로 드러났다. 격틀사전의 문형정보 부족이라는 문제는 《표준》의 문제로 귀결되었다. 그리하여 《표준》의 사전적 문제점에 대한 논의를 다시 살펴보고 그 해결 방안을 모색해 보기로 하자. 그 다음으로 실제 말뭉치 대상 작업을 할 때의 문제점으로 계량언어학적인 면에서의 문제에 대한 연구를 살펴본 후 이를 해결하기 위한 방안을 모색해 보기로 하자.

본고는 2장에서 기존의 《표준》의 문제점에 대한 논의와 실제 말뭉치의 처리에 관련된 계량언어학인 면에서의 기존의 연구를 제시하고, 3장에서는 이와 관련하여 의미역 태깅시 발생하는 문제점을 제시하겠다. 4장에서 의미역 태깅시 발생하는 문제점의 해결방안을 모색해 보고 5장에서 결론을 제시하겠다.

2. 관련연구

본고에서는 의미역 태깅시 발생하는 격틀사전의 문제점은 구축된 격틀사전의 문제이기

전에 격률사전의 기반 사전인 《표준》의 문제점에서 기인한 것임을 알고 이에 대해 기존의 연구를 살펴보고자 한다.

2.1. 사전적 문제

《표준》의 문제점에 대한 논의는 ‘표제항의 선정’과 ‘통사제시’, ‘용례’의 문제로 좁혀서 살펴보고자 한다. 이 세 가지가 본고에서 문제로 생각하고 있는 부분과 관련되기 때문이다.

표제항의 선정에 대한 문제점을 지적한 논의로는 조재수(2000), 유현경(2010), 이은령·윤애선(2007)이 있다.

조재수(2000:137)는 《표준》에서 “다른 표제어에 딸리지 않고 배열되는 말”이라고 설정하여 일부 파생용언과 파생부사 따위의 경우 그 어근을 주표제어로 삼았다고 기술했다⁶. 이는 유현경(2010)에서도 마찬가지로 문제로 언급된 바 있다. 유현경(2010:223)에서는 그 중에서 일음절 불구 어근까지를 표제어로 삼는 것을 추가로 지적하고 있다⁷.

이은령·윤애선(2007)도 표제어 제시 방법과 연계된 문제를 지적하였다. 이는 표제어와 부표제어의 원칙에 기인한 문제로 어근 명사에만 의미 정보를 주고 부표제어인 동사에는 의미 정보가 없이 문형과 용례만을 제시한 점이다. 이은령·윤애선(2007:181,190)에서는 《표준》의 다의구분과 의미정보에서 의미정보의 구성에 문제가 있음을 지적하였다. 《표준》은 공통 문형 정보를 앞세워 동일 부류의 의미를 함께 묶고, 이후 하위 단계에서 각 어휘의 의미를 번호로 구분하며 각각의 의미에 해당하는 표제어 용례를 제시하고 있다. <명사+접사>동사의 파생현상이 국어에서 매우 빈번하기 때문에 의미 기술의 경제성에 근거하여 <명사+접사>파생동사의 의미를 호도하는 결과를 낳았다고 했다⁸.

다음은 사전에서의 문형정보 제시에 대한 문제를 지적한 논의로 고석주(2003), 이은령·윤애선(2007)을 들 수 있다.

고석주(2003)에서는 《표준》이 《연세》⁹에 비해 문형정보의 단순제기가 미흡한데, 이는 ‘교차정보어구문’을 이루는 서술어가 가지는 문법적 특성을 명확히 밝히지 못하고 있기 때문이라고 지적하였다. 또한 격률(문형)정보를 단순히 조사의 형태로만 제시하는 방식은 실제 문장에서 조사를 대입했을 때 문장의 정문 여부를 확인하기 어렵다고 했다¹⁰.

이은령·윤애선(2007:175~181)은 《표준》의 통사정보 제시는 “주어를 제외한 용언의 필수 성분만을 격조사나 어미로 표시한다.”라는 원칙에 따르기 때문에 문형정보에 쓰인 격조사의 통사, 의미적 주해는 일러두기에 명시될 뿐 이를 통해 해당 동사가 생성하는 모든 통사 구조는 알 수 없다고 하였다. 게다가 용례의 문형은 제시된 문형정보만을 반영하는 데에 지나치게 충실하여 일반적인 문형과는 거리가 있다고 보았다¹¹.

6 조재수(2000), <문제점이 많은 표준국어대사전>, 《새국어생활》 10-1, 국립국어연구원, pp.133~149.

7 유현경(2010), <한국어대사전 편찬에 대한 새로운 제안 - 표준국어대사전에 대한 평가를 기반으로>, 《한국사전학》 15, pp.220~246.

8 이은령, 윤애선(2007), <표준국어대사전의 동사 정보 개선을 위한 연구-한국어 어휘의미망의 구축에서 나타난 문제점을 중심으로->, 《한민족어문학》 31, pp.157~194.

9 연세대학교 언어정보개발연구원 편(1998), 《연세한국어사전》, 두산동아.

10 고석주(2000), <사전의 문법 정보에 대하여>, 《언어사실과 관점》 12-13, pp.181~216.

11 이은령, 윤애선, 앞의 책. pp.175~181.

《표준》의 용례의 미흡함을 제시한 논의로는 정호성(2000), 한영균(2006)을 들 수 있다. 정호성(2000)에 의하면 《표준》의 경우 주표제어와 부표제어를 합하면 50만 8천 여 항목이 된다. 이 중 용례가 제시된 주표제어는 74,092항목, 부표제어는 25,651항목으로 전체 표제어의 19% 수준에 머무르고 있다. 《표준》의 표제어 중 용례가 제시된 19% 중에서도 인용례가 전체의 10%에 불과하다¹²고 하였다.

한영균(2006)은 현대국어의 표준어 용언 표제항 중에서 용례가 있는 것은 48%에 불과하고, 51.4%에 달하는 항목에는 용례가 없다고 밝히고 있다. 용례가 있더라도 용례수는 평균 2개 안팎이며 다의어의 경우에는 주표제항과 부표제항 다의어의 경우에 차이가 있다고 했다. 또한 부표제항의 경우는 소홀하게 다루어졌음을 지적하고 있다. 용언의 용례가 1개인 항목이 전체의 37.3%를 차지했고 인용례는 22.85%, 작성례는 77.15%이어서 사전 용례의 충실성을 판단하기가 어렵다고 지적하였다.¹³

이를 정리해보면 《표준》의 표제항은 주표제항과 부표제항으로 나누면서 부표제항을 부차적으로 다루어졌고, 이러한 처리 기준에 의해 실제로 주표제항에 위치해야 할 과생동사들이 대다수 부표제항에 오면서 중요한 뜻풀이나 용례 제시, 통사정보제시 등의 정보가 미흡하게 처리되었다는 것을 알 수 있다. 또한 용언의 통사정보제시에서 주어부를 생략된 격들(문형) 정보를 제시하면서 실제 가능한 최소한의 정보만 제시했지 확장가능한 데까지 격들(문형)정보¹⁴를 제공하지 못하고 있다는 점을 알 수 있다.

2.2. 실제 말뭉치 처리에서의 문제

실제 말뭉치에서 드러나는 문제는 대부분 비문법적인 문장이다. 그러나 비문법적인 문장이 아님에도 의미역 태깅이나 전산언어 처리에 문제가 되는 것이 있다. 그 중 가장 두드러진 문제점이 띄어쓰기이다. 이에 대해서는 한영균(2003)에서 말뭉치를 통한 어휘 계량적 분석의 문제점으로 지적한 것이 관련성이 커 살펴보고자 한다¹⁵. 한영균(2003)은 특히 어휘 빈도 조사에 있어 가장 큰 골칫거리 중 하나가 띄어쓰기 문제이며 이를 해결해야 할 과제이지만 그 해결이라는 과정이 참으로 어렵다고 했다.

한영균(2003)은 띄어쓰기의 문제는 합성어의 판별기준이 모호하기 때문이라고 지적했다. 한영균(2003:73)에서 어휘 빈도 조사에서는 어떤 사건의 표제항을 준거로 하든 표제어로 등재되지 않은 형태와 등재된 형태가 공존하며 복합 구성 문법 단위는 사전 표제어로의 등재 여부와 관계없이 항상 띄어쓴 예와 붙여 쓴 예가 출현한다고 하였다. 따라서 왜곡되지 않은 어휘 빈도 조사를 위해서는 조사 대상 단위를 한정할 필요가 있고, 아울러 하나의 단위를 띄어 쓴 예들을 처리하기 위한 별도의 방법이 개발되어야 한다고 주장하였다.

위 2.1.과 2.2.의 관련 연구를 통해 《표준》의 표제어 선정과 구성의 문제가 의미역 태깅을 위해 구축된 격들사전의 문제점과 관련됨을 알 수 있다. 또한 어절 단위로 의미역을 태깅하는 현 방식대로라면 말뭉치 대상 태깅시 띄어쓰기의 문제가 앞으로 해결해야 할 과제를 알

12 정호성(2000), <『표준국어대사전』 수록 정보의 통계적 분석>, 《새국어생활》 10-1, 국립국어연구원, pp.55~72.

13 한영균(2006), <한국어 어휘 교육·학습 자료 개발을 위한 계량적 분석의 한 방향>, 《어문학》 94, pp.119~146.

14 3장부터는 '문형정보'로 하기로 함.

15 한영균(2003), <어휘 계량적 분석과 띄어쓰기 문제>, 《한국문화》 31, pp.49~71.

수 있다. 이들 문제점이 격틀사전과 실제 작업말뭉치에서 드러나는 것을 3장에서 살펴보고도
록 하겠다.

3. 의미역 태깅의 문제점

2장에서 살펴본 사전의 정보제시 문제가 격틀사전의 문제이다. 그리고 실제 의미역을 태
깅할 말뭉치의 다양한 문제점 중에서 띄어쓰기에 의해서 발생하는 어휘의 단위와 문장 내
서의 해석의 중의성 등의 문제가 있다.

3.1.에서는 구축된 격틀사전의 문제점을 정리해 제시하고 3.2.에서의 실제 말뭉치의 문장
에 드러난 어휘와 문장 구조적인 문제점을 정리해 제시하도록 하겠다.

3.1. 구축된 격틀사전에서의 문제

격틀 사전의 문제점은 2.1.에서 살펴본 바와 같이 우선 정보 제시 유무에 따른 처리 과정
의 문제를 들 수 있다. 그 대상은 문형정보의 부족과 용례의 부족이라는 두 가지 문제점에서
기인했다고 볼 수 있다.

다의수준까지 용언을 세분한 결과물을 대상으로 문형 정보 유무를 따져보았을 때는 문형
정보가 있는 것이 41,522개, 문형 정보가 없는 것이 48,578개로 문형정보가 없는 것이 좀더
많았다. 이를 다시 정리해보면 문형 정보가 있는 것 중에 용례 정보가 있는 것은 2,318개,
용례 정보가 없는 것이 39,204개이고, 문형 정보가 없는 것 중에 용례 정보가 있는 것은
388개이고 용례 정보가 없는 것이 48,190개이다.

