국립국어원 2020-01-16

발간등록번호
11-1371028-000831-01

2020년 어휘의미 말뭉치 연구 분석 사업

연구 책임자

김 일 환

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '어휘의미 말뭉치 연구 분석'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업 기간: 2020년 5월 ~ 2020년 12월

2020년 12월 11일

연구 책임자: 김일환(성신여자대학교)

연구 기관 성신여자대학교 산학협력단
고려대학교 산학협력단
(주) 언어과학
홍익대학교 산학협력단
연구 책임자 김일환
공동 연구원 박진호, 유현조, 윤태진, 이규범, 이도길, 장원철, 정슬아, 정연주

국문 초록

어휘의미 말뭉치 연구 분석

이 사업은 인공지능 발전을 위한 우리말 기초 자원으로 활용될 고품질의 한국어 어휘의미 분석 말뭉치 및 형태 분석 말뭉치를 구축하고, 분석 말뭉치 구축을 위한 표준적인 지침을 개발·정비하는 데 주요 목적이 있다.

사업의 범위는 크게 네 부분으로 나눌 수 있다. 첫째는 어휘의미 분석 말뭉치 구축 지침 수립으로, 2019년도에 마련된 체언류 어휘의미 분석 지침에 더해 용언류 어휘의미 분석 지침을 새로이 마련한다. 둘째는 어휘의미 분석 말뭉치 구축으로, 어휘의미 분석 말뭉치 구축 지침을 바탕으로 총 400만 어절 규모(문어 200만 어절, 구어 100만 어절, 메신저 대화 100만 어절)의 어휘의미 분석 말뭉치를 구축한다. 셋째는 형태 분석 말뭉치 구축 지침 수립으로, 2019년도에 마련된 문어·구어 형태 분석 말뭉치 구축 지침에 더해 메신저 대화 형태 분석 지침을 새로이 마련한다. 넷째는 형태 분석 말뭉치 구축으로, 형태 분석 말뭉치 구축 지침을 바탕으로 총 100만 어절 규모(메신저 대화 100만 어절)의 형태 분석 말뭉치를 구축한다.

○ 용언류 어휘의미 분석 말뭉치 구축 지침 수립

용언류의 어휘의미 분석을 위하여 2019년에 마련한 체언류의 어휘의미 분석 지침을 기반으로 삼되 비유 표현 등에 나타나는 용언의 비유적 의미를 허용하는 방향으로 분석 지침을 마련하였다. 그리고 용언류의 어휘의미를 분석하는 과정에서 활용할 수 있는 <우리말샘>의 뜻풀이와 예문, 문장 구조, 공기하는 체언의 의미 부류 등의 다양한 기준을 지침에 명시하였다.

이렇게 수립된 체언류와 용언류의 어휘의미 분석 말뭉치 구축 지침은 메신저 대화 말뭉치에 나타나는 체언과 용언을 분석하는 데에도 크게 문제가 되지 않았다. 다만 초성

단어 등과 같은 메신저 대화에 나타나는 다양한 표기형의 분석 방안을 세부 지침으로 추가하였다. 이를 통해 다른 영역의 형태 분석 말뭉치를 대상으로 어휘의미 분석 말뭉치를 만들더라도 올해 마련한 어휘의미 분석 말뭉치 구축 지침이 적용될 수 있음을 확인하였다.

○ 어휘의미 분석 말뭉치 구축

본 사업에서는 2019년에 구축된 국립국어원 어휘의미 분석 말뭉치(300만 어절)에 나타나는 용언류의 어휘의미를 분석하여 분석 대상 범주가 확장된 어휘의미 분석 말뭉치를 구축하였다. 또한 본 사업에서 올해 구축한 메신저 대화 형태 분석 말뭉치를 대상으로 메신저 대화 어휘의미 분석 말뭉치를 구축하였다.

어휘의미 분석 말뭉치 구축은 어휘의미 분석 지침 수립 → 분석 도구(워크벤치) 구현 → 작업 교육 → 의미번호 부착 → 말뭉치 검증 → 최종 결과물 산출의 순으로 이루어졌다.

이 중 말뭉치 검증은 분석이 완료된 어휘의미 분석 말뭉치와 형태 분석 말뭉치에서 무작위로 5,000개 어절을 추출하여 상위 작업자 그룹이 만든 정답 말뭉치와 비교하는 방식으로 진행하였다. 그 결과 문어 어휘의미 분석 말뭉치는 93.01%, 구어 어휘의미 분석 말뭉치는 95.41%, 메신저 형태 분석 말뭉치는 99.37%, 메신저 어휘의미 분석 말뭉치 95.96의 일치율을 보였다.

○ 메신저 대화 형태 분석 말뭉치 구축 지침 수립

2019년도에 구축된 메신저 대화 원시 말뭉치를 대상으로 형태 분석을 수행하기 위하여, 2019년도에 마련된 형태 분석 말뭉치 구축 지침을 기반으로 삼되 메신저 대화의 특수성을 고려한 메신저 대화 형태 분석 지침을 새로이 마련하였다.

메신저 대화는 문자를 통하여 이루어진다는 점에서는 문어의 속성을 지니지만 실시간으로 즉각적인 양방향 소통이 일어난다는 점에서는 구어의 속성을 지니는데, 이에 따라

메신저 대화에는 전형적인 문어나 전사된 구어와 다른 특수한 언어 현상들이 포함된다. 이를 고려하여 본 사업에서는 아래의 세 가지 내용을 골자로 하는 메신저 대화 형태 분석 지침을 마련하였다.

① 원시 말뭉치에 포함된 표지 및 기호의 처리
- 개인정보를 치환한 표지의 처리, 이모티콘의 처리 지침을 명시하였다.

② 메신저 대화에서 자주 나타나는 언어 현상의 처리
- 사전에 등재되지 않은 다양한 의성의태어와 감탄사, 어미, 각종 신어의 처리 지침을 명시하였다.

③ 메신저 대화에서 나타나는 특수한 표기법의 처리
- 초성만으로 표기한 단어, 음절을 첨가하여 장음을 표시한 경우, 표음주의 표기법을 적용한 경우, 오타가 발생한 경우, 하나의 형태 내부에 한글 외의 기호가 삽입된 경우 등의 처리 지침을 명시하였다.

○ 100만 어절 규모의 메신저 대화 형태 분석 말뭉치 구축

본 사업에서는 2019년도에 구축된 메신저 대화 원시 말뭉치를 대상으로 100만 어절 규모의 형태 분석 말뭉치를 구축하였다. 원시 말뭉치의 각 어절을 대상으로 형태를 분리하고 형태 분류 표지(세분류 47종)를 부착하는 작업을 하였는데, 형태 분리의 기준이 되는 단위는 기본적으로 <우리말샘>에 등재된 단어이되, 생산성이 비교적 높은 접사도 분리하는 것을 원칙으로 삼았다. 또한 메신저 대화의 특수성을 고려하여 마련된 메신저 대화 형태 분석 지침의 내용을 적용하여 형태 분석을 수행하였다.

형태 분석 말뭉치 구축은 형태 분석 지침 수립 → 분석 도구(워크벤치) 구현 → 작업 교육 → 자동 형태소 분석 → 분석 오류 수정 → 최종 결과물 산출의 순으로 이루어졌다.

이 중 분석 오류 수정은 3단계로 이루어졌다. 1단계는 작업자가 원시 말뭉치에 대한

자동 형태 분석 결과를 수정하는 단계이다. 2단계는 자동 형태 분석 결과와 작업자의 오류 수정 결과를 비교하며 검수자가 형태 분석 결과를 검수하는 단계이다. 3단계는 전체 작업 결과물에 대해 상위 작업자 그룹이 형태 결합 오류 목록, 어절 분석 중의성 목록 등을 검토하며 오류를 수정하는 단계이다.

본 사업에서는 말뭉치 구축의 편의를 도모하고 정확성을 높이기 위하여 높은 분석 정확률을 갖춘 형태소 분석기(서울대 형태소 분석기)를 사용하였다. 서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다.

한편으로 형태 분석 말뭉치 구축에 최적화된 워크벤치를 사용하였다. 워크벤치에서는 서울대 형태소 분석기의 어절 분석 결과를 보여 주되 그것을 손쉽게 수정할 수 있게 하였고, 드롭다운 선택 방식 및 오류 검사를 통해 입력 오류를 원천적으로 차단함으로써 형태 분석 및 검수의 효율을 높였다.

차례

제1장 서론

1. 사업의 목적

○ 어휘의미 분석 말뭉치 및 형태 분석 말뭉치 분석 지침 정비
○ 문어, 구어, 메신저 대화 어휘의미 분석 말뭉치(400만 어절) 구축
○ 메신저 대화 형태 분석 말뭉치(100만 어절) 구축

이 사업은 인공지능 발전을 위한 우리말 기초 자원으로 활용될 고품질의 한국어 어휘의미 분석 말뭉치 및 형태 분석 말뭉치를 구축하고, 분석 말뭉치 구축을 위한 표준적인 지침을 개발·정비하는 데 주요 목적이 있다.

한국어 처리를 전제로 하는 인공지능 기술의 발전을 위해서는 높은 정확성을 갖춘 대규모의 분석 말뭉치가 요구된다. 이러한 필요성에 따라 2019년에 형태 분석 말뭉치와 어휘의미 분석 말뭉치, 메신저 대화 원시 말뭉치를 비롯한 다양한 종류의 말뭉치가 구축된 바 있다.

본 사업에서는 2019년도 어휘의미 분석 말뭉치, 메신저 대화 원시 말뭉치 구축 사업의 결과물을 기반으로 하여, 한편으로는 품사 면에서, 한편으로는 장르 면에서 확장된 어휘의미 분석 말뭉치와 형태 분석 말뭉치를 구축한다. 즉 2019년도에 체언류를 대상으로 하여 구축한 문어·구어 어휘의미 분석 말뭉치 총 300만 어절을 바탕으로 하여 용언류의 의미 분석이 추가된 어휘의미 분석 말뭉치를 구축함으로써 품사 면에서 확장된 말뭉치를 산출하고, 2019년도에 구축된 메신저 대화 원시 말뭉치 100만 어절을 대상으로 하여 형태 분석 말뭉치를 구축함으로써 장르 면에서 확장된 말뭉치를 산출한다. 이 메신저 대화 형태 분석 말뭉치를 대상으로 체언류, 용언류의 어휘의미 분석도 수행한다.

아울러, 용언류의 어휘의미 분석을 위해, 또 독특한 언어 현상을 많이 보여 주는 메신저 대화의 형태 분석을 위해, 2019년도에 마련된 어휘의미 분석 지침과 형태 분석 지침을 보완한다. 이를 바탕으로 분석의 일관성을 높인 총 500만 어절 규모의 어휘의미 분석 말뭉치와 형태 분석 말뭉치를 구축하며, 이를 통해 공공 자원으로서의 대규모 분석 말뭉치 구축의 기반을 다시금 마련하고자 한다.

2. 사업의 범위

사업의 범위는 크게 네 부분으로 나눌 수 있다. 첫째는 어휘의미 분석 말뭉치 구축 지침 수립으로, 2019년도에 마련된 체언류 어휘의미 분석 지침에 더해 용언류 어휘의미 분석 지침을 새로이 마련한다. 둘째는 어휘의미 분석 말뭉치 구축으로, 어휘의미 분석 말뭉치 구축 지침을 바탕으로 총 400만 어절 규모(문어 200만 어절, 구어 100만 어절, 메신저 대화 100만 어절)의 어휘의미 분석 말뭉치를 구축한다. 셋째는 형태 분석 말뭉치 구축 지침 수립으로, 2019년도에 마련된 문어·구어 형태 분석 말뭉치 구축 지침에 더해 메신저 대화 형태 분석 지침을 새로이 마련한다. 넷째는 형태 분석 말뭉치 구축으로, 형태 분석 말뭉치 구축 지침을 바탕으로 총 100만 어절 규모(메신저 대화 100만 어절)의 형태 분석 말뭉치를 구축한다.

2.1. 어휘의미 분석 말뭉치 구축 지침 수립

<우리말샘>의 의미 체계에 따라 2019년에 마련된 어휘의미 분석 말뭉치 구축 지침의 분석 대상 어휘를 용언류로 확장하는 방향으로 어휘의미 분석 지침을 보완하였고, 메신저 대화 말뭉치에 나타나는 특성에 대한 처리 방안을 마련하는 방향으로 어휘의미 분석 말뭉치 구축 지침을 수립하였다.