의미역을 주석하는 데에 있어서 적당량의 용례 정보가 있고 그에 해당하는 문형 정보가
제공되어야 한다. 그래야만 상정된 의미역에 맞춰 격틀사전을 구축할 수 있다. 그러나 문형
정보가 없다는지, 적당량의 용례가 없다는지의 문제는 격틀사전을 온전히 구축하는 데에 큰
걸림돌이라고 볼 수 있다.

본 연구의 격틀사전 구축 단계에서 용언을 다의수준으로 세분화한 것은 실제 문장에 드러
나는 각 용언의 뜻에 가장 적합한 의미역을 주석하기 위해서이다. 사전에서 용언의 형태를
동형어의 수준까지만 고려했다면, 다의 수준의 맥락정보가 필요할 때 격틀사전 또한 쓸모가
없어지기 때문이다. 즉, 다의 단계까지 구분을 하여 개별 어휘의 개별 문형과 의미정보, 용
례정보를 대입한 의미역을 주석하여 정확한 의미역을 태깅할 수 있도록 하기 위함이다. 격
틀사전에 구축된 9만 개의 용언¹⁶ 중에서 단의어는 37,866개로 적지 않은 분량을 차지하고
있지만, 그래도 나머지는 모두 다의어라고 볼 때 다의수준까지의 분석적인 의미역 주석은
반드시 필요한 작업이다.

문형정보가 있는 41,522개 중에서 문형정보가 2개 이상인 것이 3,989개로 9%밖에 없다.
나머지 91%는 문형 정보가 해당 용언의 다의 정보에 한 개만 적용되어 있다는 것이다. 격틀
사전이 미시기준으로서의 준거틀로 사용될 것이라고 하더라도 가능한 많은 문형정보가 제
공되어 여러 가지 맥락적 상황과 구문적 변수를 고려할 수 있어야 한다. 정확하게 문형까지
맞아떨어지는 논항의 형태가 제공되어 의미역이 자동으로 태깅될 수 있게 된다면 이는 전

16 격틀사전의 용언 수는 다의어 수준으로 계산하면 90,100개이고, 동형어의어 수준으로 계산하면
67,941개이다.

산언어학적 처리에서 큰 의미가 있는 작업이라고 볼 수 있다.

문형정보와 용례정보가 없는 용언의 경우는 기본적으로 처리할 수 있는 방안을 마련하였다. 이에 대해서는 4장에서 다시 언급하기로 하겠다.

이렇게 구축된 격틀사전을 “의미역SR”을 통해 실제 말뭉치에서 격틀사전으로 구축된 용언의 출현빈도가 0인 용언은 58,547개로 동형이의어 수준 전체 용언의 86%를 차지하고 있다. 실제 의미역 태깅 작업에 사용된 용언의 수는 9,394개로 전체 용언(동형이의어 수준) 중 14%만이 사용된 것이다. 사용된 용언도 출현빈도가 10 이상인 용언 3,007개로 32%의 비중을, 출현빈도가 1인 용언은 2,385개로 25%의 비중을 차지하고 있다. 격틀사전의 용언 중 한정된 14%의 용언만이 작업에 사용되고 나머지는 출현하지 않는다는 것은 사전에 구축된 용언의 특징적인 면이 가장 큰 이유이다. 즉, 사전에 등재된 단어 중 그 활용 가능성이 떨어지면서 사전에만 존재하는 단어가 많다고도 유추해 볼 수 있다¹⁷. 말뭉치적인 면에서의 특징을 고려한다면 신문기사, 잡지기사, 소설과 같이 유사한 장르에서의 글을 말뭉치로 구축하였기 때문에 사용되는 용언의 형태가 한정될 수 있다는 특징을 들 수 있다. 또 다른 특징으로는 한국어를 사용해서 말하거나 문장을 쓸 때 주로 사용하는 단어가 다양한 여러 표현으로 바뀌기보다는 일관된 어휘를 사용하도록 규정하고 있고, 기초적인 단어로 쉽게 글을 써야 한다는 지침서적인 부분이 적용된 결과라고 볼 수도 있을 것이다.

격틀사전에서 실제 말뭉치에 사용된 용언 중 상위빈도 10위까지 정리한 대상을 보면 <표 1>과 같다.

<표1> 격틀사전의 용언 중 실제 말뭉치 출현 빈도 상위 10위

순위	용언	출현 빈도
1	하_01	13,314
2	있_01	11,429
3	되_01	7,461
4	없_01	6,437
5	아니	4,315
6	보_01	3,891
7	대하_02	3,637
8	같	3,609
9	위하_01	2,669
10	받_01	2,327

<표1>에 의하면 격틀사전의 용언 중 실제 말뭉치에 출현 빈도가 상위 10위 안에 드는 것으로는 ‘하다, 있다, 되다, 없다, 아니다, 보다, 대하다, 같다, 위하다, 받다’이다. 이들 용언의 특징은 한국어 문장에서 가장 많이 출현하는 동사이지만, 한국어의 기초동사인 ‘먹다, 자

17 현실에 사용하기 위한 실용적인 단어라기보다 문어에서만 사용되었던 단어이거나 한자어 결합형(2음절 한자+파생접사)인 경우, 또는 옛 문헌에서는 사용했으나 현대의 문헌에서는 드러나지 않는 단어가 많다는 의미이다.

다, 가다, 사다 등의 어휘적인 역할이 중요한 동사와는 다른 기능적인 면이 중요한 단어라는 점이다.

출현빈도 상위 10위 중 1위를 차지한 ‘하_01’을 살펴보도록 하겠다. ‘하_01’은 동형이의 수준에서는 하나의 정보를 담고 있다. 그래서 하나의 정보가 빈번하게 출현했다고 생각할 수 있다. 그러나 ‘하_01’은 다의수준으로 나누면 <표2>에 제시한 바와 같이 본용언만 10개의 문형으로 나뉘고 각각의 문형별로 다의정보가 나뉘어 전체 32개의 다의정보로 구성되어 있다. 보조용언의 다의정보까지 합하면 41개나 된다. 이렇게 세분화된 복잡한 의미와 문형 정보를 가지고 있으니 당연히 가장 출현빈도가 높을 수밖에 없을 것이다. 이는 기계적으로 정확하게 정보에 맞춰 분석하는 것도 복잡하고 사람이 바로 보고 바로 의미역을 주석하는 것도 어렵다는 점을 알 수 있다.

본고에서 격틀사전에 이용한 격틀(문형)정보와 다의적으로 구분된 용례정보를 중심으로 ‘하_01’을 다시 살펴보면 다음과 같다.

《표준》에서 ‘하_01’의 본동사의 격틀(문형) 정보는 10개가 있다. 10개의 격틀(문형) 정보 하위에 각각의 문형별 다의정보를 따로 구분하고 있다. 보조용언은 보조동사와 보조형용사를 모두 ‘하_01’ 동사 하위에 두고 있다. 보조용언도 각각 나누어 해당 뜻풀이와 용례를 제시하고 있다. 이렇게 ‘하_01’의 전체 격틀(문형)정보는 12개이고 다의정보까지 세분하면 41개의 개별 의미를 가지고 있다. 이렇게 다의정보 수준으로 의미역을 세분화하여 태깅한 결과를 탑재한 격틀사전과 동형이의어 수준으로만 용언을 찾아서 제시해주는 형태소분석 시스템 사이의 차이로 인해 기계 처리를 하든 사람이 작업을 하든 41개의 다의정보를 암기하고 있지 않은 한 계속해서 혼동이 발생할 우려가 있다¹⁸. 형태소분석에서부터 본용언과 보조용언의 경계가 애매해서 틀리게 된 경우는 작업자의 언어학적 지식과 사전의 정확한 구분 기준이 필요하다. 보조용언은 태깅의 대상이 아니라고 간과하고 넘어갔다가 대상 용언이 보조용언인데 본용언으로 태깅되어 잘못 드러난 것은 아닌지 살필 일이 종종 발생하게 된다.

<표2> 격틀사전 내 ‘하_01’의 의미역 주석 정보

대상 단어	품사 정보	격틀(문형) 정보	태깅된 의미역 정보
1 하다 010101 ¹⁹	[동사]	<...을>	{X:행동주 Y:대상-을/를}
2 하다 010102	[동사]	<...을>	{X:행동주 Y:대상-을/를}
3 하다 010103	[동사]	<...을>	{X:행동주 Y:대상-을/를}
4 하다 010104	[동사]	<...을>	{X:행동주 Y:대상-을/를}
5 하다 010105	[동사]	<...을>	{X:행동주 Y:대상-을/를}
6 하다 010106	[동사]	<...을>	{X:행동주 Y:대상-을/를}
7 하다 010107	[동사]	<...을>	{X:행동주 Y:대상-을/를}
8 하다 010108	[동사]	<...을>	{X:행동주 Y:대상-을/를}
9 하다 010109	[동사]	<...을>	{X:행동주 Y:대상-을/를}
10 하다 010110	[동사]	<...을>	{X:행동주 Y:대상-을/를}
11 하다 010111	[동사]	<...을>	{X:행동주 Y:대상-을/를}

18 이에 대해 상호비교군으로 참고한 두 개의 사전이 있다. 하나는 《고대》이고, 다른 하나는 《연세》이다. 《고대》는 《표준》과 동일한 방식을 택하였으나, 《연세》에서는 본용언과 보조용언을 각각 01은 본용언 02는 보조용언으로 구분하여 다른 단어로 등재하고 있다. 이런 점에서는 《연세》의 등재방법이 기계처리에 훨씬 합리적인 것이다.

12	하다 010112	[동사]	<...을>	{X}행동주 Y:대상-을/를
13	하다 010200	[동사]	<...을 -게>	{X}행동주 Y:대상-을/를 Z:방법-게
14	하다 010301	[동사]	<...을...으로>	{X}행동주 Y:대상-을/를 Z:착점-으로/로
15	하다 010302	[동사]	<...을...으로>	{X}행동주 Y:대상-을/를 Z:방향-으로/로
16	하다 010303	[동사]	<...을...으로> <-기로>	{X}행동주 Y:대상-을/를 Z:내용-으로/로 {X}행동주 Y:대상-을/를 Z:내용-기로
17	하다 010400	[동사]	<...을 -고>	{X}행동주 Y:대상-을/를 Z:내용-고
18	하다 010501	[동사]	<...으로>	{X}행동주 Z:원인-으로/로
19	하다 010502	[동사]	<...으로>	{X}행동주 Z:경로-으로/로
20	하다 010601	[동사]		{X}대상
21	하다 010602	[동사]		{X}대상
22	하다 010701	[동사]	<-고>	{X}행동주 Z:내용-고
23	하다 010702	[동사]	<-고>	{X}행동주 Z:내용-고
24	하다 010800	[동사]	<...에/에게 -게>	{X}행동주 Z0:착점-에/에게 Z1:방법-게
25	하다 010900	[동사]	<-게>	{X}행동주 Z:내용-게
26	하다 011001	[동사]		{X}대상
27	하다 011002	[동사]		{X}대상
28	하다 011003	[동사]		{X}대상
29	하다 011004	[동사]		{X}대상
30	하다 011005	[동사]		{X}대상
31	하다 011006	[동사]		{X}대상
32	하다 011007	[동사]		{X}대상
33	하다 011101	[동사][보조]		의미역 태깅 대상 아님
34	하다 011102	[동사][보조]		
35	하다 011103	[동사][보조]		
36	하다 011104	[동사][보조]		
37	하다 011105	[동사][보조]		
38	하다 011106	[동사][보조]		
39	하다 011107	[동사][보조]		
40	하다 011201	[형용사][보조]		
41	하다 011202	[형용사][보조]		

위 <표2>의 ‘하다_01’에서 본용언에 해당하는 부분만 다시 살펴보겠다. 본용언 부분에서 문형정보가 10개로 구분되어 있고, 1번 문형(<...을>)은 12개, 2번 문형(<...을 -게>)은 1개, 3번 문형(<...을...으로>/<-기로>)은 3개, 4번 문형(<...을 -고>)은 1개, 5번 문형

19 부여된 번호는 차례대로 두 자리씩 나눠서 맨 앞의 두 자리는 동형이의 정보, 두 번째 두 자리는 문형 정보, 마지막 두 자리는 다의정보이다. 사전에 제시된 기호에 따라서 다시 살펴보면 [01 01 01]은 맨 앞의 01은 동형이의 정보이고, 가운데 01은 문형정보이고, 마지막 01은 다의정보이다. 이 중에서 동형이의정보를 따로 분리해서 제시할 때는 01과 같이 숫자로만 제시하고, 문형정보를 분리해서 제시할 때는 대괄호를 사용해서 [1]처럼 제시하고, 다의정보를 제시할 때는 소괄호를 사용해서 (1)로 제시한다. 간혹 다의정보는 원문자로 제시하기도 한다. 대괄호엔 로마자대문자로 표시한 경우는 문형정보가 아니라 한 단어가 다른 품사정보를 가지고 있을 구분하기 위해 사용한다. ‘하다01’의 경우는 품사 정보가 동사, 보조동사, 보조형용사로 나뉘어 있다. 즉, 부여된 [01 01 01] 번호만으로는 본용언인지 보조요언인지 혹은 동사인지 형용사인지 알 수가 없다.

(〈…으로〉)은 2개, 6번 문형(문형없음)은 2개, 7번 문형(〈-고〉)은 2개, 8번 문형(〈…에/에게 -게〉)은 1개, 9번 문형(〈-게〉)은 1개, 10번 문형(문형없음)은 7개의 다의정보로 구분된다. 그런데 문형 6번과 10번은 격틀(문형)정보가 없다. 즉 다의정보 배열순서가 의미적인 배열순서인지 문형의 구분에 의한 순서인지 제시된 문형정보에 의해서 구분되었다고 보기에는 체계적이지 못한 부분이 있다. 격틀사전을 구축하기 위한 위 정보대로라면 문형정보 없이 논항값을 찾고 해당 논항에 의미역을 부착하는 작업을 해야 한다고 봐야 한다. 다의정보를 구분하는 기준은 중심의미에서 의미적으로 떨어진 의미를 순서대로 나열하는 것을 원칙으로 삼고 있다. 그런데 그 기준이 애매하거나 증거의 틀이 명확하지 않다는 것을 <표2>의 격틀(문형)정보가 빈 곳을 통해서 알 수 있다. 즉, 의미역을 태깅할 때 기계든 사람이든 자료로 제시된 문맥에서의 의미 정보에 의해 정보를 파악하고 접근하는 것이 사전과 적확하게 맞아떨어지지 않는다는 것을 알 수 있다.

사전에서 격틀(문형)정보가 없다고 제시된 문형 6번과 10번의 다의정보²⁰는 실제로 문장에서는 앞뒤로 문장을 통사적으로 확대하면서 더 폭넓게 문장이 작성될 수 있어서 문형정보가 없다고 단정지을 수 없다는 점도 문제삼을 수 있다. 그러나 이 부분은 문형정보가 없는 것으로 처리될 수밖에 없다. 왜냐하면 본고에서 작업을 위해 기준으로 삼은 것이 격조사와 사전의 문형정보에 의한 격틀사전이기 때문이다. 즉 격틀사전은 사전에서 명확하게 제시해주는 정보가 있는, 정해진 것에 한해서만 의미역을 태깅하고 있고, 또한 사전에 제시된 정보에 의존해서 의미역을 태깅할 수 있도록 하고 있다. 또한 격틀사전만으로는 모든 응용된 확장된 문맥에 대처할 수 없기 때문에 기존의 작업을 통해 구축된 의미역 태깅 말뭉치를 통해 응용된 문장에 대처할 수 있는 방안이 세워져야 한다. 이 부분에 대해서는 4장에서 좀더 기술하기로 하자.

보조용언은 의미역 태깅대상이 아니다. 그러나 이 또한 태깅 작업자의 직관이나 형태소분석기의 분석 중 오류로 인해 간혹 대상 서술어로 연결될 때가 있다. 이때 드러나 오류나 작

20 국립국어원, 《표준국어대사전》.

<http://stdweb2.korean.go.kr/search/List_dic.jsp>

문형 6번과, 10번의 경우는 격틀(문형) 정보만 없는 것이 아니라 하위 용례 또한 없어 의미역 태깅을 위한 격틀사전에서 배제하였다. 그러나 실제 작업 말뭉치에서의 출현 빈도는 그리 낮지 않은 편이다.

[6]「1」((주로 ‘하여(서)’ 꼴로 ‘쫘’, ‘경’ 다음에 쓰여)) 일정한 시각이나 시기에 이른다.

「2」((‘하면’ 꼴로 명사 다음에 쓰여)) 이야기의 화제로 삼다.

「3」((‘하고’ 꼴로 명사의 단독형과 함께 쓰여)) ‘그것에 그치지 않고 거기에 더하여’의 의미를 나타내는 말.

[10]「1」((‘-거나 -거나, -든지 -든지, -고 -고, -다가 -다가, -거니 -거니, -둥 -둥, -까 -까’ 따위의 구성 뒤에 쓰여)) 나열되거나 되풀이되는 둘 이상의 일을 서술하는 기능을 나타내는 말.

「2」((‘-ㄴ가/-ㄹ까/-나, -려니, -려나, -거니’ 구성 뒤에 쓰여)) 생각하거나 추측하다.

「3」((‘-니 -니’ 구성 따위 뒤에 쓰여)) 이리저리하게 말하다.

「4」((‘-다’ 구성 뒤에서 ‘하면’ 꼴로 쓰여)) 만일 어떤 상황이 일어나면 그에 따르는 어떤 상황이 반드시 뒤따라움을 나타내는 말.

「5」((의성어 뒤에 쓰여)) 그런 소리가 나다. 또는 그런 소리를 내다.

「6」((인용 조사 없이 발화를 직접 인용하는 문장 뒤에 쓰여)) 인용하는 기능을 나타내는 말.

「7」((문장 앞에서 ‘하나’, ‘하니’, ‘하면’, ‘하여’, ‘해서’ 따위의 꼴로 쓰여)) ‘그러나’, ‘그러니’, ‘그러면’, ‘그리하여’, ‘그래서’의 뜻을 나타내는 말.

업자의 직관에 혼동을 주는 경우는 사전이 판별기준이 된다. 즉 격률사전에는 의미역이 주석된 단어가 아니어도 기본적으로 이러한 형태에 구별을 할 줄 알아야 하고, 혼동이 있을 때는 다시 사전으로 돌아가 이를 구별해 내어야 한다.

실제 말뭉치에 제시된 형태 중 사전의 격률정보의 적용이 용례와 맞지 않는 사례도 왕왕 나타나고, 사전의 용례 부족으로 모든 문장을 다 맞춰서 태깅할 수 있는 것도 아니기 때문에 구축된 격률사전대로 의미역 태깅이 되지 않는 경우가 발생하게 된다. 이 부분이 격률사전을 수정보완해야 하는 이유 중 하나이다.

‘하-01’의 문형 중에서 4, 7, 11, 12번은 구문 정보가 유사하다. 사전대로라면 명확하게 구별이 되고 있어서²¹ 문제가 없지만, 실제 문장에서 약간의 문형이 추가되기만 해도 명확한 구분 기준은 다시 의미역을 태깅할 때 혼선을 빚게 된다. 그러나 실제 문장에서는 문형 4번과 7번의 「1」의 간접 인용의 ‘고’를 구분하기가 어렵다. 왜냐하면 문형 4번을 간접인용에서 명사로 인용의 내부문장이 끝난 것과 구별하기가 어렵기 때문이다.

예를 들어, ‘저 사람은 직업이 외교관이라고 해요.’라는 문장이 있다고 치자. 이때 문형을 4번의 문형에 맞춰서 의미역을 태깅해야 할지 아니면 7번의 「1」의 문형에 맞춰서 의미역을 태깅해야 할지 고민하게 되어 있다. 왜냐하면 문형의 형태적인 면에서는 4번 문형에 맞고 의미정보로는 7번 「1」의 것과 맞기 때문이다.

21 국립국어원, 《표준국어대사전》
 <<http://stdweb2.korean.go.kr/search/View.jsp>>

‘하다01’의 [4], [7], [11], [12] 정보

[4][…을 -고]

이름 지어 부른다.

☞ 꿀을 얻기 위해 벌을 치는 것을 양봉이라고 한다.

[7][-고]

「1」((간접 인용의 경우에는 ‘고’가, 직접 인용의 경우에는 ‘라고’가 쓰인다))이르거나 말하다.

☞ 그 책에서는 세계는 이제 정보화의 전쟁에 돌입했다고 했다./경찰은 도망간 범인이 잡혔다고 하였다./친구가 영화 구경 가자고 했다./선생님께서 학생들에게 숙제는 해 왔냐고 하셨다./어떤 철학자는 “시간은 금이다.”라고 하여 시간의 중요성을 강조했다./피의자는 분명히 경찰에게 “나는 그 시간에 집에 있었습니다.”라고 하며 자신의 결백을 주장했다.

「2」((주로 ‘하는’ 꼴로 쓰이는데 ‘-고 하는’은 ‘-는’으로 줄기도 한다))다른 사람의 말이나 생각 따위를 나타내는 문장의 내용을 받아 뒤에 오는 체언을 꾸미는 기능을 나타내는 말.

☞ 그가 거짓말을 했다고 하는 증거는 있다./다음번에는 소고기를 먹자고 하는 제안이 나왔다./그동안 무엇을 했느냐고 하는 사장의 질책에 직원들은 모두 고개만 숙이고 있었다.