2019년도에 마련된 국립국어원 어휘의미 분석 말뭉치 구축 지침은 <우리말샘>의 의

미 체계를 따름으로써 다의어 수준으로 어휘의미를 분석하였다. 또한 형태와 의미 차원을 구분하여 형태 미등재어(777)와 의미 미등재어(888)를 위한 별도의 어휘의미 표지를 마련하고, 말뭉치의 표기 오류를 표시하기 위한 표지(999)를 설정하였다. 또한 다양한 사례를 제시하여 분석 지침의 각 항목을 설명함으로써 실제 구축 과정에서 도출되는 오류와 비일관성을 줄였다.

이에 본 사업에서는 2019년도에 마련된 어휘의미 분석 지침의 기본 원칙과 어휘의미 분석 표지를 유지한 채, 분석 대상 어휘를 용언류로 확장하는 방향으로 어휘의미 분석 지침을 수립하였다. 구체적으로 용언류에 나타나는 형태 미등재어와 의미 미등재어, 표기 오류의 사례를 추가하고, 갈래뜻과 문장 구조, 공기하는 체언의 의미 부류에 근거하여 어휘의미를 분석하는 세부 지침을 마련하였다. 또한 관용 표현 등에 나타나는 용언의 어휘의미를 분석하는 방안을 지침으로 제시하였다.

또한 메신저 대화 말뭉치의 특성, 구체적으로 개인정보를 치환한 표지와 초성 단어 등의 어휘의미를 분석하는 지침을 마련하였다. 또한 메신저 대화 말뭉치에 나타나는 다양한 표기 양상을 표기 오류로 다룰 것인지 비표준어형으로 다룰 것인지를 구분하는 방안을 유형별로 제시하였다. 이 과정에서 질의응답 게시판에 수집된 다양한 사례를 지침에 추가하였다.

2.2. 어휘의미 분석 말뭉치 구축

수립된 어휘의미 분석 말뭉치 구축 지침을 바탕으로 2019년도에 구축된 국립국어원 어휘의미 분석 말뭉치(300만 어절)에 나타나는 용언류의 어휘의미를 분석하여 분석 대상 어휘가 확장된 어휘의미 분석 말뭉치를 구축하였다. 또한 2019년도에 구축된 국립국어원 메신저 대화 원시 말뭉치(100만 어절)를 대상으로 체언류와 용언류의 어휘의미를 분석한 메신저 대화 어휘의미 분석 말뭉치를 구축하였다. 이 과정에서 동일한 어휘의미 분석 말뭉치 구축 지침으로 문어와 구어, 메신저 등의 다양한 유형의 말뭉치의 어휘의미를

분석할 수 있음을 확인하였다.

어휘의미 말뭉치 구축 과정에서 분석 결과의 정확성을 높이고 구축의 편의를 도모하기 위하여 두 개의 고성능 어휘의미 분석 도구를 활용하였다. '서울대 체언 어휘의미 분석기'는 LSTM 기반의 신경망 모델을 통해 『우리말샘』의 단의 명사 346,047개, 다의 명사 센스 185,087개를 대상으로 용례 9만 문장, 88만 어절 규모의 학습 데이터를 학습한 분석기이고, '서울대 용언 어휘의미 분석기'는 세종 형태의미 분석 말뭉치(1,200만 어절)와 2019년도 형태 분석 말뭉치(300만 어절)를 학습 데이터로 만들어진 분석기이다. 이 두 분석기는 <우리말샘>의 의미 식별번호에 대응이 가능하기 때문에 이를 활용하여 작업자에게 초벌 분석을 제공해 주었다.

어휘의미 분석 작업은 2019년도에 개발된 워크벤치에서 이루어졌다. 작업자는 워크벤치에서 서울대 어휘의미 분석기의 분석 결과를 확인하고 드롭다운 방식으로 어휘의미 번호를 선택하였다. 검수자는 워크벤치에서 동일 어휘에 대한 자동 분석기와 작업자의 분석 결과를 확인하고 드롭다운 방식으로 최종 어휘의미 번호를 선택하였다. 이때 분석기와 작업자의 분석 결과가 불일치할 경우 바탕색을 다르게 표시하여 검수 작업의 효율을 높였다. 이처럼 동일한 분석 대상 어휘에 대해 작업자와 검수자가 한 차례씩 의미를 분석하여 직관에 의한 오분석을 최소화하고, 작업자와 검수자의 결과 차이 역시 시각적으로 표시함으로써 상호 간의 피드백이 자연스럽게 이루어질 수 있도록 제작하였다. 또한 작업자와 검수자의 건의 사항을 적극 반영하여 워크벤치 기능을 지속적으로 향상시켰다.

이러한 과정을 거쳐 각 말뭉치에 나타나는 분석 대상 어휘를 대상으로 <우리말샘>의 의미 번호를 부여하는 작업을 하였다. 다만 메신저 대화 말뭉치의 형태 분석 과정에서 분리된 접사 중 일부를 앞말과 뒷말에 통합하는 과정을 거쳤다. 또한 <우리말샘>에 해당 형태가 없는 경우 어휘의미 표지 '777'을, 형태는 있되 해당 의미가 없는 경우 어휘의미 표지 '888'을 부여하여 사전 미등재 유형을 형태와 의미 차원으로 구분하고, 말뭉치 원어절의 오타, 탈자 등에 어휘의미 표지 '999'를 부여하여 표기 오류를 고려하였다.

2.3. 메신저 대화 형태 분석 말뭉치 구축 지침 수립

형태 분석 말뭉치 구축은 기본적으로 2019년도에 마련된 형태 분석 말뭉치 구축 지침에 따라 이루어진다. 2019년도에는 비교적 정제된 문어를 분석하는 지침을 먼저 수립하고, 구어의 분석도 그 지침을 바탕으로 하되 구어에서 나타나는 특수한 언어 현상(준말, 형태 변이 등)을 분석하는 지침을 추가로 마련하여 적용하도록 한 바 있다.

2020년도에 구축된 메신저 대화 형태 분석 말뭉치도 2019년도에 마련된 형태 분석 말뭉치 구축 지침에 따라 이루어짐은 물론이다. 그러나 메신저 대화에는 이모티콘, 초성 단어, 의도적인 표기법 변형, 오타 등이 빈번하게 나타나므로 이러한 표기형의 분석 방법을 별도로 마련할 필요가 있다.

이에 본 사업에서는 2019년도 형태 분석 말뭉치 구축 지침에 '메신저 대화' 부분을 추가하여 원시 말뭉치에 포함된 표지 및 기호의 처리 방법, 메신저 대화에서 자주 나타나는 언어 현상의 처리 방법, 메신저 대화에서 나타나는 특수한 표기법의 처리 방법으로 나누어 메신저 대화 형태 분석 지침을 정리하고 적용하였다.

그리고 형태 분석 질의응답 게시판을 운영하여 자주 질문되는 문제를 해결할 수 있는 설명과 사례를 지침에 추가하며 지침을 보완하였다.

2.4. 메신저 대화 형태 분석 말뭉치 구축

위의 방법으로 보완된 형태 분석 말뭉치 구축 지침을 바탕으로 100만 어절 규모의 메신저 대화 형태 분석 말뭉치를 구축하였다.

말뭉치 구축의 편의를 도모하고 정확성을 높이기 위하여 높은 분석 정확률을 갖춘 형태소 분석기(서울대 형태소 분석기)를 사용하였다. 서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자

료로 삼아 개발한 것이다.

한편으로 형태 분석 말뭉치 구축에 최적화된 워크벤치를 개발하여 사용하였다. 워크벤치에서는 서울대 형태소 분석기의 어절 분석 결과를 보여 주되 그것을 손쉽게 수정할 수 있게 하였고, 오류 검사를 통해 입력 오류를 원천적으로 차단하였다.

이를 바탕으로 원시 말뭉치의 각 어절을 대상으로 형태를 분리하고 형태 분류 표지(세분류 48종)를 부착하는 작업을 하였다. 형태 분리의 기준이 되는 단위는 기본적으로 <우리말샘>에 등재된 단어이되, 생산성이 비교적 높은 접사도 분리하는 것을 원칙으로 삼았다.

제2장 어휘의미 분석 말뭉치의 구축

1. 어휘의미 분석 말뭉치의 구성

1.1. 분석 말뭉치의 규모

이 사업에서 구축한 분석 말뭉치의 총 규모는 500만 어절이다.

말뭉치 종류		분석 대상 어휘	규모	비고
어휘의미 분석 말뭉치	문어	용언류	약 200만 어절	체언류는 2019년도에 구축됨.
	구어	용언류	약 100만 어절	체언류는 2019년도에 구축됨.
	메신저 대화	체언류, 용언류	약 100만 어절	
형태 분석 말뭉치	메신저 대화	전체	약 100만 어절	
총계			약 500만 어절	

<표 2> 2020년도 구축 말뭉치의 구성 및 규모

문어·구어 어휘의미 분석 말뭉치는 2019년도에 구축된 문어·구어 어휘의미 분석 말뭉치를 기반으로 하여 구축되었다. 2019년도에는 체언류(일반명사, 고유명사, 의존명사, 대명사, 수사, 어근)에 <우리말샘>의 의미 번호를 부착하여 문어·구어 어휘의미 분석 말뭉치를 구축하였는데, 2020년도에는 이에 더하여 용언류(동사, 형용사, 보조용언, 긍정지정사, 부정지정사)에도 <우리말샘>의 의미 번호를 부착하여, 결과적으로 체언류와 용언류에 어휘의미 번호가 부착된 문어·구어 어휘의미 분석 말뭉치 총 300만 어절(문어

200만 어절, 구어 100만 어절)을 구축한 것이다.

이에 더해 2019년도에 구축된 메신저 대화 원시 말뭉치를 바탕으로 형태 분석 말뭉치를 구축하고(총 100만 어절), 구축된 말뭉치를 어휘의미 분석용 말뭉치로 변환한 후 체언류와 용언류 모두에 <우리말샘>의 의미 번호를 부착하여 메신저 대화 어휘의미 분석 말뭉치(총 100만 어절)를 구축하였다.

1.2. 분석 대상 어휘의 규모

어휘의미 분석 말뭉치의 분석 대상 어휘는 기능어(function words)를 제외한 내용어(content words) 중 체언류과 용언류를 중심으로 선정하였다. 다시 말해 형태 분석된 어휘 중 일반명사(NNG), 고유명사(NNP), 의존명사(NNB), 대명사(NP), 수사(NR), 어근(XR), 동사(VV), 형용사(VA), 보조용언(VX), 긍정지정사(VCP), 부정지정사(VCN)가 어휘의미 분석 대상이 된다. 다만 문어 말뭉치와 구어 말뭉치에 출현한 체언류의 어휘의미 분석은 2019년도에 이루어졌기 때문에 두 말뭉치의 경우 분석 대상 어휘가 용언류에 한정된다. 분석 대상 어휘의 빈도를 품사별로 정리하면 <표 2>와 같다.[1)]

	문어	구어	메신저	합계	비율(%)
일반명사(NNG)	1,247,008	315,393	161,679	161,679 (1,724,080)	13.47 (51.61)
고유명사(NNP)	202,595	20,882	17,221	17,221 (240,698)	1.43 (7.21)
의존명사(NNB)	181,287	63,400	31,470	31,470 (276,157)	2.62 (8.27)
대명사(NP)	22,903	53,539	29,206	29,206 (105,648)	2.43 (3.16)

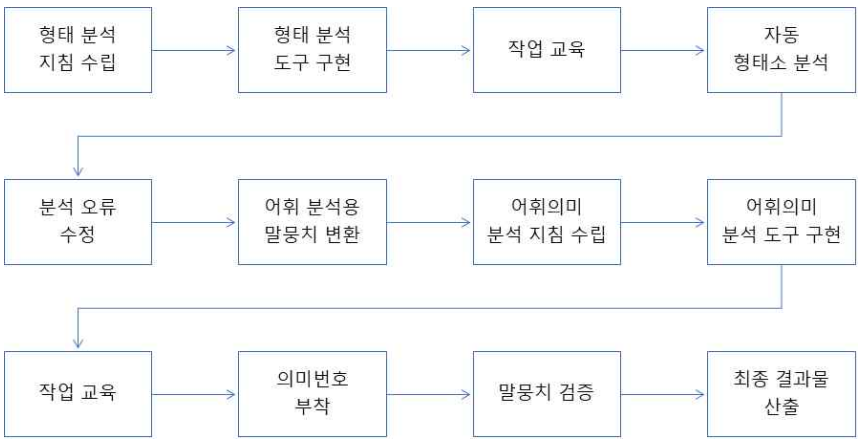
1) <표 2>에 음영 처리된 부분은 2019년도에 분석이 이루어진 분석 대상 품사의 빈도이다. 괄호 안에 적힌 숫자는 구축된 어휘의미 분석 말뭉치 전체를 대상으로 한 빈도와 비율이고, 괄호 없이 적힌 숫자는 2020년도에 작업한 분석 대상 어휘의 빈도와 비율이다.