[II]「보조동사」

「6」((동사 뒤에서 ‘-고 하다’ 구성으로 쓰여))앞말의 사실이 뒷말의 이유가 됨을 나타내는 말.

☞ 눈도 오고 해서 일찍 귀가했다.

「7」((동사 뒤에서 ‘-고는 하다’, ‘-곤 하다’ 구성으로 쓰여))앞말이 뜻하는 행동을 습관처럼 하거나 앞말이 뜻하는 상황이 반복되어 일어남을 나타내는 말.

☞ 그 사람은 점심을 먹고 난 후에 고궁을 산책하고는 한다./이 지역은 가끔 돌풍이 불곤 한다.

[III]「보조형용사」

「2」((형용사 뒤에서 ‘-고 하다’ 구성으로 쓰여))앞말의 사실이 뒷말의 이유가 됨을 나타내는 말.

☞ 길도 멀고 하니 일찍 출발해라./집도 가깝고 한데 더 놀다 가지그래.

또한 문형 4번과 7번 「1」의 문형정보의 차이는 ‘을’ 존재여부인데, 문형 4번과 7번 「1」에 ‘을’이 반드시 나타나야 한다는 제약이 없다는 점도 혼선을 빚는 이유이다. 게다가 보조용언 「6」의 ‘-고 하다’ 또한 간접 인용문과의 구분이 어려워 대상 말뭉치에서 본용언으로 볼 것인지 보조용언으로 볼 것인지를 혼동하는 경우도 있다. 이처럼 실제 말뭉치의 문장이 사전의 용례대로만 나타나는 것은 아니라는 점이 기계적으로 문장을 처리하는 데에 쉽지 않은 문제임을 알 수 있다. 고석주(2003:192)에서 지적한 것처럼²² ‘교차정보보어구문’을 이루는 문법적 특성을 명확히 전달하지 못하고 있다는 《표준》의 문제점을 지적한 것처럼 모든 사전이 모든 등재어의 교차정보를 등재할 수 없다는 점에서 기인한 것이라고 볼 수 있다.

이상의 비교 결과를 통해 알 수 있는 것은 하나의 단어 정보 안에 너무 많은 다의정보가 세분화되어, 해당 다의정보마다 개별 문형과 뜻풀이가 다 되어 있지만, 제대로 용례가 제시되지 않거나 교차정보가 제공되지 않아, 문제가 되고 있다는 것이다.²³

3.2. 실제 작업에서 드러나는 문장 구조적 문제점

실제 문장의 구조적 문제점으로는 어디까지를 단어로 볼 것인가의 문제와 어디까지를 문장의 논항으로 인정하고 문형으로 볼 것인가의 문제라고 볼 수 있다.

실제로 문장이 가지는 비문법적인 것들의 대다수가 오타거나 띄어쓰기 오류이다. 문장을 완전히 비문으로 만든 경우는 의미역을 주석한다고 해도 불구적인 단계에서 머물게 된다. 정말 딱 들어맞는 논항에만 의미역을 태깅하게 되는데 이는 온전한 의미역 태깅이 될 수 없으므로 의미역을 태깅하는 계획 단계에서부터 배제대상이 되었다. 그렇다면 완전한 비문이 아니면서 비문법적인 부분이 되는 것은 띄어쓰기 문제라고 볼 수 있다. 띄어쓰기는 여러 가지 문제점을 가지고 있다. 실제 어문규정에서도 띄어쓰기에 대해서는 원칙인 띄어써야 할 것과 허용인 붙여써도 된다는 항목을 복수로 가지고 있어 쓰는 사람에 따라 형태가 달라질 수 있다는 기본적인 문제점을 안고 있다. 의미역 태깅 프로그램은 대상 말뭉치에서 한 문장을 기본 단위로 놓고 문장 내에서 서술어 위치에 있는 용언의 형태정보를 찾아 그 기본형을 의미역태깅 대상으로 추출해 낸다. 그리고 해당 문장의 구문 구성은 어절 단위로 인식을 하게 되어 있어서 각 어절의 형태소 분석에 의해 구문 관계를 유추하고 구문 레이블링이 되게 된다. 이에 문장 내에서 기본적으로 서술어가 요구하는 필수논항이 찾아지도록 프로그램이 되

²² 고석주, 앞의 글.

《표준》에서와 같은 ‘격틀(문형) 정보’를 단순히 ‘조사’의 형태로만 제시하는 방식이 가지는 문제를 아래와 같은 예에서 더 분명하게 드러난다고 했다. (18)에서의 문형정보로는 (19)에서 볼 수 있는 사실에 대한 파악이 가능하지 않다는 점을 지적하였다.

(18) 가다1 [...에/에게][...으로/[...을] 《표준》

(19) a. 지방에 사는 친구0에게/*로/*를/에게를(에겐) 간다.

b. 아침 일찍 서울-로/에/을 가셨다.

c. 새벽에 친구 집-을/에/?으로 가 본 적은 없다.

²³ 여기에는 두 가지 이유가 있다. 형태소분석 단계에서부터 본용언인지 보조용언인지 판별이 어려운 것에서 오류가 생겨서 제시된 경우가 발생한다. 이 경우 만약 본용언인데 보조용언으로 태깅이 되어 있다면 태깅 정보를 수정하고 의미역을 그에 맞춰 태깅해야 한다. 또 다른 경우는 본용언으로 태깅이 되어 있는 경우는 보조용언으로 태깅 수정을 하면 된다. 이 과정에서 의미역 태깅의 오류가 발생하게 된다.

어 있다. 이 부분이 기계처리가 되도록 하는 기본적인 프로그램상의 구성이다.

프로그램은 실제 문장의 형태가 정해진 대로 태깅할 수 있도록 바탕을 깔아서 준비를 해준다. 즉, 실제 문장의 어떠한 부분도 기계가 가공해서 다른 결과물을 내놓지는 않는다는 것이다.

띄어쓰기의 문제는 다음과 같다. 서술어 형태의 불균형성을 유도한다. 동일한 형태의 단어도 어떤 필자는 붙여 쓰고 어떤 필자는 띄어 쓴다. 크게 문제되지 않는다고 보는 사람들이 많다. 본용언과 보조용언은 붙여써도 띄어써도 상관없다. 그래서 어떤 사람은 붙여쓰는 습관이 있고 보조용언이라는 직관이 강한 어떤 사람은 항상 띄어쓰는 경향이 있다. 그러다보니 붙여쓴 결과물에 대해서 하나의 단어라는 착각을 하게 되는 경우도 상당히 많은 것이 현실이다.

띄어쓰기 처리는 사람의 인지단계에서는 문장을 이해하는 데에 크게 문제가 되지 않는다. 그러나 기계처리에서는 동일 형태의 띄어쓴 것과 붙여쓴 것의 공존으로 형태소분석에서부터 차이가 발생하게 되고 이 둘을 다른 것으로 처리하게 된다는 점이 적지 않은 문제점이다. 특히 복합어의 경우와 구구성을 하고 있는 형태들에서 두드러진다. 반드시 붙여써야 할 단어인 경우는 단어로 인정되어 사전에 표제어로 등재되게 되어 있다. 그러나 모든 단어를 사전에 실을 수 없다는 점이 이 부분에 대한 또다른 문제가 되는 것이다.

이 문제는 의미역을 태깅한 결과물을 통해 서술어가 가지는 논항이 무엇인지를 결과물로 추출하고 그 결과를 온전하게 맞춰 본다는 관점에서는 상당히 미흡한 결과물이 도출될 수도 있다는 우려가 남게 된다. 즉, 동일한 서술어 앞의 동일한 성분이 띄어쓰기에 따라 의미역이 달리 태깅되는 것은 그리 바람직하지 않다.

한영균(2003)은 어휘의 계량적 처리를 위해서는 이 부분이 선처리되든 후처리되든 띄어쓰기 문제를 해결하지 않고서는 제대로 된 처리가 될 수 없음을 지적한 바 있다. 본고에서도 이 부분이 의미역 태깅의 문제로 계속 제기되고 있어 프로그램상에서의 처리가 절실히 필요한 상태이다.

선처리를 한다면 원시말뭉치를 가공하는 과정이 필요하고²⁴, 후처리를 한다면 의미역을 태깅하는 단계에서의 표시를 남겨 의미역 태깅 말뭉치에 대한 결과물을 추출해 낼 때 이들을 묶는다든지 분리해낸다든지 방안을 마련해야 한다. 현재로서는 원시말뭉치를 훼손하지 않고 작업을 하고 있어 프로그램상의 후처리가 적합한 방안이다.

띄어쓰기의 문제는 단어의 문제에서 구구성의 문제로 확대되어 나타나는데, 먼저 대상이 되는 서술어에 대해서 살펴보면 다음과 같다.

서술어로 선정되는 것은 문장에서 동사는 'VV'로 형용사는 'VA'로 품사가 부착된 것이다²⁵. 형태분석에서 태깅오류는 'Utagger-SR'에서 작업자가 수작업으로 수정이 가능하기 때문에 작업 과정에서 즉각적으로 수정하여 작업에 반영하고 있으므로 크게 문제 삼지 않는다.

그러나 다음과 같은 경우는 의미역에 태깅의 문제가 되고 있다. 서술어의 형태적인 정보

24 오진영(2013)과 같은 경우에는 전산언어학에서의 높은 정확률과 정확한 결과물의 도출을 위해 구문분석의 대상으로 이러한 문제가 될 소지가 있는 문장을 아예 배제하고 바람직한 문장으로만 연구를 하기도 했다.

25 본고에서의 품사 태그는 세종태그셋에 의한 결과물이다.

가 사전에 등재되어 있지 않아 서술어로는 잡히지만 격틀사전에 주석된 의미역 정보가 없다든지, 동일 어형이 사전에는 합성어로 등재되어 있는데 띄어쓰기가 되어 각각이 대등한 단어로 인식이 된다는가, 부사와 동사의 결합형인데 띄어쓰기를 해서 두 어절로 처리가 되고 있다는가 여러 가지 실현형이 있다.

합성어인 사전 등재어를 각각의 형태별로 띄어 쓰면 형태소분석 단계에서부터 각각을 독립된 서술어로 형태소분석하게 되고 그 결과를 통해 의미역 태깅 프로그램에서는 각각을 서술어로 상정하게 된다. 상정된 서술어에 맞춰 논항을 찾고 이에 맞춰 의미역을 태깅하게 되므로 여기서는 온전한 결과물이 도출될 수가 없다.

반면에, 각각 독립된 단어인 것을 붙여쓴 경우도 발생하는데 이때는 형태소분석 단계에서 두 단어가 합성어가 아니므로 각각을 서술어로 태깅하게 되고 개별 서술어의 형태가 의미역 태깅프로그램에 별개의 서술어로 상정된다. 그러므로 붙여쓴 상태지만 문장에서의 개별 서술어에 맞춰 의미역을 태깅할 수 있어 크게 문제가 되지 않는다.

<그림1>에 사전 등재어인 합성어 ‘몰아넣다’의 띄어쓴 사례를 제시하였다.

<그림1> 합성어의 실현 사례 - 몰아넣다

의미역부작		형태소-의존관계											
순서	의존	어절	3	마지_02	7	몰_01	8	넣	10	하_01	12	통일_02	시키
1	2	이_05/MM											
2	3	동작_03/NNG+을/JKO	THM										
3	4	마지_02/VV+고/EC											
4	7	나_01/VX+아서/EC											
5	7	화기_02/NNG+를/JKO				THM	THM						
6	7	단전_01/NNG+에/JKB				GOL							
7	8	몰_01/VV+아/EC											
8	9	넣/VV+는/ETM											
9	10	호흡/NNG+을/JKO							THM				
10	12	하_01/VV+어/EC											
11	12	정신_12/NNG+을/JKO									THM		
12	12	통일_02/NNG+시키/XSV+L 다/EF+./SF											

‘몰아넣다’는 격틀사전에 등재된 합성어이다. 그러나 한 단어인 것을 실제 작업 말뭉치에서는 <그림1>에서처럼 ‘몰아’와 ‘넣을’로 띄어쓰기를 하는 경우가 많다. 이런 경우 프로그램에서는 각각을 형태소 단위로 보고 ‘몰다’와 ‘넣다’를 개별 단어로 인지한다. 이런 경우 지침서대로 ‘몰다’와 ‘넣다’ 각각에 ‘화기를’을 대상역으로 ‘단전에’를 착점역으로 태깅하고 있다. 이런 경우에는 서술어별 태깅된 의미역을 추출해 내는 과정에서 잘못된 결과물로 도출되게 된다. 즉 ‘몰아넣다’의 하위 의미역으로 결과물이 나오는 것이 아니라, ‘몰다’와 ‘넣다’의 의미역으로 결과물이 도출되는 것이다. 이에 대한 해결방안으로는 후처리를 고려중에 있다.

격틀사전에 있는 합성어(동형의의어 수준)는 총 726개이다²⁶. 이 중에서 실제 작업 말뭉치

26 다의어 수준으로 나누면 1484개가 나온다. 그러나 실제 말뭉치에 비교해 보기 위해서는 동형이

에 출현한 개수는 338개이다. 338개 중에서 형태가 합성어로 제대로 인지되어 띄어 쓴 형태 없이 존재하는 것은 198개이고, 이와 반대로 합성어로 인식되지 않은 그러니까 개별 단어로 분리되어 띄어 쓴 것이 17개이다. 그리고 나머지 123개는 합성어인 형태와 합성어가 아닌 분리되어 띄어 쓴 형태가 각각 공존하고 있다. 의미역 태깅의 문제 대상은 338개 중 140개 즉 합성어와 띄어 쓴 형태가 공존하는 것과 띄어 쓴 형태가 존재하는 것이다.

이와 반대인 경우로 사전 미등재어를 들 수 있다. 단어의 구성에 의해 마땅히 합성어로 처리할 수 있는 것인데, 사전에는 등재되지 않은 것이다. 이러한 단어는 의미역 태깅 프로그램에서는 형태소 태깅 단계에서 대상 단어로 추출되어 제시되지만, 격틀사전에 존재하지 않는 형태가 되므로 기계처리가 불가능하다. 다만, 작업자의 직관에 의해 의미역이 태깅되게 되어 있어 객관적 기준이 필요한 상황이다. 이에 대해서는 방안으로 격틀사전에 미등재어를 추가해서 의미역을 주석한 결과물을 마련해야 한다.

이러한 예는 대체로 사전에서 접사로 인정한 ‘-하다, -되다, -받다, -당하다’류에 의해 파생된 단어들이다. 이러한 단어들은 격틀사전에 존재하지 않으므로 의미역 정보 또한 존재하지 않는다. 이런 류의 단어들은 형태적으로 사전에 등재되지 않은 단어라서 실제 말뭉치에서의 출현 시 띄어쓰기에 일관성이 없다.

<그림2> 세종구구조분석말뭉치 내 《표준》 미등재어 : 작품하다

의미역부착	형태소-의존관계							
순서	의존	어절	4 보이_01	7 보_01	8 작품_01하	10 일하	11 보_01	14 보이
1	2	개인적/NNG+이/VCP+L/ETM						
2	8	자원_01/NNG+에서/JKB						
3	8	논_01/NNG+에/JKB	THM					
4	5	보이_01/VV+지/EC						
5	6	않/VX+는/ETM						
6	7	관객/NNG+들_09/XSN+과/JKB		COM	COM			
7	8	마주_01/MAG+보_01/VV+며/EC						
8	21	작품_01/NNG+하/XSV+다가/EC+/_SP						
9	24	HIF의/NNG+에서/JKB				com		

<그림2>에 제시한 ‘작품하다’는 미등재어이다. 보는 바와 같이 실제 문장에서는 하나의 단어로 붙여 사용되고 있다. 문형을 유추해 보면 [○○가 작품하다/ ○○가 ○○와/과 작품하다/ ○○가 ○○에서 작품하다/ ○○가 ○○으로 작품하다] 등 많은 유형을 만들어 볼 수 있다. 원래는 ‘작품을 하다’로 사용했을 단어인데 사용빈도가 높아지면서 ‘작품하다’로 줄여 사용하게 되고 그 표현이 동시에 익숙해지면 ‘작품을 하다’와 ‘작품하다’의 너앙스 차이가 발생하게 된다. ‘작품을 하다’는 서술어가 ‘하다’에 초점이 맞춰지고 행위의 대상이 ‘작품’이 된다. 그러나 ‘작품하다’로 사용되면서는 서술어가 ‘작품하다’가 되고 행위의 대상은 필요없고 주체만 필요한 단어로 바뀌게 된다. 즉 ‘작품을 하다’와 ‘작품하다’를 동일한 형태로 다룰 수 없게 된다는 의미이다. 기존의 방식대로라면 ‘작품하다’가 출현빈도가 더 높아지고 단어로서 하나의 의미를 가지는 단어로 굳어지게 되면 바로 사전 등재어가 될 것이다. 그러나 현

의어 수준으로 가공해야 한다. 실제 말뭉치에서 형태정보를 제공하는 것은 동형이의어수준까지만 제시해주기 때문이다.

제로서는 미등재어이므로 의미역을 태깅하기 위해서는 의미역 태깅용 사전의 목록에 이를 올려서 의미역 태깅이 반자동으로 원활하게 이루어질 수 있도록 해주어야 한다.

실제 작업 말뭉치 내에서 존재하는 사전미등재어 중에 '-하다, -되다, -당하다, -받다'류를 추출한 결과를 <표3>에 제시하였다.

<표3> 1, 2차 의미역 태깅 작업 결과 중 사전미등재 '-하다, -되다, -당하다, -받다' 개수

미등재어	개수	-하-	-되-	-당하-	-받-
	누적출현율	505	363	120	6
중복어 삭제	171	157	85	6	

미등재어임에도 실제 작업 말뭉치에 출현한 개수는 상당히 많은 편이다. 이외에도 더 많은 단어들도 존재하지만, 우선 '-하다, -되다, -받다, -당하다'류의 파생어만을 추출하여 개수를 알아보았다. 이들의 개별 출현 개수별로 보면 '-하다'가 171개로 가장 많이 나타났다. 누적출현율을 비교해 보면 '-받다'가 357개로 가장 많다. 사전미등재어 중에서 '-하다, -되다, -받다, -당하다'에 의해 파생된 단어가 미등재어 전체 1,091개 중 419개로 38%를 차지한다. 단독으로 제일 많은 비중을 차지하는 것은 '시키다'이다, '시키다'는 파생접사가 아님에도 불구하고 접사처럼 인식되어 형성된 단어가 438개나 된다. 이는 '-하다, -되다, -받다, -당하다'를 다 더한 것보다도 많은 비중을 차지한다. 이렇게 실재하지는 않지만 사전에는 존재하지 않는 단어들의 특징은 비규범적인 조어에 의한 결과물이다. 그러므로 비문의 일종으로 돌리고 넘어갈 수도 있다. 그러나, 비규범적이라고 해서 실제 작문에서 불가능한 어휘라고 볼 수는 없다. 기술문법 차원에서는 오타라고 인식된 것을 빼고는 조어법적으로 문제가 없는 형태들이 많기 때문이다. 또한 실제 작업 말뭉치의 문장에서 이 부분이 차지하는 비중이 적지 않은 관계로 완전 오타라고 인식되는 것을 빼고는 미등재어 사전을 구축하여 'UTagger-SR'에 탑재하는 것이 'UTagger-SR'의 반자동화를 활성화시키는 데에 훨씬 효과적인 방안이다.

다음으로 제기되는 문제는 문장 내부에서 드러나는 실제 문장의 종류와 추출되는 용언의 형태정보와 관계에서 발생한다. 현재 'UTagger-SR'에 의해 대상 용언으로 추출되는 형태는 어절 단위로만 되게 되어 있다. 즉, 본용언과 보조용언의 형태로 구성되어 있다면 본용언만 대상으로 추출한다는 뜻이다. 단적으로 그냥 보조용언은 의미만 보조하기 때문에 의미역을 태깅하는 데에 큰 걸림돌이 되지 않는 경향이 많다. 그러나 피동문과 사동문처럼 문장의 구조 차이가 큰 경우, 보조용언이 구조적으로 빠진 상태로 의미역이 태깅되는 것은 의미가 없다고 볼 수 있다. 아래 (1)의 ㄱ은 접사에 의한 사동문이고 ㄴ은 '-게 하다'에 의해 통사적으로 사동이 된 문장이다.

- (1) ㄱ. 선생님께서 철수에게 책을 읽히셨다
- ㄴ. 선생님께서 철수에게 책을 읽게 하셨다.

현재 'UTagger-SR'에서 대상 용언을 추출하는 방식에 의하면, (1ㄱ)은 <그림3>에서와 같

이 ‘읽히다’이고, (1ㄴ)은 <그림4>에서와 같이 ‘읽다’가 된다. 이렇게 용언이 추출되면 해당 용언의 앞뒤를 살펴 필수논항을 찾는다. 그 결과물을 다음과 같다.