수사(NR)	18,349	14,207	2,520	2,520 (35,076)	0.21 (1.05)
어근(XR)	374	179	312	312 (865)	0.03 (0.03)
동사(VV)	242,287	161,729	102,358	506,374	42.18 (15.16)
형용사(VA)	71,223	76,078	52,360	199,661	16.63 (5.98)
보조용언(VX)	58,979	39,613	22,460	121,052	10.08 (3.62)
긍정지정사(VCP)	58,202	37,540	22,991	118,733	9.89 (3.55)
부정지정사(VCN)	4,531	5,521	2,209	12,261	1.02 (0.37)
합계	435,222 (2,107,738)	320,481 (788,081)	444,786	1,200,489 (3,340,605)	100.00 (100.00)

<표 3> 어휘의미 분석 말뭉치의 분석 규모

2. 어휘의미 분석 말뭉치 구축 절차

어휘의미 분석 말뭉치의 구축 절차는 다음과 같다.



<그림 2> 어휘의미 분석 말뭉치 구축 절차

2020년도 구축 말뭉치 중 문어·구어 용언류 어휘의미 분석 말뭉치는 이미 어휘 분석용으로 변환된 말뭉치를 바탕으로 구축되므로, <그림 1> 중 '어휘의미 분석 지침 수립' 단계를 시작점으로 하여 말뭉치 구축이 진행된다.

이에 비해 메신저 대화 분석 말뭉치는 형태 분석이 먼저 이루어진 후 그것을 바탕으로 어휘의미 분석 말뭉치를 구축하게 되므로, <그림 1>에 제시된 전 단계를 거쳐 구축된다. 아래에서는 <그림 1>에 제시된 전 단계에 대해 순차적으로 보이고자 한다.

2.1. 형태 분석 지침 수립

어휘의미 분석 말뭉치를 구축하기 위해서는 그에 앞서 형태 분석 말뭉치를 구축해야 한다. 형태 분석 말뭉치 구축의 첫 번째 단계는 형태 분석 지침을 수립하는 것이다. 이미 2019년도에 <21세기 세종계획>의 형태 분석 말뭉치 구축 지침을 수정·보완하여 형태 분석 말뭉치 구축 지침을 수립한 바 있으므로, 그것을 바탕으로 하되 메신저 대화 말뭉치의 형태 분석을 위한 지침을 추가해야 한다.

메신저 대화는 문자를 통하여 이루어진다는 점에서는 문어의 속성을 지니지만 실시간으로 즉각적인 양방향 소통이 일어난다는 점에서는 구어의 속성을 지니는, 전형적인 문어와도 전사된 구어와도 다른 하나의 특수한 장르이다. 이에 따라 메신저 대화에는 전형적인 문어, 전사된 구어와 달리 아래와 같은 특성이 나타난다.

○ 이미지로 된 이모티콘이 사용됨.
○ 한글과 기호를 이용한 이모티콘이 사용됨. 예) ㅠㅠ, ^^;
○ 초성만으로 작성한 단어가 자주 쓰임. 예) ㅇㅇ, ㅋㅋ
○ 의도적이거나 비의도적인 오타가 자주 나타남. 예) 약속이 잇어서
○ 발음 나는 대로 표기하는 경우가 있음. 예) 시러, 조아
○ 의도적으로 표기법을 변형한 경우가 있음. 예) 춧현(추천), 젭알(제발)

○ 한글과 기호를 함께 사용하여 표기한 단어가 있음. 예) 알G, 빠2

‘형태 분석 지침 수립’ 단계에서는 이러한 메신저 대화의 특성을 형태 분석에 어떤 방식으로 반영할 것인지를 결정하여 2019년도 형태 분석 말뭉치 구축 지침을 보완한다.

형태 분석 지침 수립 단계는 형태 분석 작업을 시작하기 전에 우선적으로 수행되어야 할 단계이지만, 이를 바탕으로 분석 오류를 수정하는 단계에서 발생하는 문제들을 반영하면서 반복적으로 수행되어야 할 단계이기도 하다. 본 사업에서는 질의응답 게시판을 운영하여 형태 분석 작업 시 지침으로 해결되기 어려운 부분에 대한 질문을 상시 수합하였으며, 그러한 문제를 해결할 수 있도록 지속적으로 지침을 수정하였다.

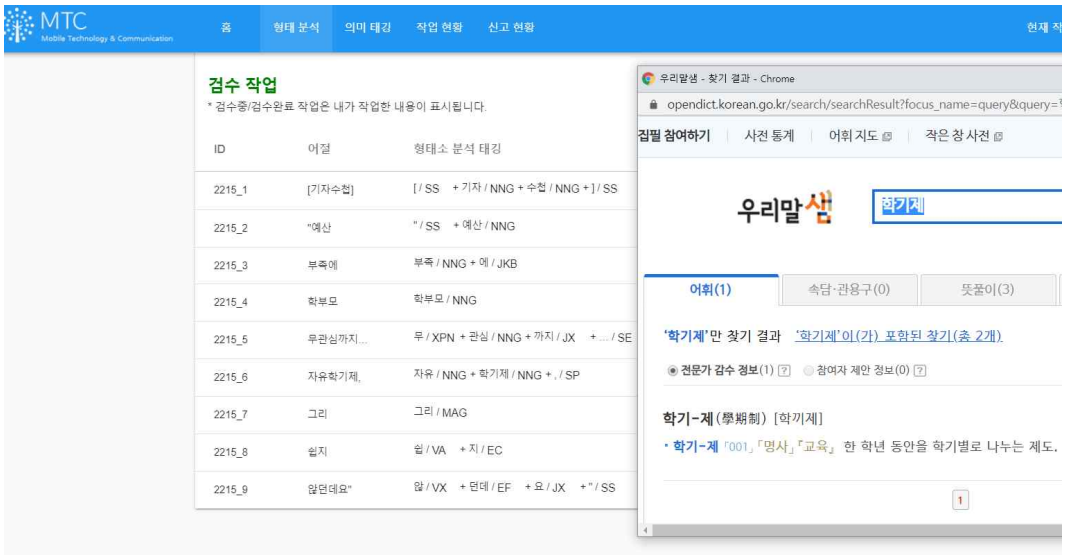
이러한 과정을 통해 마련된 최종 지침의 구체적인 내용과 주요 보완 사항은 제3장에서 보일 것이다.

2.2. 분석 도구(워크벤치) 구현

형태 분석 말뭉치 구축의 두 번째 단계는 형태 분석 오류를 효율적으로 수정하고 작업 진도를 모니터링할 수 있는 분석 도구를 구현하는 것이다.

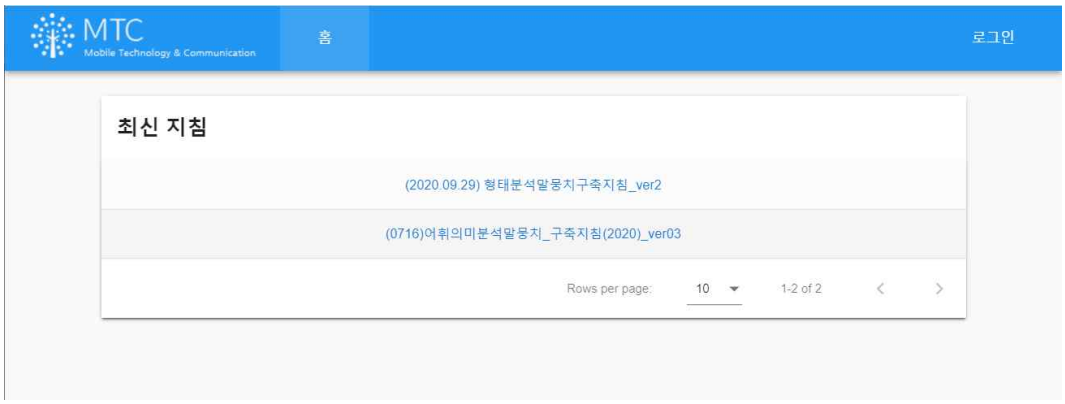
본 사업에서는 형태 분석 오류를 효율적으로 수정하고 작업 진도를 관리하기 위해 웹 기반의 워크벤치를 구축, 활용하였다. 웹 기반의 워크벤치는 작업자들의 동시 접속과 다중 작업 수행을 돕고 <우리말샘>과 연동하여 사전을 쉽게 참조할 수 있게 하는 등 효율적인 수정 작업을 지원하였다. <그림 2>는 작업자 화면에서 ‘학기제’라는 단어에 마우스 포인터를 두었을 때 단어 아래에 돋보기 모양의 아이콘이 생기는 것을 보여 주는데, 이 아이콘을 클릭하면 <그림 3>과 같이 별도 창이 열려 <우리말샘>에서 ‘학기제’를 검색한 결과를 보여 준다.

<그림 3> 워크벤치 작업자 화면과 <우리말샘>의 연동(1)



<그림 4> 워크벤치 작업자 화면과 <우리말샘>의 연동(2)

워크벤치의 홈 화면에는 최신 지침과 작업 시 유의 사항을 제시하여 모든 작업자가 공통된 지침을 확인할 수 있도록 하였다.



<그림 5> 워크벤치의 홈 화면

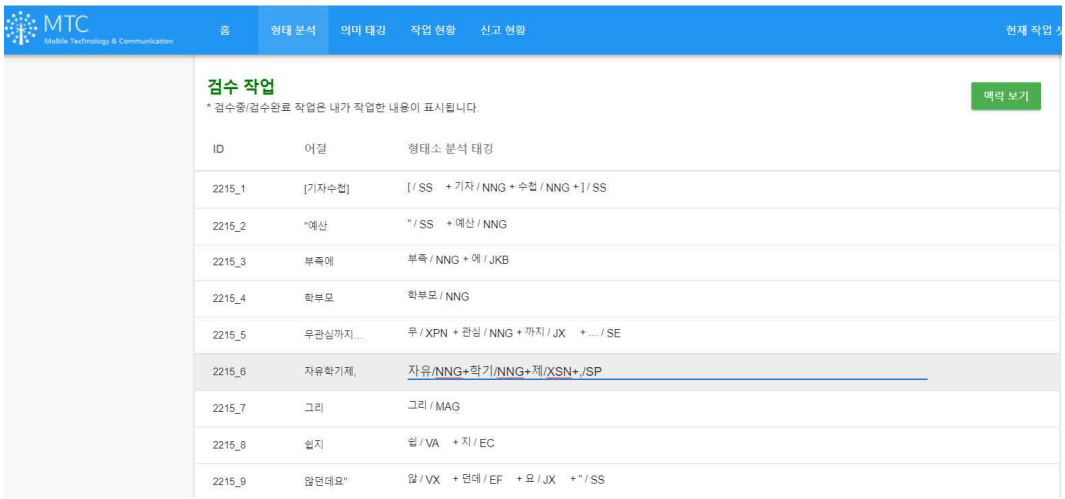
또한 워크벤치 사용자의 역할에 따라 최적화된 기능을 활용할 수 있도록 하였다. 워크벤치 사용자는 작업자, 검수자, 운영자로 나뉘며, 각 부류의 사용자는 담당하는 역할이 다른 만큼 워크벤치에서 활용할 기능도 서로 다르다.

먼저 작업자는 자신이 맡은 작업 분량에 대하여 서울대 형태소 분석기의 어절 분석 결과를 확인하고 그것에서 보이는 오류를 수정하는 역할을 한다. 이에 워크벤치는 작업자들이 맡은 분량에 대하여 서울대 형태소 분석기의 어절 분석 결과를 화면에서 실시간으로 보여주는 역할을 하였으며, 그 분석 결과를 드롭다운 선택 방식 또는 직접 수정 방식으로 수정할 수 있도록 지원하였다. 분석 결과 수정 시 의도치 않은 오류가 발생하는 것을 원천적으로 차단하기 위하여 형태 표지를 수정해야 할 경우에는 드롭다운 메뉴를 통해 수정하도록 하였고, 분석 방식을 수정해야 할 경우에는 직접 입력하여 수정할 수 있게 하되, 형식에 대한 유효성 제약을 두어 잘못된 방식으로 수정되었을 경우 오류 메시지를 산출하도록 하였다.