<그림3> (1ㄱ)의 의미역 태깅 결과

반자동 태깅결과		수작업 태깅결과	
의미역부작	형태소-의존관계	의미역부작	형태소-의존관계
순서	의존 어절	순서	의존 어절
1	4 선생님/NNG+께서/JKS	4	읽히_02
2	4 절수_99/NNP+에게/JKB	AGT	
3	4 책_01/NNG+을/JKO	GOL	
4	8 읽히_02/VV+시/EP+었/EP+다/EF+./SF	THM	
		4	4 읽히_02/VV+시/EP+었/EP+다/EF+./SF

<그림4> (1ㄴ)의 의미역 태깅 결과

반자동 태깅결과		수작업 태깅결과	
의미역부작	형태소-의존관계	의미역부작	형태소-의존관계
순서	의존 어절	순서	의존 어절
1	4 선생님/NNG+께서/JKS+는/JX	4	읽
2	4 절수_99/NNP+에게/JKB	AGT	
3	4 책_01/NNG+을/JKO	THM	
4	5 읽/VV+게/EC		
5	5 하_01/VX+시/EP+었/EP+다/EF+./SF		

이처럼 통사적 과생의 경우는 그 형태가 앞의 본용언의 형태 정보만 보여주기 때문에 실제 문장 구조에 맞춘 의미역이 태깅되지 않는다. 또한 이러한 부분에 대해서는 태깅 지침을 보조용언까지 고려한 의미역을 태깅하도록 처리하고, 이 부분에 대한 후처리 방안도 마련해야 한다.

동일한 격틀정보에 의한 문장이지만, 격틀 앞에 오는 체언이 무엇이냐에 따라 의미역 태깅이 달라지므로 격틀정보만으로는 완전한 의미역 태깅이 되지 않아서 완전한 기계처리에는 계속해서 어려움이 따른다. 해당 사례는 다음과 같다.

- (2) 김치를[대상] 안주로[착점] 하다.
- (3) 유럽 통합의 꿈은[기점] 신중한 발걸음을[대상] 필요로[방법] 하다.
- (4) {투표를 바탕으로}[방법] 하다.

(2)~(4)의 예문은 술어가 ‘하다’이고, 격틀이 [~를 ~로 ~하다]로 동일한 문장 구조를 가지고 있다. 그러나 동일한 구조를 가진다고 해서 동일한 의미역으로 처리되는 것은 아니다. 격틀사전에 주석된 의미역은 사전의 용례와 문형에 기반한 것이어서 주석된 의미역은 매우 한정적이다. 이 부분이 실제 말뭉치에 대입했을 때에 'UTagger-SR'의 반자동 프로그램에 의한다면 사전대로만 태깅된 결과물이 도출되게 된다. 그러나 실제로 작업자의 수작업에 의해서 태깅기준이 달라지고 수정된 결과물이 새롭게 도출되어 말뭉치로 저장되게 된다. 현재까

지는 수작업에 의해서 의미역이 태깅되고 있으나 본 작업의 미래의 모습은 기계에 의해서 대다수의 문장이 자동으로 처리되게 하는 데에 있다. 그러므로 이렇게 다른 단어에 의해서 다른 의미를 전달해야 할 여러 상황에 대해 많은 자료를 구축한 후 다양한 상황에 잘 대처하도록 해야 한다.

(2)의 예는 구조만으로도 의미역이 태깅되는 가장 기본적인 유형이다. 그리고 의미역도 격틀사전에 주석된 대로 대상과 착점역으로 태깅하면 된다. 그러나 (3)의 예문은 드러난 문장 대로 의미역을 태깅하기에는 무리가 있다. 즉 서술어 ‘하다’의 대상이 ‘신중한 발걸음을’이고 착점이 ‘필요로’라고 태깅되기 때문이다. 이때 ‘하다’는 독립적인 서술어로서 보기 보다는 ‘필요하다’라는 서술어의 어근과 접사 사이에 조사가 개입되어 형태가 분리된 것으로 보아야 한다. 문장의 서술어를 ‘필요하다’라고 보고 ‘유럽 통합의 꿈을 이루기 위해서는 신중한 발걸음이 필요하다’라는 문장으로 재구성 해본다면 서술어 ‘필요하다’의 대상인 ‘신중한 발걸음이’만 의미역을 태깅하면 된다. 그러나 실현한 문장은 이와 다르게 되어 있으니 제시된 문장대로 의미역을 태깅할 수밖에 없다.

이렇게 태깅된 결과물은 우리 문장에서 온전한 형태의 서술어인 ‘필요하다’의 의미역 태깅 결과물로 도출되지 않고 형태가 분리되어 접사의 기능을 하는 ‘하다’의 형태에 잘못된 본용언으로서의 ‘하다’에 준해서 의미역을 태깅하게 되고 이러한 결과물이 도출되게 된다. 즉, ‘하다’의 의미역 태깅 결과물만 양적으로 더 느는 결과물만 생성하게 되는 것이다.

이 문제에 한국어의 단어의 경계에 대한 애매모호함과 하나의 단어 내부에 조사가 개입되고 띄어쓸 수 있다는 특이한 특징까지 더해지면서 온전한 서술어의 형태를 갖춘 문장으로 처리하기가 어렵다.

그 다음으로는 긴 문장(주로 복문에서 발생)에서 의존관계가 멀어지면서 해당 논항을 정확하게 찾아 의미역을 태깅하는 데에 어려움이 발생하는 경우를 들 수 있다.

실제 1차 작업 말뭉치의 문장 중에 복문 중 의미역 태깅의 중의성이 발생하는 문장을 살펴보자.

‘이러한 주장은 초기의 데카르트와 그의 친구이자 신부였던 메르센트가 잘 보여주는데 여기서는 메르센트를 중심으로 살펴보기로 한다.’ <BSHO0127.txt 623번 문장>²⁷

위 예문을 ‘이러한 주장은 초기의 데카르트와 그의 친구이자 신부였던 메르센트가 잘 보여준다.’와 ‘여기서는 이러한 주장을 메르센트를 중심으로 살펴보기로 한다.’는 문장으로 나눌 수 있다. 다시 후행문은 ‘여기서는 이러한 주장을 메르센트를 중심으로 살펴본다.’와 ‘여기서는 살펴보기로 하자.’로 나눌 수 있다. 후행문만을 대상으로 다시 살펴보면 다음과 같다. ‘UTagger-SR’에서는 ‘살펴보다’와 ‘하다’를 각각 서술어로 추출한다. 의미역을 태깅하는 방법을 구분해서 제시하였다.

- ㄱ. 여기서는[처소] 메르센트를[대상] 중심으로[방법] 살펴보다.
- ㄱ'. 여기서는[처소] (이러한 주장을) {메르센트를 중심으로}[방법] 살펴보다.
- ㄴ. 여기서는[처소] 살펴보기로[내용] 한다.

27 세종구구조분석 말뭉치 내 작업 대상으로 선정된 파일.

위 예문의 서술어 ‘살펴보다’의 논항을 두 가지 관점에서 구분해 보았다. 즉, 문장 자체만으로 의미역을 부여했을 때와 전체 문장의 의미를 기준으로 의미역을 부여했을 때로 구분해보았다. 선행문이 없었다면 ㄱ과 같이 ‘여기서는’은 처소역, ‘메르센스를’은 대상역, ‘중심으로’는 방법역이 된다. 선행문과 후행문이 결합할 때, 두 문장의 동일 정보로 존재한 ‘이러한 주장을’이 선행문에는 남고 후행문에는 탈락되었다. 동일 정보인 ‘이러한 주장을’을 재구한 문장이 ㄱ’이다. ㄱ’와 같이 의미역을 태깅하면 ‘여기서는’은 처소역, ‘이러한 주장을’이 대상역, ‘메르센스를 중심으로’가 방법역이 된다. 동일한 문형을 가지고 있지만, 가시적인 문장의 구성성분만으로 분석할 때와 선행문의 관계를 조합해서 생략된 형태를 복원해서 분석할 때의 결과물이 달라지는 것이다. 이처럼 의미역을 태깅하기 위해서는 서술어와 인접한 논항의 형태정보로만 되는 것이 아니라 전체 문장에서의 의존관계에 의해서도 고려되어야 정확한 의미역을 태깅할 수 있다.

4. 격틀사전과 문장 구조적 문제에 대한 해결방안

4장에서는 3.1.과 3.2.에서 논의된 문제점들을 한 문제점들에 대한 해결방안을 제시하고자 한다.

4.1. ‘문형정보와 용례정보의 부족 문제’에 대한 방안

3.1.에서 언급한 바와 같이 격틀사전에서 대상으로 삼고 있는 다의수준의 용언 90,100개 각각에 개별적인 문형정보와 용례정보가 적당량 제시되지 않고 있다. 사실 사전에 모든 정보가 충분할 정도로 많이 제시되어야 한다는 것은 현실적으로는 불가능하다는 것을 알 수 있다.

우선 격틀사전이 의미역 태깅 준거틀의 역할을 하려면 기본적인 조건이 충족되어야 한다. 즉, 대상 용언으로서 격틀을 제공해 주어야 하는 것이다. 격틀이 제공되어야 자동으로 의미역 태깅을 할 수가 있다. 이에 근거하여 격틀사전 구축 대상 용언 중에서 문형정보와 용례정보가 없거나 부족한 대상만을 모아 재작업을 시행하였다.

재작업 대상 중에서 문형정보가 없는 것은 자동사와 형용사이다. 자동사와 형용사는 사전에 용례가 있더라도 문형이 없는 경우가 많다. 이 두 종류의 대상 용언에 대해 의미역을 주석한 방법은 다음과 같다.

기본적으로 문형은 없으나 용례가 제시된 것은 용례를 기반으로 하여 의미역을 주석하였다. 주어항이 서술어의 행동주라고 하더라도 동작성이 없는 경우는 대상역으로 주석하였다. 아래에 제시한 표는 자동사인 ‘가감되다000000’와 형용사 ‘가공하다030000’의 격틀사전정보이다. 두 용언은 문형정보가 없으므로 의미역 정보를 제시된 용례정보를 통해 주석하였다. ‘액수가 가감되다’, ‘가감된 부분’과 같은 용례를 통해 ‘가감되다’ 앞에 오는 주어항에는 대상역이 옴을 알 수 있다. 또한 ‘가공할 파괴력’, ‘가공할 만한 위력’, ‘언론의 위력은 가공할 만한다’를 통해 ‘가공하다030000’의 주어항을 대상역으로 정하였다.

<표4> ‘가감되다000000, 가공하다 030000’의 격틀사전 정보

표제어	품사 정보	문형 정보	의미 정보	용례	의미역
가감되다 000000	동사		가감01(1). >>더하거나 더는 일. 또는 그렇게 하여 알맞게 맞추는 일. '더하고 빼기로 순화.'	실적에 따라 월급의 액수가 가감될 것이다. 사방을 돌아보아도 운동장이 축소된 흔적은 없었다. 울타리도 여전했고 건물도 여전했다. 조금도 가감된 부분이 없었다. >>월급은 능력에 따라 가감이 있을 수 있어요.	{X:대상}
가공하다 030000	형용사		두려워하거나 놀랄 만하다.	핵무기의 가공할 파괴력. 가공할 만한 위력을 지니다. 언론의 위력은 가공할 만하다.	{X:대상}

다음으로 문형정보와 용례정보가 없고 의미정보는 있는 경우이다. 이 경우에는 의미정보에 의지하여 의미역을 주석하였다. 의미정보는 표제어의 의미를 정확하게 풀어놓은 경우와 다른 표제어의 정보로 연결해 놓은 경우를 볼 수 있다. 표제어 자신의 의미 정보를 가진 경우 최대한 의미 정보에 맞게 의미역을 주석하려고 노력했다. <표5>의 예에 제시된 '싱경싱경하다010000'은 문장을 유추하면 '방이 싱경싱경하다'가 나오므로, 의미역은 대상역을 주석하였다. '묵다010002'는 문장을 유추하면 '방이/논이 묵다'가 나오므로, 의미역은 대상역을 주석하였다.

<표5> '싱경싱경하다010000, 묵다010002'의 격틀사전 정보

표제어	품사 정보	문형 정보	의미 정보	용례	의미역
싱경싱경하다 010000	형용사		방이 차고 서늘하다.		{X:대상}
묵다 010002	동사		방이나 논 따위가 사용되지 않은 채 그대로 남다.		{X:대상}

위와 같이 의미정보가 정확하게 존재하는 경우에는 의미역을 쉽게 유추하여 주석할 수 있으나, 이러한 경우는 어근형으로 의미정보가 제시되는 것에 비해 상당히 적은 편이다. 의미정보를 어근에 기대어 놓은 것이 많은 이유는 앞에서도 언급했듯이 많은 파생동사와 형용사를 그 어근형의 하위에 부표제어로 제시했기 때문이다. 해당 표제어의 어근형인 주표제어는 대다수가 명사이다. 의미는 주표제어의 의미에 접사의 의미를 더해서 '~을 하다' 또는 '~이 되다' 식으로 정하게 되어 있다. 이러한 경우도 해당 표제어의 주표제어를 찾아 그 의미에 의해 유추된 예문을 만들고 이를 통해 의미역을 주석하였다. 이러한 형태의 대부분은 2음절에 파생접사가 결합된 파생어이다. 게다가 사전에서 이들 파생어의 위치를 주표제어가 아닌 부표제어의 위치에 두면서 해당 표제어의 의미정보와 용례정보를 제대로 제시해주지 못한 부분이 상당하다. 그런데 이렇게 부표제어에 위치한 단어들 중에서 단어로는 인정되어 사전에 등재되어 있으나 실제 말뭉치에 출현하지 않는 것들이 꽤 많다는 것을 알 수 있다. 이에 대해서는 3장에서 이미 제시한 바 있다.

<표6> ‘방전되다010000, 방전되다020001, 방전되다 020002, 방전하다020000’의 격틀사전정보

표제어	품사 정보	문형 정보	의미정보	용례	의미역
방전되다 010000	동사		방전03. >>전파방해. >>무선 전신에서 여러 가지 전자기적 영향으로 전파가 방해받는 일.		{X:대상}
방전되다 020001	동사		방전04(1). >>전지나 축전기 또는 전기를 띤 물체에서 전기가 외부로 흘러나오는 현상.		{X:대상}
방전되다 020002	동사		방전04(2). >>기체 따위의 절연체를 사이에 끼인 두 전극 사이에 높은 전압을 가하였을 때, 전류가 흐르는 현상. 불꽃 방전, 진공 방전 따위가 있다.		{X:대상}
방전하다 020000	동사		방전05. >>화살을 쏘.		{X:행동주}

그 다음으로 의미정보마저도 없이 표제어만 존재하는 것도 상당수가 있다. 어떠한 사전적 정보제시가 없는 경우는 가장 기본이 되는 대상역을 주석하였다. 이는 한국어의 기본 문장으로 주어와 서술어로 가능한 문장인 ‘무엇이 어떠하다’, ‘무엇이 어찌하다’, ‘무엇이 무엇이다’를 기준으로 정한 것이다. 행동주를 의미역으로 주석하지 않고 대상역으로 정해놓은 것은 동사라고 하더라도 모두 주동적인 주체가 주어의 위치에 오는 것은 아니기 때문에 중의적인 입장에서 대상역을 기본으로 정하였다. 그러나 실제 문장에서는 대상역이 아닌 다른 성격의 논항이 출현할 수 있고 그때마다 해당 논항의 역할에 맞게 의미역을 태깅하도록 하였다. 이 부분은 앞으로 기계가 자동으로 의미역을 처리하는 데에는 걸림돌이 될 소지가 있다. 여전히 사람의 인지에 의해서 작업이 이루어져 할 부분이 존재하고 있음을 의미한다.

다음으로 격틀사전에 주석된 말뭉치의 한정성을 들 수 있다. 제한된 문형정보와 용례정보로 인해 주석된 의미역의 유형이 최소의 정보만을 제공한다는 점을 해결하기 위해서 본고에서는 고석주(2000)에서 언급한 바 있는 교차정보보어구문을 고려하였다. 즉, 사전의 제시 정보 외에도 용언이 출현할 수 있는 다양한 구문을 고려해 문형정보를 추가하고 추가된 문형정보에 의미역을 주석하여 탑재하는 것이다. 실례로 ‘밀리다010002’를 살펴보자.

<표7> ‘밀리다010002, 뒤처지다

표제어	품사 정보	문형 정보	의미정보	용례	의미역
밀리다 010002	동사		어떤 이유로 뒤처지게 된다.	교통사고로 차가 밀려 제시간에 약속 장소에 도착하지 못했다. 어려서부터 나는 항상 형에게 밀려 뒷전이었다. 별떡 일어나 앉는 춘보이의 기세에 옹 구네가 뒤로 밀리는 소리로 묻는다.	{X:대상}
뒤처지다 020001	동사	…보다 …에/에	어떤 수준이나 대열에 들지	성적이 남들보다 뒤처지다. 그는 심장이 약해 친구들보다 걸음이 뒤처진	{X:대상 Z: 비교기준}

		게	못하고 뒤로 쳐지거나 남 게 된다.	다. 이번 입찰 경쟁에서 다른 회사보 다 뒤쳐져서 는 안 되니 열심히 하지 오. 시대의 변화에 뒤쳐지다. 그는 친 구에게 뒤쳐진 성적표를 보고 무척 실망하였다.	보다 {X:대상 Z: 비교기준- 에/에게}
--	--	---	---------------------------	---	----------------------------------

‘밀리다010002’는 문형정보가 없으므로 가장 기본적인 의미역인 ‘대상역’을 주석하였다. 그러나 용례정보에 의해서 ‘차가 밀리다’, ‘나는 형에게 밀리다’, ‘기체에 옹구네가 뒤로 밀리다’의 문형을 찾을 수 있다. 이때 ‘차가 밀리다’는 ‘어떤 이유’로 함의하려면 ‘교통사고로 차가 밀리다’가 되어야 의미정보를 담고 있는 문형이 된다. 즉 이유를 나타내는 문형정보인 ‘-(으)로’가 필요하다는 점이다. 이때 이유를 나타내는 ‘-(으)로’는 원인역이 된다. 또한 두 번째 용례에서 ‘누가 누구에게 밀리다’라는 기본 문형을 찾을 수 있다. 이때 주어항의 ‘누가’는 밀리는 대상역이고 부사어항의 ‘누구에게’는 원인역이 된다. 세 번째 용례에서 ‘기체에’는 원인역이 되고 ‘옹구네’는 대상역, ‘뒤로’는 방향역이 된다. 또한 의미정보에 의해서 용언 ‘뒤쳐지다’와의 관계를 참조할 필요가 있다. 격틀사전의 ‘뒤쳐지다’의 의미역은 [X:대상 Z:비교기준-에/에게]로 주석되어 있다. ‘뒤쳐지다’는 문형정보에 ‘-에/에게’가 있고 이를 비교기준역으로 주석하고 있다. ‘밀리다010002’의 두 번째 용례와 뒤쳐지다의 의미를 유사패턴으로 볼 수 있다. 그리하여 ‘나는 형에게 밀리다’의 의미역에서 ‘형에게’를 원인역에서 비교기준역으로 수정하여 반영하였다. 이처럼 제시된 의미정보와 용례정보는 문형정보가 부재한 영역을 채워줄 수 있는 매우 중요한 정보이다.

또한 위와 같은 수정 보완 작업은 격틀사전의 전체 용언 중 실제로 말뭉치에 출현한 19%의 용언에 한해서 진행하였다. 이렇게 수정된 격틀사전을 통해서 의미역 태깅의 결과물을 비교해 본 결과 일치율이 83%로 수정 전 70% 수준에서 13%의 향상을 보였다.

4.2. 문장 구조적 문제에 대한 방안

문장 구조적 문제로는 띄어쓰기 문제가 있다. 3.2.에 제시된 <그림1>을 통해 드러난 합성어의 일관성 없는 출현에 대해서 한영균(2003)에서는 작업 전 말뭉치의 가공이 필요하나 그 또한 어려움이 있다고 언급하였다. 본고는 원시말뭉치를 가공하지 않은 상태에서 작업을 시행하고 있어 이러한 띄어쓰기 문제는 후처리를 통해 결과물을 도출하는 수밖에 없다. 프로그램 상에서의 처리 방안은 마련하지 못하였다. 다만, 3.2에서 언급한 것과 같이 결과물을 도출해 낼 때 자료를 가공하는 방안을 고려한 작업 지침서를 마련하였다. 그리하여 차후 결과물을 추출해 낼 때 합성어인 형태가 띄어쓰기가 되어 잘못 쓰인 형태를 합성어인 형태와 동일하게 처리하여 단어별 결과물이 잘못 나오는 정도를 줄이고자 한다.

3.2에서 예로 든 합성어 ‘몰아넣다’와 같은 단어가 ‘몰다’와 ‘넣다’로 띄어쓴 결과물이 대상 용언으로 선정되었을 경우 각각의 다르게 보지 않고 하나의 단어로 보고 동일한 의미역을 연결시켜 놓기로 정하였다. 이렇게 지침서를 정한 후 현재 기준 사전에서 용언 중에 합성어이거나 파생어인 단어를 추출하여 목록화 하였다. 이렇게 목록을 만들어놓은 용언의 의미역태깅된 결과물에서 띄어쓰기 오류로 출현된 경우 의미역 결과물이 동일하게 태깅된 것은 하나의 용언으로 처리하도록 후처리 방안을 정하였다.

두 번째로 미등재어 처리 문제가 있다. 우선 미등재어인 경우는 띄어쓰기 오류도 포함

된다. 그러나 띄어쓰기 오류는 합성어를 띄어쓴 위의 사례가 아닌 합성어가 아닌 것을 붙여 합성어처럼 사용한 경우가 여기에 해당한다. 그 중에서 큰 비중을 차지하고 있는 것이 파생어인데, 사전에서 접사로 인정된 ‘-하다, -되다, -받다, -당하다’류와 결합한 경우가 그것이다. 이들은 접사의 기능을 하고 있지만 실제 사전에는 모두 등재가 되지 못한 경우이다. 이러한 미등재어의 처리 방안은 두 가지가 필요하다. 하나는 미등재어이므로 비문으로 처리하고 말뭉치에 추가할 의미역 태깅 대상으로 삼지 않는 것이다. 그러나 이들 단어가 극소수라면 무시하고 넘길 수 있으나 그 수가 상당히 남아 무시하기에는 무리가 있는 경우에는 두 번째 안으로 미등재어에 대한 목록화와 이들의 의미역 주석을 추가적으로 진행하는 작업을 하는 것이다. 본고에서는 미등재어를 격틀사전에 추가하는 안을 선택하였다. 의미역을 기계가 자동으로 처리하게 하려면 미등재어를 단어로 인정하고 격틀사전에 추가하는 것이 더 나은 방안이기 때문이다.