아래의 <그림 5>는 워크벤치의 작업자 화면에서 형태 표지 부분을 클릭하면 드롭다

운 메뉴가 나오는 것을 보여 준다. <그림 6>은 형태 표지뿐 아니라 분석 방식을 수정할 필요가 있을 때 해당 라인을 더블클릭하여 분석 방식을 직접 수정할 수 있음을 보여 준다. 이때 만약 분석 결과에 형식 오류가 포함되면 오류 메시지가 나오면서 작업 결과를 제출할 수 없도록 하였다.

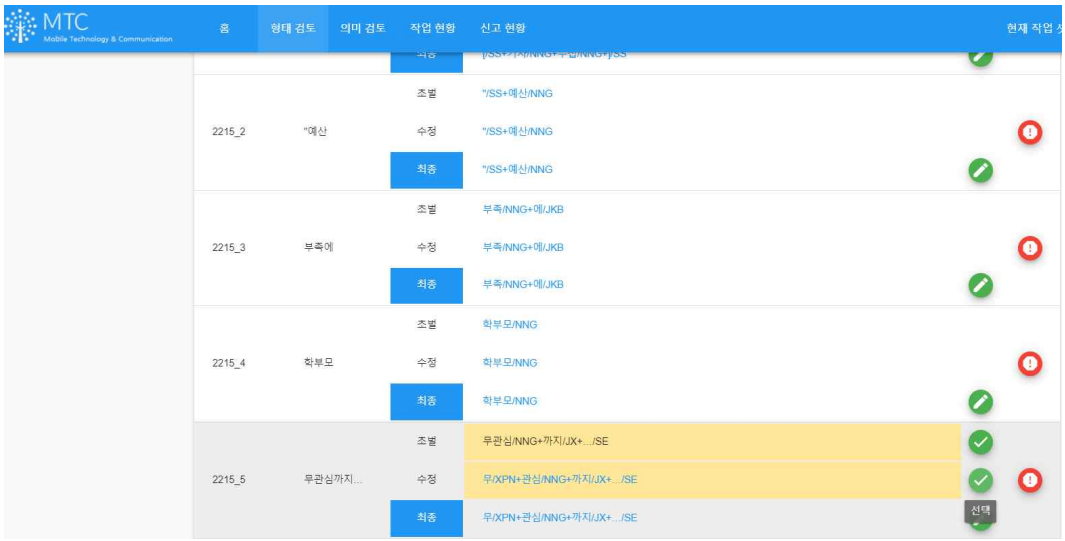
<그림 6> 작업자 화면의 드롭다운 메뉴



<그림 7> 작업자 화면에서의 형태 분석 결과 직접 수정

다음으로 검수자는 작업자의 형태 분석 수정 결과를 검토하고 그것에서 보이는 오류를 수정하는 역할을 한다. 워크벤치는 동일 어절에 대한 자동 분석기의 초벌 분석과 그에 대한 작업자의 수정 결과를 비교하여 보여주었고, 그 중 올바른 분석을 선택하는 방식으로 손쉽게 검수할 수 있도록 하였다. 자동 분석기의 초벌 분석과 작업자의 수정 결과 모두에 오류가 있을 때에는 검수자가 직접 입력하여 분석 결과를 수정할 수 있게 하되, 역시 형식에 대한 유효성 제약을 두어 잘못된 방식으로 수정되었을 경우 오류 메시지를 산출하도록 하였다.

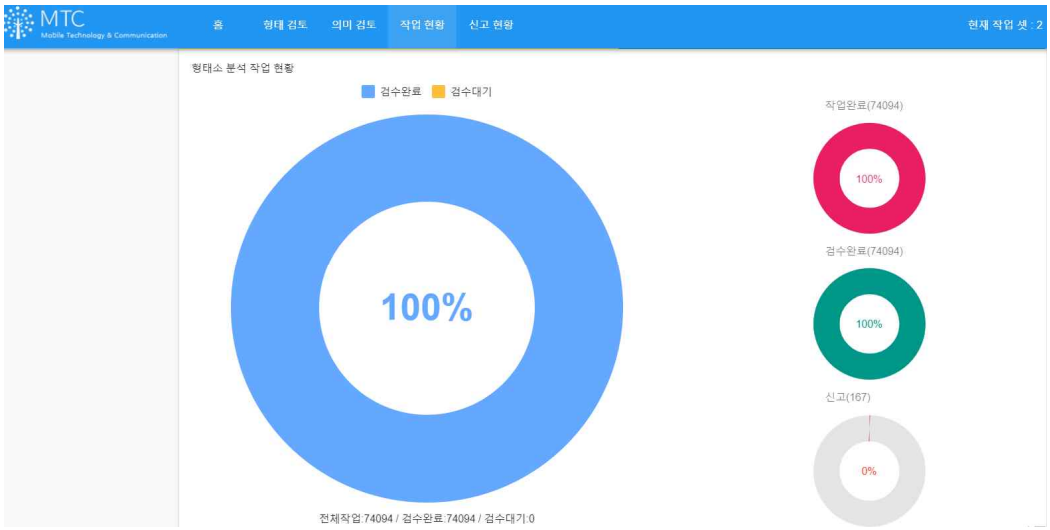
<그림 7>은 워크벤치의 검수자 화면으로, 각 어절에 대한 자동 분석기의 초벌 분석과 그에 대한 작업자의 형태 분석 수정 결과를 비교하여 보여주고 있다. 검수자는 두 분석 결과 중 올바른 분석을 클릭하여 최종 결과로 반영할 수 있다. 만약 두 분석 결과 모두에 오류가 있다면 '최종' 라인을 더블클릭하여 직접 최종 결과를 수정할 수 있는데, 이때 만약 분석 결과에 형식 오류가 포함되면 오류 메시지가 나오면서 작업 결과를 제출할 수 없도록 하였다.



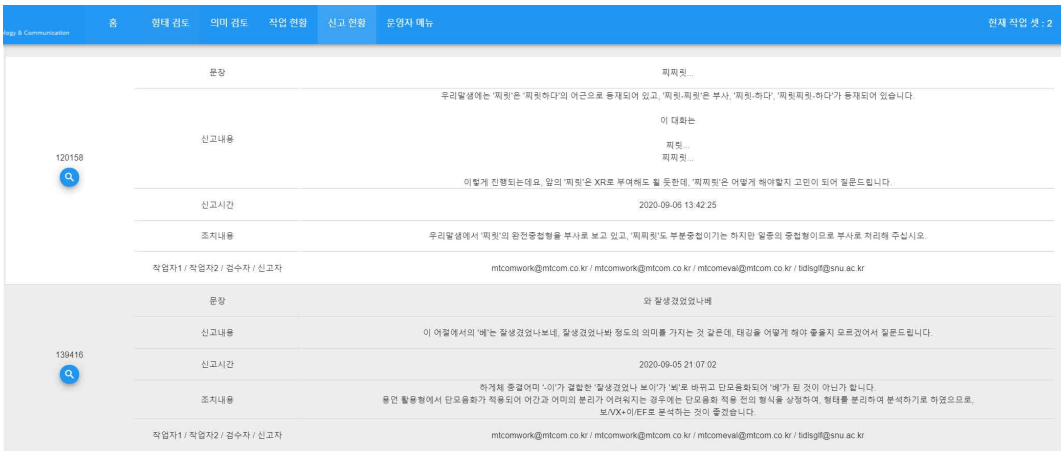
<그림 8> 워크벤치의 검수자 화면

작업자, 검수자의 작업 중 질문이나 논의가 필요한 사안이 생기는 경우에는 각 어절 옆에 있는 신고 버튼을 눌러 해당 어절에 대한 질문·논의 사안을 기록할 수 있게 하였다. <그림 7>에서 보이는 빨간색의 느낌표 아이콘이 신고 버튼에 해당한다.

마지막으로 운영자는 작업자, 검수자의 작업 진행 상황을 모니터링하고 형태 분석 시 발생하는 여러 문제 및 질문 사안을 해결하는 역할을 한다. 이에 워크벤치에 '작업 현황' 탭을 두어 전체 작업자·검수자의 작업 진척 정도 및 개별 작업자·검수자의 작업 진척 정도를 확인할 수 있게 하였다. 또한 '신고 현황' 탭을 두어 작업자, 검수자가 제기한 질문 및 논의 사안을 확인하고, 각 사안에 대한 해결 방법을 입력하여 모든 작업자, 검수자가 그 내용을 확인할 수 있게 하였다. '신고 현황' 탭은 작업자·검수자와 운영자 사이의 질의응답 게시판의 역할을 수행하였으며, 여기에서 확인된 질의응답 내용을 바탕으로 형태 분석 말뭉치 구축 지침을 지속적으로 수정하고 보완하였다. <그림 8>은 작업 현황 화면을, <그림 9>는 신고 현황 화면을 보인 것이다.



<그림 9> 작업 현황 화면



<그림 10> 신고 현황 화면

이와 같이 작업자, 검수자, 운영자 각자의 업무를 돕는 웹 기반의 워크벤치를 구현하여 작업의 효율성을 도모하였다. 워크벤치의 기능 및 화면 구성은 형태 분석 작업 진행 과정 중 작업자, 검수자, 운영자의 요청에 따라 지속적으로 수정되고 보완되었다.

2.3. 작업 교육

형태 분석 말뭉치 구축의 세 번째 단계는 형태 분석 말뭉치 구축 지침과 분석 도구 사용에 대한 교육을 실시하는 것이다. 아울러 비밀 유지와 자료 보안, 문서 보안 등과 관련한 보안 교육도 필요하다. 이에 사업 전체 참여자를 대상으로 보안 교육 및 형태 분석 말뭉치 구축 지침, 워크벤치 사용에 대한 교육을 실시하였다.

특히 형태 분석 말뭉치 구축 지침에 대한 교육은 1회의 교육만으로는 불충분하며 수시로 개별 작업자의 작업 수행에 대한 피드백이 이루어져야 한다. 본 사업의 형태 분석 작업은 작업자가 수행한 형태 분석 수정 결과를 검수자가 검토하는 방식으로 이루어졌으므로, 개별 작업자가 빈번히 만들어 내는 오류를 검수자가 파악하여 지속적으로 피드

백하였다. 또한 전체 분석 결과의 오류와 일관성을 검토하는 최종 검수 단계에서 작업자들이 공통적으로 만들어 내는 오류를 파악하여 수시 교육을 실시하였다.

2.4. 자동 형태소 분석

형태 분석 말뭉치 구축의 네 번째 단계는 자동 형태소 분석이다. 작업자는 자동 형태소 분석 결과를 토대로 오류를 수정하게 되며, 따라서 높은 분석 정확률을 갖춘 자동 형태 분석 도구를 사용하는 것이 작업의 효율을 높이기 위한 관건이 된다. 이에 본 사업에서는 높은 분석 정확률을 갖춘 자동 형태 분석 도구(서울대 형태소 분석기)를 사용하였다.

서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다. 딥러닝(특히 LSTM)의 sequence tagging 알고리즘을 적용하였고, 형태소 분절과 품사 태그 부착의 두 단계로 나누어 각 어절을 처리한다. 딥러닝 적용을 위해 형태소 분절 문제를 한글 음절의 유형 분류 문제로 변형하여 접근하였는데, 이를 위해 1200만 어절 형태의미분석 말뭉치 전체에 대한 검토를 통해 총망라적 목록으로서 한글 음절의 200개 유형을 정리하였다.

한편 Mecab 고유명사 사전의 30여만 개 표제어에 대해 다른 품사와 중의성을 야기할 수 있는 문제 단어들을 수작업으로 제외하여 고유명사 사전을 구축하였고, <우리말샘>의 복합명사 중 하이픈(-)으로 연결된 것과 ^으로 연결된 것을 각각 사전으로 구축하여 복합명사의 분절 여부 판단에 활용하였다. 하이픈으로 연결된 것도 중의성 야기 우려가 있는 것들은 수작업으로 검토하여 제외하였다.

이와 같은 과정을 통해 개발된 서울대 형태소 분석기는 문어 자료에 대해 98%의 F1 score를 보였다. 딥러닝 기반의 형태소 분석기이므로, 딥러닝의 훈련 자료가 추가되면 될수록 형태소 분석기의 성능은 높아진다. 이에 본 사업에서는 2019년도에 구축된 형태 분석 말뭉치를 추가 학습 자료로 삼아 형태소 분석기를 다시금 업그레이드하였고, 이를

적용함으로써 형태 분석 작업의 효율성을 더욱 높였다.

2.5. 분석 오류 수정

서울대 형태소 분석기의 어절 분석 결과는 웹 기반 워크벤치의 작업자 화면에서 확인되며, 작업자는 이 분석 결과의 오류를 수정하는 방식으로 형태 분석 말뭉치 구축 작업을 진행하였다.

작업자들은 2인 1조(작업자 1인-검수자 1인)를 이루어 형태 분석의 오류를 수정하였으며, 오류 수정은 다음과 같이 3단계로 이루어졌다.

○ 1단계: 작업자의 자동 형태 분석 오류 수정

각 조는 일정 분량의 원시 말뭉치를 분배받는다. 조에 할당된 원시 말뭉치의 각 어절에 대한 자동 형태 분석 결과가 작업자의 작업 화면에 제시된다. 이를 토대로 작업자는 자동 형태 분석 결과의 오류를 수정하는 작업을 진행한다.

○ 2단계: 작업자의 오류 수정 결과에 대한 검수자의 검수

검수자는 작업자의 오류 수정 결과를 검토하면서 거기에서 발견되는 형태 분석의 오류를 수정한다.

○ 3단계: 전체 작업 결과물에 대한 최종 검수

전체 조의 형태 분석 결과 검수가 끝나면, 공동연구원을 중심으로 한 상위 그룹이 전체 형태 분석 말뭉치에 대하여 형태 결합 방식의 오류(예: 체언에 어미가 결합하는 것으로 분석된 오류)나 원어절과 분석 결과 사이의 자소 불일치 오류가 있는지 검토하여 수정한다. 또한 동일한 형식의 어절을 둘 이상의 방식으로 분석한 어절 중의성 목록을 검토하여 진정한 중의성과 분석 오류로 인한 중의성을 구별하고, 후자에 대해서는 분석 중의성이 발생하지 않도록 수정한다. 또한 형태 분석 지침의 변경 이력에 따라 변경된 지침의 적용 여부를 검토하여 수정한다.

이 중 3단계 최종 검수 과정에 대해 자세히 보이기로 한다. 전체 작업 결과물에는 지침에 대한 숙지 또는 이해의 차이, 개별 작업자의 단순 실수 등으로 인한 오류가 포함되어 있다. 우선 형태 결합 방식에 대한 메타 지식을 토대로 알고리즘을 구성하여 아래와 같은 결합 오류 유형을 추출하고 수정하였다.

2392-432353	228195	알듯...	알/VV+듯/NNB+.../SE
2392-435861	223674	일본가서	일 본/NNP+가/VV+서/EC
2392-439740	298084	해준다는거굼	하/VV+아/EC+주/VX+ㄴ 다는/ETM+거/NNB+굼/EC
2392-445977	440578	솔론데	솔로/NNG+ㄴ 데/EC
2392-445993	440597	팔자누	팔자/NNG+누/EF
2392-446195	426422	개소리야	개 소리/NNG+야/EF
2392-446219	426454	있는거지	있/VV+는/ETM+거/NNB+지/EC
2392-447165	389587	힘들듯	힘 들/VA+듯/NNB

<그림 11> 형태 결합 오류

위는 동사+명사, 명사+어미, 형용사+명사와 같이 문법적으로 올바르지 않은 결합으로 분석된 어절이 있음을 보여 준다. 이처럼 문법적으로 불가능한 방식으로 분석된 어절을 추출하는 알고리즘을 구성하여 형태 결합 오류를 포함한 어절을 자동으로 추출하고 수정하였다. 특히 위의 예에서 '알듯(알+ㄹ+듯)', '솔론데(솔로+이+ㄴ데)', '가서(가+아서)'처럼 생략된 요소를 복원하여 분석해야 하는 어절에서 생략 요소를 복원하지 않은 오류가 발생하기 쉬운데, 형태 결합 오류 어절을 추출하는 알고리즘을 통하여 효과적으로 오류 어절을 발견할 수 있다.

아래는 동일한 형식의 어절을 둘 이상의 방식으로 분석한 어절 중의성 목록의 예를 보인 것이다. 첫 번째 열에는 원어절, 두 번째 열과 네 번째 열에는 원어절의 분석 결과, 세 번째 열과 다섯 번째 열에는 각 분석의 빈도가 제시되어 있다. 이 목록을 통하여, 지침에 따르면 '3/SN+시간/NNB'로 분석되어야 할 것이 '3/SN+시간/NNG'로 분석된 오류를 발견할 수 있다. 이러한 분석 오류로 인한 중의성은 올바른 분석으로 수정함으로써 중의성 없이 하나의 분석 방식으로 통일되도록 하였다.

3시간	3/SN+시간/NNB	32	3/SN+시간/NNG	5
3주	3/SN+주/NNB	8	3/SN+주/NNG	1
3주하	3/SN+주/NNB+하/VV	1	3/SN+주/NNG+하/VV	1
3천	3/SN+천/NR	12	3/SN+천/SN	1
3천언	3/SN+천/NR+언/NNB	2	3/SN+천/SN+언/NNB	1
3키로	3/SN+키로/NNB	4	3/SN+키로/NNG	1
3학년되	3/SN+학년/NNG+되/VV	1	3/SN+학년/NNG+되/XSV	1

<그림 12> 어절 분석 중의성 검토 자료

본 사업에서는 위와 같은 자동 형태 분석과 3단계의 오류 수정 공정을 거쳐 100만 어절 규모의 메신저 대화 형태 분석 말뭉치를 구축하였다.

2.6. 어휘의미 분석용 말뭉치 변환

어휘의미 분석 말뭉치는 형태 분석 말뭉치 구축 결과를 바탕으로 한다. 그러나 어휘의미 분석의 기본 단위를 <우리말샘>에 등재된 단어로 설정하였기 때문에 형태 분석 과정에서 분석된 일부 접사를 앞말 또는 뒷말에 통합하여 어휘의미 분석에 적합한 말뭉치로 변환하는 과정이 요구되었다. 이러한 접사 처리(통합) 과정을 거쳐 만들어진 말뭉치를 어휘 분석용 말뭉치라 부르는데, 여기에서는 메신저 대화 형태 분석 말뭉치의 접사 처리(통합) 과정을 예로 들어 어휘 분석용 말뭉치로의 변환하는 과정을 설명하겠다.

2.1.1. 통합 접사의 선정

"형태 분석 말뭉치 구축(2019년)"에서 수립한 형태 분석 지침에서 분석 대상으로 삼은 접사의 목록은 다음과 같다.

● 체언접두사			
가(假)	가건물	소(小)	소강당
고(高)	고물가	신(新)	신정당
과(過)	과보호	왕(王)	왕족발
구(舊)	구소련	재(再)	재충전

날	날음식	저(低)	저임금
노(老)	노부부	제(第)	제13차
대(大)	대선배	준(準)	준전시
맏	맏아들	초(超)	초만원
맨	맨몸	최(最)	최고급
무(無)	무의식	친(親)	친러시아
미(未)	미완성	탈(脫)	탈냉전
반(反)	반독재	폐(廢)	폐광산
범(汎)	범세계	풋	풋살구
부(不)	부도덕	피(被)	피고소인
불(不)	불합리	한	한가운데
비(非)	비논리	헛	헛고생
생(生)	생김치		

<표 4> "형태 분석 말뭉치 구축(2019년)"의 분석 대상 접두사 목록

<표 3>에 제시된 것처럼 형태 분석 말뭉치 구축 과정에서 어근과 분리하여 분석한 체언접두사는 총 33개이다. 이들의 경우 어근과 결합한 형태가 어휘의미의 기본 단위인 <우리말샘>에 등재된 단어이기 때문에 접사 처리(통합) 과정이 필요하다. 여기에 메신저 대화 말뭉치의 특성으로 인해 출현하는 다양한 표기형을 고려할 필요가 있다. 실제로 2019년도에 구축된 국립국어원 메신저 대화 원시 말뭉치에는 '생/XPN'이 경음화된 '쌩/XPN'과 '대/XPN'이 문자 모양의 유사성에 의해 변형된 '머/XPN'이 발견된다. 이들 역시 형태 분석 과정에서 접두사로 분석된 요소이기 때문에 접사 처리(통합) 과정에서 뒷말과 결합하였다.

체언접두사의 통합 과정에서 주의해야 할 점은 어근의 품사가 자동으로 승계되지 않는다는 것이다. 예를 들어 '대/XPN+부분/NNG'이 명사로 쓰이기도 하나 부사로 사용되는 경우도 존재한다. 이처럼 전후 맥락에 따라 품사가 달라지는 것에 유의하여 변동된 품사로 재분석해야 한다.

● 명사파생접미사			
가(哥)	김가	분지(分之)	삼분지 일
가(價)	매매가	뻘	조카뻘
가량	1시간가량, 다섯 명가량	산(産)	중국산

간(間)	한 달간	상(上)	역사상
경(頃)	두 시경	생1(生)	갑자생
계(界)	교육계	생2(生)	견습생
계(系)	몽고계	성(性)	인간성
광(狂)	메모광	시(視)	영웅시
권(券)	만 원권	씩	만원씩
권(圈)	운동권	어치	만원어치
권(權)	참정권	여(餘)	삼십여
기(氣)	기름기	용(用)	전쟁용
께	10분께	율(率)	출산율
꼴	십 원꼴	장이	간판장이
꾼	노름꾼	쟁이	심술쟁이
끼리	전우끼리	적(的)	사상적
네	동이네	정(整)	일만 원정
님	선생님	제(制)	봉건제
당(當)	한 사람당	질	서방질
대(臺)	억대	짜리	백 원짜리
댁(宅)	청주댁	째1	이틀째
들	우리들	째2	옹기째
들이	1ℓ들이	쯤	내일쯤
론(論)	비평론	층(層)	선수층
류(類)	자연류	치(値)	기대치
률(率)	경쟁률	치레	인사치레
리(裡)	비밀리	투성이	먼지투성이
발(發)	서울발	풍(風)	복고풍
배기	열 살배기	하(下)	지배하
별(別)	가구별	형(型)	기본형
부(附)	12일부	형(形)	계란형
분(分)	3분의 일	화(化)	도구화

<표 5> "형태 분석 말뭉치 구축(2019년)"의 분석 대상 명사파생접미사 목록

<표 4>은 형태 분석 말뭉치 구축 지침에서 분석 대상으로 삼은 명사파생접미사이다. 이 외에도 메신저 대화 말뭉치의 특성으로 인해 출현하는 표기형들이 있다. 예를 들어 '네/XSN~내/XSN', '들/XSN~덜/XSN~둘/XSN~돌/XSN~드/XSN', '씩/XSN~식/XSN~씬/XSN', '짜리/XSN~따리/XSN~짜이/XSN', '째/XSN~째/XSN', '쯤/XSN~즘/XSN~끔/XSN~쯔음/XSN~쯩/XSN', '님/XSN~니/XSN' '기/XSN~끼/XSN', '적/XSN~석/XSN~젇/XSN~전/XSN'이 실제 메신저 대화 원시 말뭉치에서 발견된다.

이들은 체언접두사와 달리 앞말과의 결합형이 <우리말샘>에 등재될 만한 한 단어를 이루지 못하는 경우가 있다. 이러한 경우는 크게 두 가지로 구분된다. 첫째는 주로 구와 결합하여 앞말과 합쳤을 때 사전에 등재될 만한 한 단어를 이루지 못하는 경우이다. 주

로 통사적 접사가 이에 해당한다.