사전미등재어로 실제 말뭉치에 출현한 용언은 전체 1,091개이다. 이 중에서 ‘-하다, -되다, -받다, -당하다’류는 419개로 38%를 차지하고 있다. ‘-하다, -되다, -받다, -당하다’보다 더 많은 미등재어에 존재하는 단어는 ‘NN+시키다’의 형태이다. ‘시키다’ 파생어 외에도 대부분 ‘-같다, -거러다, -답다, -대다, -드리다, -롭다, -맞다, -부리다, -스럽다, -이다, -지다, -쩍다’등의 접사와 결합한 형태가 있다. 그 중에서 ‘시키다’ 파생어가 438개로 40%를 차지해 가장 큰 비중을 차지하고 있다. 이들에 대한 의미역을 주석하기 위해서 우선 해당 용언의 어근의 의미를 찾고, 작업 말뭉치에 출현한 문장을 용례로 사용하였다. 또한 이를 기반으로 적정 문형정보를 만들고 해당 의미역을 주석하였다. 이렇게 미등재어의 의미역주석 정보는 격틀사전의 보조 사전이 된다.

세 번째로 보조 용언 구문 문제이다. 실제로 문장에서 서술어를 찾고 해당 서술어의 동사 기본형이 필수논항으로 무엇을 취하는지를 찾도록 만들어진 프로그램에서 보조적 문장 구성은 고려되지 못한 부분이다. 이 부분에 대해서도 선처리가 되지 않은 상황이기 때문에 후처리를 하도록 지침을 정하고 있다. 실제로 보조적 문장 구성은 뒤따르는 보조동사의 의미가 추가되어야만 문장 전체가 의미하는 바를 알 수 있다.

통사적 사동문과 같이 보조용언까지 고려해야 하는 문형이 가장 두드러진 사례가 된다. 의미역 태깅프로그램에서 의미역 태깅 대상 용언을 문장에서 추출할 때 본용언만 추출하지 뒤따르는 보조 용언 구문인 ‘-게 하다’는 드러나지 않는다. 그러나 실제 문장을 분석하는 사람은 통사적 사동문으로 문장을 분석하고 필수논항을 찾고 의미역을 태깅하는 작업을 하였다. 즉 후처리 과정에서 본용언 뒤에 따르는 보조 용언 구문을 모두 고려해서 결과물을 처리해야만 한다는 것이다.

관용구의 경우 또한 ‘미역국을 먹다’를 그냥 ‘먹다’의 대상역이 ‘미역국’이라고만 처리하고 나면 ‘힘에 떨어지다’는 의미역 태깅 결과물에서는 찾아볼 수 없게 된다. 즉, 올바른 문장 구성의 분석에 의해 의미역이 태깅된 결과물이 도출되지 않도록 하기 위해서는 반드시 통사적으로 서술어의 구성이 확대된 경우를 모두 고려 대상으로 삼아야 할 것이다.

5. 결론

본고에서는 두 차례에 걸쳐 의미역 태깅 말뭉치를 구축하였다. 이를 위해 격틀사전을 구축하고 'UTagger-SR'에 탑재하여 기존의 태깅 속도를 절반으로 줄이는 성과를 내었다. 격틀사전의 구축으로 인해 작업 속도가 빨라진 점은 좋은 성과이지만, 격틀사전만으로 의미역 태깅이 완전히 진행되지는 않았다. 이에 그 문제점을 찾고 해결방안을 마련해보았다.

우선 'UTagger-SR'에 탑재한 격틀사전이 작업 중 어떤 문제가 있는지를 찾아보았다. 이때 드러난 문제는 문형정보와 용례정보가 부족해서 주석된 의미역 결과물이 매우 적었다는 점이다. 또한 사전에 등재되지 않은 용언들이 출현하여 격틀사전이 적용되지 못하는 경우도 상당히 많았다는 점이다. 게다가 사전의 모든 용언이 실제 작업 말뭉치에서는 19% 정도만 출현했다는 점이다.

격틀사전의 문제점을 해결하기 위한 방안은 다음과 같다.

격틀사전의 문형을 추가하고 의미역을 다양한 문장에 대비할 수 있도록 확대하여 주석하는 방안을 마련하였다. 또한 미등재어를 격틀사전의 보조 사전으로 등재하여 의미역을 주석할 수 있도록 대비를 하도록 하였다.

다음으로 의미역을 태깅하는 데에 문제가 되는 것은 실제 작업 말뭉치의 문장의 문제점이었다. 단어의 띄어쓰기 문제에서부터 통사적으로 보조 용언 구문을 가지고 있는 문장까지 살펴보았다.

실제로 띄어쓰기에 대한 문제는 작업 전에 작업 말뭉치를 가공하지 않도록 되어 있어 작업 중 지침서와 작업 후 결과물 도출 과정에서의 후처리의 필요성을 언급하였다. 특히 문제가 되고 있는 것은 의미역을 태깅하기 위해 추출된 용언의 형태가 합성어인 것이 띄어쓰기가 되어 각각 독립된 단어로 제시된 경우이다. 그냥 문장을 읽고 이해하는 데에는 문제가 되지 않지만, 기계 처리에서는 큰 차이를 가지고 있으므로 이에 대해서 고민하게 되었다. 의미역 태깅 작업자는 이런 경우 반드시 사전을 통해 합성어 여부를 확인 한 뒤 작업 일지에 기록하고, 작업을 한 결과물에 대한 후처리 과정에 반영하도록 하였다. 각각이 독립된 단어로 의미역 태깅이 된 것이지만, 동일한 의미역을 부여해 합성어로서의 의미역으로 도출될 수 있도록 하였다.

또한 통사적 사동 구성과 같이 본용언만 의미역 태깅의 대상으로 추출된 경우가 있다. 이 문제는 눈에 띄게 다른 결과물이 도출되므로 반드시 본용언과 보조 용언을 하나의 구성으로 묶어서 필수논항에 의미역을 태깅해야 한다. 그러므로 작업자가 통사적 사동 구성으로 의미역을 태깅하였으므로 그 결과물은 또한 반드시 통사적 사동 구성문의 의미역 태깅 결과물이 도출되어야 한다.

이처럼 작업 말뭉치의 문장 내에서 드러나는 문제점은 대체로 작업자의 직관에 의해 올바른 구성에 맞춰 의미역을 태깅하고 그 결과에 대해 결과물을 도출하는 과정에 반드시 이러한 부분이 반영되어 바른 결과물이 도출되어야 한다.

감사의 글

"이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(R0101-15-0176) Symbolic Approach 기반 인간모사형 자가학습지능 원천기술 개발"

<참고 문헌>

- 고려대학교 민족문화연구원. 2009. 고려대한국어대사전. 고려대 출판부.
- 고석주. 2000. 사전의 문법 정보에 대하여. 언어사실과 관점 12-13: 181~216.
- 김완수, 옥철영. 2015. 한국어 격틀 사전과 의미역 빈도 정보를 사용한 한국어 의미역 결정. 한국정보과학회 학술발표논문집. 한국정보과학회. 651-653.
- 김윤정, 김완수, 옥철영. 2014. 전산언어학에서의 한국어 필수논항의 의미역 상정과 재고. 언어와 정보 18-2: 169-199.
- 남승호. 2008. 한국어 술어의 사건 구조와 논항 구조. 서울대학교출판부.
- 송복승. 1995. 국어의 논항구조 연구. 보고서.
- 연세대학교 언어정보연구원. 1998. 연세 한국어사전. 두산동아.
- 우형식. 1996. 국어타동구문연구. 박이정.
- 유현경. 2000. 한국어대사전 편찬에 대한 새로운 제안 - 표준국어대사전에 대한 평가를 기반으로. 한국사전학 15: 220~246.
- 윤준태, 정의석, 송만석. 1998. 명사간 어휘 정보를 이용한 한국어 복합 명사 분석. 정보과학회논문지(B) 25-11: 1716~1725.
- 이은령, 윤애선. 2007. 표준국어대사전의 동사 정보 개선을 위한 연구-한국어 어휘의미망의 구축에서 나타난 문제점을 중심으로-. 한민족어문학 31: 157-194.
- 이정훈. 2011. 통사구조 형성과 명사구 및 동사구. 국어학. 60: 264~421.
- 이희자·이종희. 2010. 한국어 학습 전문가용 어미·조사사전. 한국문화사.
- 임채경. 1993. 심리술어구문의 통사적 특성. 현대문법연구 3: 59~90.
- 임홍빈. 2007. 한국어의 주제와 통사 분석. 서울대학교출판부.
- 정주리. 2004. 동사, 구문, 그리고 의미. 국학자료원.
- 정현기, 김유섭. 2011. 확장된 격틀 사전을 이용한 한국어 부사격 논항의 의미역 결정. 한국정보기술학회논문집 9-10: 167~176.
- 정호성. 2000. 『표준국어대사전』수록 정보의 통계적 분석. 새국어생활 10-1: 55~72.
- 조재수. 2000. 문제점이 많은 표준국어대사전. 새국어생활 10-1: 133~149.
- 최형강. 2006. 피동문의 조건과 ‘받다, 당하다, 되다’ 구문의 재고. 국어학 92: 159-190.
- 한영균. 2003. 어휘 계량적 분석과 띄어쓰기 문제. 한국문화 31: 49~71.
- 한영균. 2006a. 한국어 어휘 교육·학습 자료 개발을 위한 계량적 분석의 한 방향. 어문학 94: 119~146.
- 한영균. 2006b. 《표준국어대사전》의 용례에 대한 사전학적 검토. 국어학 48: 289~312.
- 홍재성 외. 2002. 21세기 세종계획 전자사전 개발분과 연구보고서. 문화관광부.
- Chomsky. 1986. Knowledge of Language, Greenwood, 이선우 역. 2000. 언어지식-그 본질, 근원 및 사용. 아르케.

<참고 사이트>

- 국립국어원. 표준국어대사전.
<http://stdweb2.korean.go.kr/search/View.jsp>

포털사이트 다음 어학사전.

<<http://dic.daum.net>>

연세대학교 언어정보연구원. 연세 현대 한국어사전.

<<https://ilis.yonsei.ac.kr/dic>>

접수 일자: 2016년 05월 24일

수정 일자: 2016년 06월 28일

게재 결정: 2016년 07월 14일