(1) 주로 구와 결합하여 등재어를 이루지 못하는 경우

가량	다섯 명가량, 10%가량, 30세가량
간(間)	한 달간, 삼십 일간, 주말간
경(頃)	두 시경, 오전 9시경, 16세기경
께	10분께, 이달 말께, 서울역께
꼴	십 원꼴, 100원꼴, 한 명꼴, 열 개꼴
끼리	전우끼리, 우리끼리, 자매끼리
당(當)	한 사람당, 40명당, 일인당
대(臺)	만 원대, 70프로대, 수천억대
들~덜~둘~돌~드	우리들, 친구덜, 그분둘, 눈돌, 인간드
들이	1L들이, 한 말들이, 1리터들이
발(發)	3월 12일발, 열 시발
배기	두 살배기, 다섯 살배기
분지	삼분지 일
생1(生)	1960년 1월 1일생, 이십 년생
씩~식~씬	하나씩, 조금씬, 가끔식
어치	사천원어치, 2처넌어치, 얼마어치
여(餘)	이십여 년, 백여 개, 십오 년여의
정(整)	일만 원정
짜리	얼마짜리, 턴원짜리, 10만따리
째1~쨰	반년째, 두잔째, 5년쨰
쯤~즘~끔~쯔음~쯩	내년쯤, 지금즘, 한번끔, 11시쯔음, 10개쯩
하	독재 정권하, 식민 지배하, 분단 체제하

(1)에 제시된 명사파생접미사 중 '당(當)', '발(發)', '생1(生)'은 어근과의 결합형 일부가 <우리말샘>에 등재되어 있다. 그러나 복합어 자체의 의미 특성이 크게 나타나지 않기 때문에 접사 결합형에 의미 번호를 부여하는 것과 앞말에 의미 번호를 부여하는 것이 별반 다르지 않기 때문에 접사 처리(통합)에서 제외하였다.

다만 '세 살, 다섯 살' 등 수량을 나타내는 구와 결합하는 '-배기'는 접사 처리(통합) 과정에서 제외하나, '나이배기, 알배기, 공짜배기, 대짜배기, 진짜배기' 등 <우리말샘>에 등재된 어휘는 결합형으로 처리한다. 또한 '-째1'은 일반적으로 접사 처리(통합) 과정에서 제외하나, 의존명사 '번'과 결합하는 '-째1'는 앞말과의 결합형 '번째'를 의존명사로 처리하였다.

둘째는 주로 고유명사와 결합하여 일반명사처럼 쓰이나, 결합형이 사전에 등재되어 있지 않는 경우로, 다음과 같은 명사파생접미사가 이에 속한다.

(2) 주로 고유명사와 결합하여 등재어를 이루지 못하는 경우

가(哥)	김가, 이가, 박가
계(系)	중국계, 몽고계, 외국계
네~내	얘네, 언니네, 개내
댁(宅)	안성댁, 광주댁, 상주댁, 철수댁, 참봉댁

(2)에 제시된 '가(哥)', '계(系)', '네', '댁(宅)'은 주로 고유명사와 결합하나, 결합형이 <우리말샘>에 등재되어 있지 않아 특정 의미 번호를 부여하기가 어렵다. 이처럼 의미 번호를 부여하는 데에 효용성이 떨어지는 명사파생접미사는 접사 처리(통합)에서 제외하고자 한다.

이상의 내용을 바탕으로 접사 처리(통합) 과정을 통해 앞말과 결합하여 어휘의미를 분석하는 명사파생접미사는 <표 5>과 같다. 여기에 앞에서 언급한 바와 같이 메신저 대화

말뭉치의 특성으로 인해 나타나는 '님/XSN~니/XSN' '기/XSN~끼/XSN', '적/XSN~석/XSN~젇/XSN~전/XSN' 역시 앞말에 결합하였다.

● 접사 처리(통합) 대상 명사파생접미사 목록			
가(價)	매매가	생2(生)	견습생
계(界)	교육계	성(性)	인간성
광(狂)	메모광	시(視)	영웅시
권(券)	만 원권	용(用)	전쟁용
권(圈)	운동권	율(率)	출산율
권(權)	참정권	장이	간판장이
기(氣)~끼	기름기	쟁이	심술쟁이
꾼	노름꾼	적(的)~석~젇~전	사상적
네	동이네	제(制)	봉건제
님~니	선생님	질	서방질
론(論)	비평론	째2	옹기째
류(類)	자연류	층(層)	선수층
률(率)	경쟁률	치(値)	기대치
리(裡)	비밀리	치레	인사치레
별(別)	가구별	투성이	먼지투성이
부(附)	12일부	풍(風)	복고풍
분(分)	3분의 일	형(型)	기본형
뻘	조카뻘	형(形)	계란형
산(産)	중국산	화(化)	도구화
상(上)	역사상		

<표 6> 접사 처리(통합) 대상 명사파생접미사 목록

● 동사파생접미사	
당하	아군이 공격당하는 데에는 이유가 있다.
되	아침식사가 이미 준비되어 있었다.
시키	오늘 강아지를 운동시키려고 공원에 나갔다.
하	외국에서 공부하는 일이 쉬운 것은 아니다.
받	몇몇은 집제 인상을 강요받았다.
● 형용사파생접미사	
답	사람이 사람답게 행동해야 사람이지,
되	거짓된 말은 들통 나기 마련이다.
롭	어려운 일일수록 슬기롭게 대처하라.
스럽	그녀의 사랑스러운 표정을 보거라.
하	건강한 신체에 건강한 정신이 깃든다.

<표 7> "형태 분석 말뭉치 구축(2019년)"의 분석 대상 용언화 접미사 목록

<표 6>는 형태 분석 말뭉치 구축 지침에서 분석 대상으로 삼은 동사파생접미사와 형

용사파생접미사이다. 이들은 어기 또는 어근에 붙어서 그것을 동사나 형용사로 만들어 주는 기능을 갖는 접미사이다. 이들 역시 메신저 대화 말뭉치의 특성으로 인해 다양한 표기형이 나타난다. '되/XSV~대/XSV', '시켜/XSV~시키/XSV', '하/XSV~허/XSV~햐/XSV~라/XSV~학/XSV~사/XSV~흐/XSV', '롭/XSA~럽/XSA', '스럽/XSA~스러/XSA~습럽/XSA', '하/XSA~허/XSA~라/XSA~히/XSA~사/XSA~학/XSA' 등의 다양한 표기형이 메신저 대화 말뭉치에서 발견되었다.

동사파생접미사와 형용사파생접미사는 접사의 의미가 파생어의 의미에 기여하는 바가 거의 없다. 그러므로 어휘의미 분석 말뭉치의 향후 활용성 등을 감안했을 때 결합형에 어휘의미 번호를 부여하는 것보다 어기 또는 어근의 의미를 분석하는 것이 보다 유의미하다. 따라서 접사 처리(통합) 과정을 거치지 않고 어근(어기)에 해당하는 부분에 어휘의미의 표지를 부여하였다.

2.1.2. 접사 처리(통합) 과정

형태 분석 말뭉치를 어휘 분석용 말뭉치로 변환하기 위해 2.1.1.에서 언급한 통합 대상 접사를 앞말 혹은 뒷말과 결합하는 접사 처리 과정을 진행하였다. 다만 접사 처리(통합) 대상이 되는 접사일지라도 이 과정에서 따옴표나 괄호, 띄어쓰기 등으로 어근과 직접적인 결합이 불가능한 경우 접사 처리(통합) 과정에서 제외하였다.

(3) 따옴표나 괄호, 숫자 등으로 인해 접사 처리(통합) 과정에서 제외된 경우

가. 친(親)시장적　친/XPN+(/SS+親/SH+)/SS+시장적/NNG

나. (재)재협상은　(/SS+재/XPN+)/SS+재협상/NNG+은/JX

다. 비(非)수술적　비/XPN+(/SS+非/SH+)/SS+수술적/NNG

라. 노(老)신사가　노/XPN+(/SS+老/SH+)/SS+신사/NNG+가/JKS

마. 제2차　제/XPN+2/SN+차/NNB

(3)에 나타난 접두사 '친-, 재-, 비-, 노-'는 어휘 분석용 말뭉치에서 어근과 결합해야 하나, 괄호에 의해 뒷말과 직접적인 결합이 불가능하다. '제-'는 역시 숫자와 같은 기호와 결합하여 복합어를 구성하는 것이 불가능하다. 이러한 경우 접사 처리(통합) 과정 없이 체언류를 대상으로 어휘의미를 분석하였다.

이 밖에 접사 처리(통합) 과정에서 유의할 점은 어근의 품사가 자동으로 승계되지 않는 경우가 존재한다는 것이다. 접두사는 일반적으로 어근의 품사를 바꾸지 않으나 '대부분/MAG'처럼 어근에 체언접두사가 결합하여 부사가 되는 경우가 있다. 따라서 접사 통합형의 품사를 결정할 때에는 반드시 문장에서의 쓰임을 고려해야 한다. 아래 (4)은 메신저 대화 말뭉치에서 나타나는 사례를 중심으로 뒷말의 품사가 자동으로 승계되지 않는 경우를 정리한 것이다.

(4) 체언접두사 결합형에서 뒷말의 품사가 자동으로 승계되지 않는 경우

가. 대/XPN+부분/NNG → 대부분/MAG

예. 지금은 대부분 갖춰졌어요.

나. 무/XPN+작정/NNG → 무작정/MAG

예. 그럼 무작정 떠나는 여행은 안좋아해?

다. 무/XPN+조건/NNG → 무조건/MAG

예. 무조건 가야겠네요

라. 한/XPN+바탕/NNG → 한바탕/MAG

예. 한바탕 혼좀 났징.

바. 날/XPN+거/NNB → 날거/NNG

예. 날거를 좋아해요 ㅎㅎㅎㅎ

마. 불/XPN+확실/XR → 불확실/NNG

예. 그래서 약간 불확실 ㅂㄷㅂㄷ

명사파생접미사를 앞말과 통합시킬 때에도 결합형의 품사가 앞말의 품사를 승계하지 않는 경우가 있다. 접사 처리(통합) 과정에서 일반명사 또는 의존명사가 명사파생접미사와 결합할 때 어근의 품사가 유지되는 것이 일반적이나, 일부 복합어에서 품사가 바뀌는 경우가 일부 나타났다.

(5) 일반명사와 명사파생접미사 결합형에서 앞말의 품사가 승계되지 않는 경우

가. 비교/NNG+적/XSN → 비교적/MAG

예. 비교적 얇게 입고 왔어요 ㅋㅋㅋㅋ

나. 사실/NNG+상/XSN → 사실상/MAG

예. ㅋㅋㅋ사실상감시서비스네

다. 인/NNG+용/XSN → 인용/NNB

예. 2인용 갈치에 5만원임...ㅂㄷㅂㄷ

라. 인/NNG+분/XSN → 인분/NNB

예. 보통 몇인분드세요?!

(6) 의존명사와 명사파생접미사 결합형에서 앞말의 품사가 승계되지 않는 경우

가. 내/NNB+적/XSN → 내적/NNG

예. 내적 갈등이 잘 보여지는 거 같아서

이와 달리 고유명사가 명사파생접미사와 결합할 때에는 어근의 품사가 승계되지 않는 것이 일반적이어서 결합형은 대부분 일반명사로 나타났다.

(7) 고유명사와 명사파생접미사 결합형에서 앞말의 품사가 승계되지 않는 경우

가. 미국/NNP+적/XSN → 미국적/NNG

예. 동네구석구석이... 미국적이면서도...

나. 엘사/NNP+용/XSN → 엘사용/NNG

예. 이번 엘사용 드레스나 구두도 엄청 유행하겠다 ㅋㅋㅋ

다. 유럽/NNP+권/XSN → 유럽권/NNG

예. 유럽권은 한번도 안가봤어요!

라. 중국/NNP+산/XSN → 중국산/NNG

예. 중국산 제품도 넘 많아서 ㅜㅜ

마. 발리/NNP+풍/XSN → 발리풍/NNG

예. 정말 발리풍이였는데 ㅎ

이러한 점에 유의하여 메신저 대화 형태 분석 말뭉치를 어휘 분석용 말뭉치로 변환하였다. <표 7>은 접사 처리(통합) 과정을 거쳐 만들어진 어휘 분석용 말뭉치의 일부이다.

형태 분석 말뭉치	쉬는시간에	쉬/VV+ 는/ETM+ 시간/NNG+ 에/JKB
	친구들하고	**친구/NNG+ 들/XSN**+ 하고/JKB
	숨박꼭질하고	**숨박꼭질/NNG+ 하/XSV**+ 고/EC
	딱지치기하고	**딱지치기/NNG+ 하/XSV**+ 고/EC
	하니깐	하/VV+ 니깐/EC
	ㅎㅎ그게	ㅎㅎ/MAG+ 그거/NP+ 이/JKS
	재미있나봐	재미있/VA+ 나/EF+ 보/VX+ 아/EF
	담임선생님은	담임/NNG+ **선생/NNG+ 님/XSN**+ 은/JX
	요즘	요즘/NNG
	별말	별말/NNG
	없고??	없/VA+ 고/EF+ ?/SF+ ?/SF
↓		
어휘 분석용 말뭉치	쉬는시간에	쉬/VV+ 는/ETM+ 시간/NNG+ 에/JKB
	친구들하고	**친구/NNG+ 들/XSN**+ 하고/JKB
	숨박꼭질하고	**숨박꼭질/NNG+ 하/XSV**+ 고/EC
	딱지치기하고	**딱지치기/NNG+ 하/XSV**+ 고/EC
	하니깐	하/VV+ 니깐/EC
	ㅎㅎ그게	ㅎㅎ/MAG+ 그거/NP+ 이/JKS
	재미있나봐	재미있/VA+ 나/EF+ 보/VX+ 아/EF
	담임선생님은	담임/NNG+ **선생님/NNG**+ 은/JX
	요즘	요즘/NNG
	별말	별말/NNG
	없고??	없/VA+ 고/EF+ ?/SF+ ?/SF

<표 8> 어휘 분석용 말뭉치 변환의 예

형태 분석 말뭉치 구축 지침에 의거하여 메신저 대화 말뭉치를 형태 분석한 결과, 접사 '들/XSN, 하/XSV, 님/XSN'이 어근과 분리되었다. 그러나 접사 처리(통합) 과정을 통해 통사적 접사 '들/XSN'과 동사파생접미사 '하/XSV'를 제외한 '님/XSN'을 앞말과 결합하여 형태 분석 말뭉치를 어휘 분석용 말뭉치로 변환하였다.

2.7. 어휘의미 분석 지침 수립

어휘 분석용 말뭉치에 나타나는 분석 대상 어휘에 일관된 어휘의미 표지를 부여하기 위해서는 어휘의미 분석 지침이 수립되어야 한다. 이미 2019년에 <21세기 세종계획>의 형태의미 분석 말뭉치 구축 지침을 수정·보완하여 체언류의 어휘의미를 분석하는 지침을 수립한 바가 있으므로, 그것의 기본 원칙을 유지하며 용언류의 어휘의미 분석을 위한 세부 지침과 메신저의 특성을 반영한 지침을 추가해야 한다.

용언류, 즉 동사(VV), 형용상(VA), 보조용언(VX), 긍정지정사(VCP), 부정지정사(VCN) 등은 홀로 출현하기보다는 체언류와 함께 나타나는 경우가 많다. 따라서 뜻풀이와 예문뿐만 아니라 문장 구조와 공기하는 체언의 의미 부류를 분석 과정에서 어떻게 활용할 것인가에 대한 기준을 세워야 한다. 또한 관용구 등과 같은 비유적 표현에 나타나는 용언의 비유적 의미를 어디까지 인정할 것인가에 대한 기준을 마련하는 방향으로 2019년도 어휘의미 분석 말뭉치 구축 지침을 보완해야 한다.

메신저 대화 어휘의미 분석 말뭉치는 메신저 대화 형태 말뭉치에 나타나는 다양한 표기형들을 어떻게 처리할 것인가에 대한 기준을 제시해야 한다. 즉, 체언류나 용언류로 분석된 초성 단어나 의도적 혹은 비의도적 표기 변형을 얼마만큼 인정할 것인가를 규정하는 방향으로 어휘의미 분석 말뭉치 구축 지침을 보완해야 한다.

본 사업에서는 어휘의미 분석의 기본 원칙과 어휘분석 표지 등 대원칙을 세운 뒤, 어휘의미 분석 과정에서 나타나는 다양한 용례에 대한 세부 지침을 추가하는 방식으로 어

휘의미 분석 지침을 수정하였다. 다시 말해 질의응답 게시판을 통해 분석이 곤란한 어휘를 상시 수합하였고, 그러한 문제를 해결할 수 있도록 지침을 보완하였다. 이러한 과정을 통해 마련된 최종 지침의 구체적인 내용과 주요 보완 사항은 제3장에서 확인할 수 있다.

2.8. 어휘의미 분석 도구(워크벤치) 구현

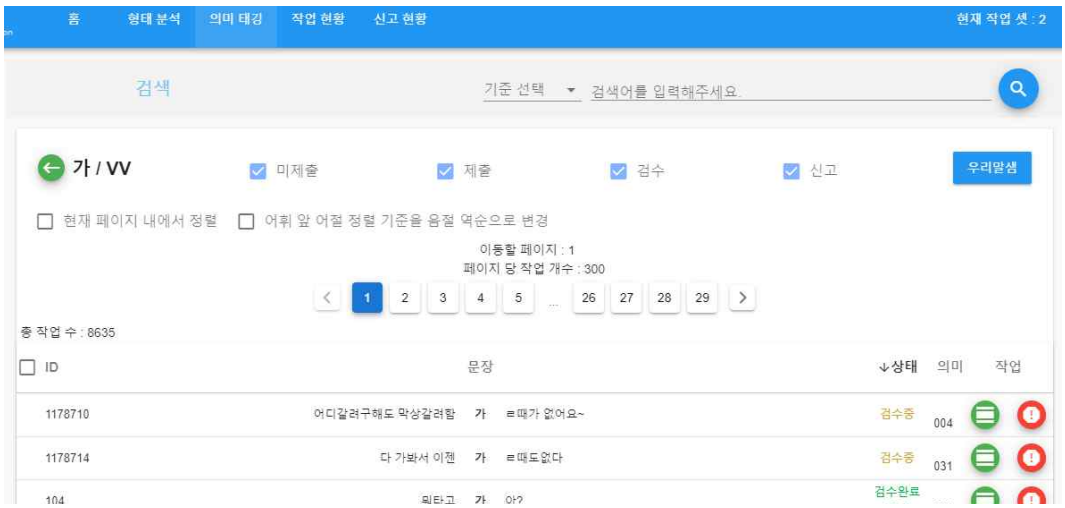
어휘의미 분석 말뭉치 구축 과정에서 어휘의미 분석 오류를 효율적으로 수정하고 작업 진도를 모니터링할 수 있는 분석 도구가 필요하다. 이에 2019년 체언류에 대한 어휘의미 분석 말뭉치를 구축하는 과정에서 개발한 웹 기반의 워크벤치를 지속적으로 사용하여 작업자에게 균등한 분량의 작업을 배분해 주고, 작업자들의 동시 접속과 다중 작업 수행을 도왔다. 또한 <우리말샘>과 연동하여 사전을 쉽게 참고할 수 있게 했고, 다양한 정렬 기능을 제공하여 분석 대상 어휘를 신속하게 검색할 수 있게 하는 등 효율적인 수정 작업을 지원하였다.

작업자는 웹 기반의 워크벤치를 통해 자신에게 배분된 어휘만을 불러오기 때문에 작업 처리 속도를 일정하게 유지할 수 있었고, 분석 결과를 드롭다운 방식으로 선택했기 때문에 오류를 야기하지 않을 수 있었다. 그리고 워크벤치를 통해 분석 결과를 취합하여 그 과정에서 발생할 수 있는 오류를 차단하였다.

워크벤치는 사용자의 역할에 따라 최적화된 기능을 활용할 수 있도록 구성되었다. 따라서 작업자와 검수자, 운영자는 자신의 역할에 따라 활용할 수 있는 기능이 서로 다르다. 우선 작업자는 자신이 의미 분석해야 할 어휘의 형태와 해당 형태의 작업 수를 <그림 12>과 같은 화면을 통하여 확인할 수 있다. 또한 작업자는 작업 선택 화면을 통해 대기, 진행, 완료 등 자신의 어휘별 작업 상태를 확인할 수 있고, '어휘 검색' 등을 통해 작업 어휘를 선택할 수 있다. '어휘, 품사, 총 작업 수, 제출 작업 수, 작업상태' 등 항목별로 정렬 기능을 제공하여 작업 어휘 선택의 편이성을 높였다.

<그림 13> 작업자의 작업 선택 화면

작업자가 작업 선택 화면에서 작업 어휘를 선택하면 <그림 13>과 같은 작업 화면으로 연결된다. 작업 화면에는 동일한 형태의 작업 어휘가 포함된 문장이 제시되어 있고 의미 부분에 자동 분석 결과를 디폴트값으로 제시하여 작업자가 어휘의미 분석기의 분석 결과를 참고할 수 있도록 하였다. 작업자는 작업 화면에 주어진 앞뒤 문맥과 자동 분석 결과를 바탕으로 작업 어휘의 의미를 분석하고 의미 번호를 부여하게 된다.

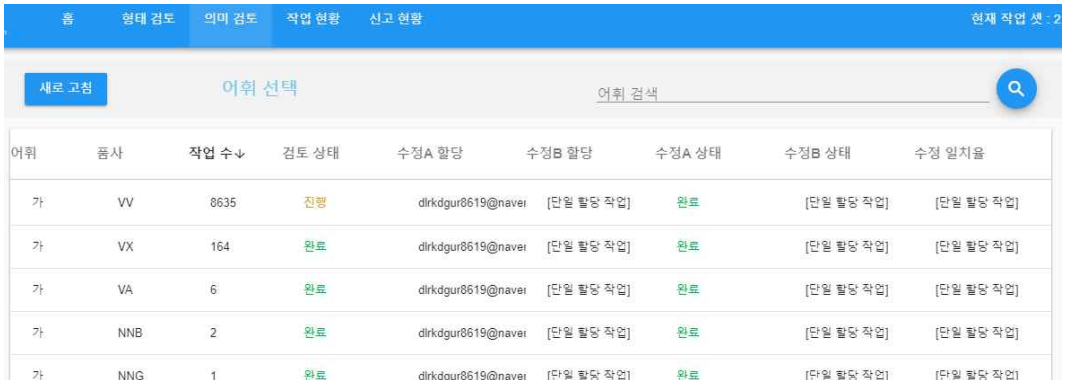


<그림 14> 작업자의 작업 화면

분석 결과는 드롭다운 선택 방식으로 입력하도록 제한함으로써 분석 결과 수정 시 의도치 않은 형식 오류가 발생하는 것을 원천적으로 차단했다. 드롭다운 메뉴의 선택항 역시 <우리말샘>에 등재된 의미 번호와 함께 지침에 의해 규정된 어휘의미 표지 777, 888, 999과 NA로 한정하였다. 이때 공존할 수 없는 어휘의미 표지인 777과 888을 분리하여 제시함으로써 오류의 발생 가능성을 최소화하였다. 여기서 'NA'는 형태 분석 오류로 인해 의미 번호를 부여할 수 없을 때 임시로 사용하는 표지로, 해당 표지가 붙은 어휘들은 형태 분석을 수정한 후 다시 의미 번호를 부여할 수 있도록 작업자에게 재분배하였다.

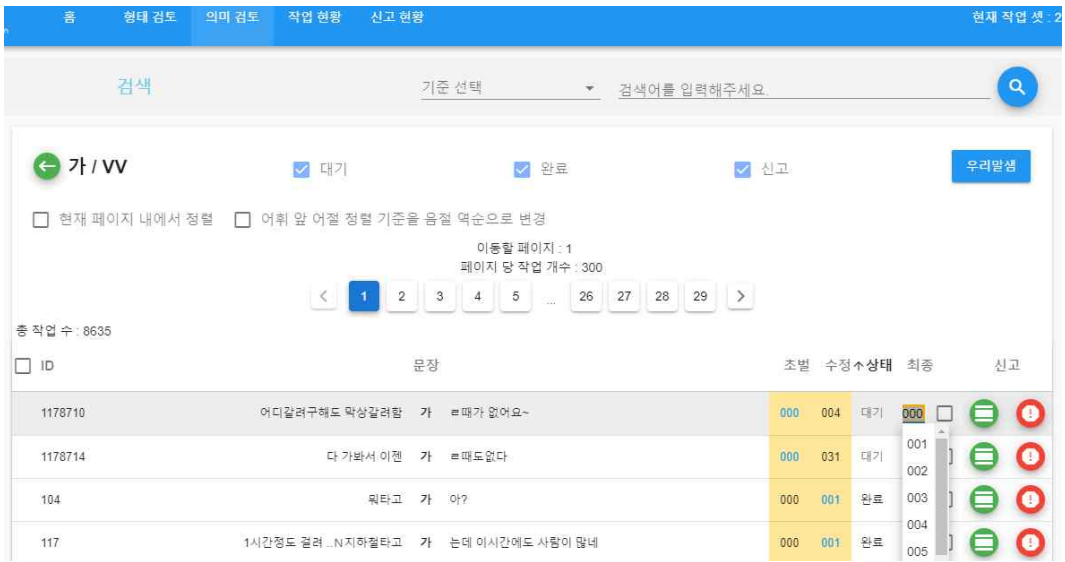
이 밖에 워크벤치는 작업자에게 다양한 기능을 제공한다. 2019년에 개발한 워크벤치와 동일하게 작업자는 '우리말샘'을 통해 <우리말샘>에 접속하고, '맥락보기'를 통해 더 많은 문맥을 확인하며, '신고하기'를 통해 작업 중 발생한 문제를 운영자 그룹에게 알릴 수 있다. 여기에 분석 대상 어휘를 빠르게 찾을 수 있도록 다양한 검색 기능을 추가하였다. 검색 기준을 '문장 ID, 앞 어절, 현재 어절, 뒷 어절, 의미 번호'로 다양화하였고, 좌우 문맥 정렬을 도입하여 비슷한 문형을 모아 분석할 수 있게 하였다. 또한 페이지 당 작업 개수를 '50, 100, 300, 500, 1000' 등 작업자가 선택하고, 여러 페이지를 한꺼번에 이동할 수 있는 기능을 제공하여 작업의 편이성을 향상시켰다.

검수자는 <그림 14>과 같은 검수 선택 화면을 통해 자신에게 분배된 검수 어휘를 확인할 수 있고, '어휘 검색' 등을 통해 검수 어휘를 선택할 수 있다. '검토 상태'에 제시된 대기, 진행, 완료, 작업미완료를 통해 어휘별 검수 상태를 확인할 수 있으며, 작업자를 가리키는 수정A의 상태를 통해 어휘별 작업자의 작업 상태를 점검할 수 있다. 이를 통해 검수자는 작업자를 독려하고 작업 속도를 조절할 수 있다. 작업자의 작업 선택 화면과 마찬가지로 '어휘, 품사, 작업 수, 검토 상태, 수정A 상태' 등 항목별로 정렬 기능을 제공하여 검수 어휘 선택의 편이성을 높였다.



<그림 15> 검수자의 검수 선택 화면

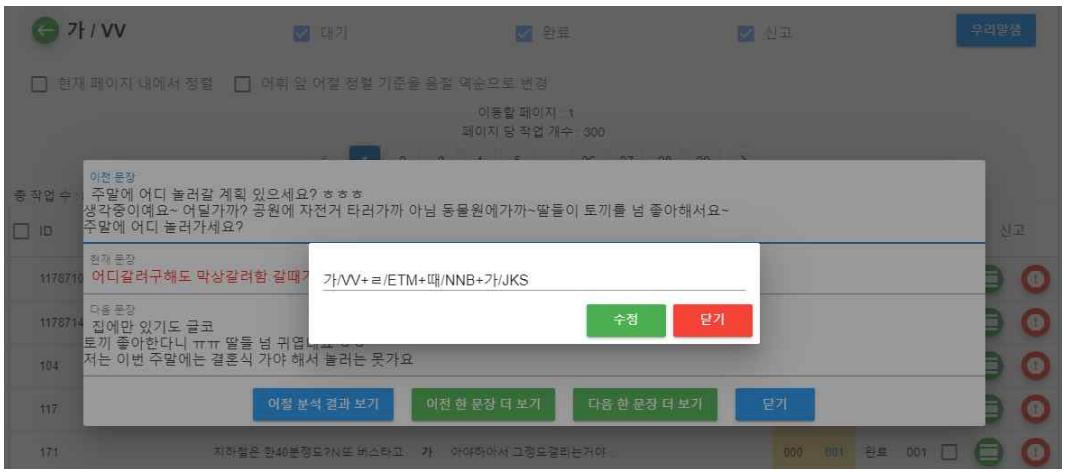
검수자가 검수 선택 화면에서 검수 어휘를 선택하면 <그림 15>와 같은 검수 화면이 나타난다. 검수 화면에는 동일한 형태의 검수 어휘가 포함된 문장이 제시되어 있다는 점에서 작업 화면과 유사하다. 검수자는 문맥과 작업자의 분석 결과를 바탕으로 어휘의 의미 번호를 최종 선택한다.



<그림 16> 검수자의 검수 선택 화면

작업자의 작업 화면과 마찬가지로 검수 화면에 다양한 기능을 제공하였다. 우선 검수 결과를 드롭다운 선택 방식으로 입력하도록 제한하고 드롭다운 메뉴의 선택항 역시 <우리말샘>을 바탕으로 한정하여 검수하는 과정에서 의도치 않은 형식 오류의 발생을 원천적으로 차단하였다.

이 밖에 검수자와 작업자는 '맥락보기'를 통해 화면에 주어진 문장 이상의 문맥을 확인할 수 있다. <그림 16>은 '맥락보기'를 선택했을 때 나타나는 화면이다. 작업 또는 검수 과정에서 작업 또는 검수 화면에 제공된 문장 이상의 맥락이 필요할 때 작업자와 검수자는 '맥락보기'를 통하여 선택 어휘가 포함된 문장의 앞뒤 문장을 확인할 수 있다. 그럼에도 불구하고 어휘의미 분석이 불가능할 경우 '이전 한 문장 더 보기' 또는 '다음 한 문장 더 보기'를 통하여 맥락을 추가적으로 확인할 수 있다. 이에 더하여 2020년에는 '어절 분석 결과 보기' 기능을 추가하였다. 이를 통해 형태 분석 말뭉치를 지속적으로 점검하고 어휘의미 분석 과정에서 파악된 형태 분석 결과의 오류를 바로 수정함으로써 형태 분석 말뭉치의 품질을 향상시켰다.



<그림 17> '맥락보기' 화면

워크벤치를 통해 어휘의미를 분석하는 모든 사용자는 형태 분석 때와 마찬가지로 작업 현황을 통해 작업 상태와 검수 비율을 확인할 수 있고, 분석 어휘와 연동된 <우리말샘>을 통하여 등재 의미를 수시로 검색할 수 있다. 또한 '신고하기' 기능을 통해 어휘의미 분석 및 검수 과정에서 발생하는 이슈를 운영자 그룹과 공유할 수 있다. 이러한 웹 기반의 워크벤치는 사용자의 요청에 따라 지속적으로 보완되어 작업의 효율성을 높였다.

2.9. 작업 교육

어휘의미 분석 말뭉치를 구축하기에 앞서 사업 전체 참여자를 대상으로 어휘의미 분석 말뭉치 구축 지침과 분석 도구인 웹 기반 워크벤치 사용에 대한 교육을 진행하였다. 이와 함께 비밀 유지와 자료 보안, 문서 보안 등의 보안 교육을 실시하였다. 특히 어휘의미 분석 말뭉치 구축 지침은 작업자와 검수자가 생성한 이슈를 바탕으로 지속적으로 보완되기 때문에 수시로 피드백이 이루어져야 한다. 본 사업팀은 주기적으로 어휘의미 분석 결과를 수합하여 형식 오류와 일관성을 검토하고 빈번한 오류를 공유하여 지침 및 교육 자료를 보완하였다.

2.10. 의미 번호 부착

작업자는 워크벤치를 통해 작업 어휘를 할당받고 어휘의미 자동 분석기의 분석 결과를 참고하여 분석 대상 어휘의 의미 번호를 부착한다. 의미 번호는 <우리말샘>에 나타난 표제어 등재 번호와 일치한다. 다만 해당 형태가 <우리말샘>에 검색되지 않을 경우 형태 미등재어로 판단하여 어휘의미 번호 777을 부착하고, <우리말샘>에 형태는 존재하나 해당 의미가 없을 경우 의미 미등재어로 판단하여 어휘의미 번호 888을 부착한다. 예를 들어 "군데군데 시멘트가 덧발려 있었고"의 '덧발리다'는 그 형태가 <우리말샘>에 등록되어 있으므로 777을 부착하고, "그는 신작에 다양한 제주 설화를 끌어들였다"의

'끌어들이다'는 "인용하다" 정도의 의미로 파악되는데, <우리말샘>에는 "남을 권하거나 꾀어서 자기편이 되게 하다"의 의미만 제시되어 있으므로 888을 부착한다. 이러한 과정을 거쳐 분석된 작업자의 결과를 바탕으로 검수자가 다시 한번 어휘의미를 분석한다. 이때도 <우리말샘>의 등재 번호와 777, 888 등을 활용하여 분석 대상 어휘의 의미를 의미 번호로 나타낸다.

본 사업에서는 말뭉치 원어절의 오타, 탈자 등 표기 오류를 표시하기 위해 의미 번호 999를 마련하였다. 이와 관련하여 메신저에 등장하는 다양한 표기형 중 어느 것을 비표준어형으로 판단하고 어느 것을 오표기로 다룰 것인가 하는 문제가 제기한다. 지침에서는 오표기의 다양한 사례를 세분하여 그 기준을 세웠는데 이에 대해서는 3장에서 살피겠다.

2.11. 분석 오류 수정 및 말뭉치 검증

어휘의미 분석 결과에 대한 검수가 끝나면, 공동연구원을 중심으로 한 상위 그룹이 전체 어휘의미 분석 과정에서 신고된 이슈를 검토하여 제대로 분석되었는가를 확인한다. 또한 하나의 형태의 여러 의미 번호가 부여된 경우를 선정하여 의미 분석이 제대로 이루어졌는가를 확인한다. 특히 공존할 수 없는 의미 번호인 777과 888이 동시에 부여된 어휘를 점검하여 잘못 부여된 의미 번호를 수정하였다. 그리고 <우리말샘> 등재 번호와 999가 동시에 부여된 어휘를 대상으로 하여 999가 적절한 분석 결과인지를 확인하였다.

```
개발원/NNG     2    777(10) 888(8)
노노/NNG 2     777(2)   888(2)
당·청/NNG 2    888(5)   777(4)
```

<그림 18> 777-888 부여 적절성 검증

<그림 17>에서 '개발원/NNG', '노노/NNG', '당·청/NNG' 등에 777과 888이 함께 부여된 것을 확인할 수 있다. 이 경우 <우리말샘>에 해당 형태를 검색하여 등재 유무를 확인하였다. '개발원, 노노, 당청'은 그 형태는 <우리말샘>에 존재하나 그 의미가 등재되어 있지 않으므로 888에 해당하는 사례이다. 대부분 작업 시기와 <우리말샘>의 업데이트 시기의 차이로 인해 발생하는데, 결과물 제출 시점을 기준으로 하나의 의미 번호로 수정하였다.

```
중/NNB   6   010(1024)   009(196)    011(71) 012(5)  999(3)  013(2)
가사/NNG 3   011(8)  005(3)  999(1)
가상/NNG 2   005(1)  999(1)
갈/NNG   2   002(2)  999(1)
```

<그림 19> 999 부여 적절성 검증

<그림 18>을 통해 '중/NNB, 가사/NNG, 가상/NNG, 갈/NNG' 등에 <우리말샘> 등재 번호와 999가 함께 부여된 것을 확인할 수 있다. 이 경우 999가 부여된 어형을 확인하여 오표기가 맞는가를 확인하였다. 즉 '중__999/NNB'이 포함된 문장 "팝송만 나오는 중 알았는데"와 '가사/NNG가 분석된 "저와 대화해 주셔서 가사합니다!"를 확인하여 999 부여의 적절성을 검증하였다.

분석 오류 수정을 거쳐 구축된 어휘의미 분석 말뭉치를 대상으로 자체 검증을 실시하였다. 자체 검증은 무작위로 5,000개 어절을 추출하여 상위 작업자 그룹이 만든 정답 말뭉치와 비교하는 방식으로 진행되었는데, 그 결과는 (8)과 같다.

(8) 자체 검증 결과

가. 문어 어휘의미 분석 말뭉치

용언 1188개 중 1105개 일치 = 일치율 93.01%

나. 구어 어휘의미 분석 말뭉치

용언 1548개 중 1477개 일치 = 일치율 95.41%

다. 메신저 형태 분석 말뭉치

5119개 어절 중 5087개 일치 = 일치율 99.37%

라. 메신저 어휘의미 분석 말뭉치

용언 2363개 중 2228개 일치 = 일치율 94.29

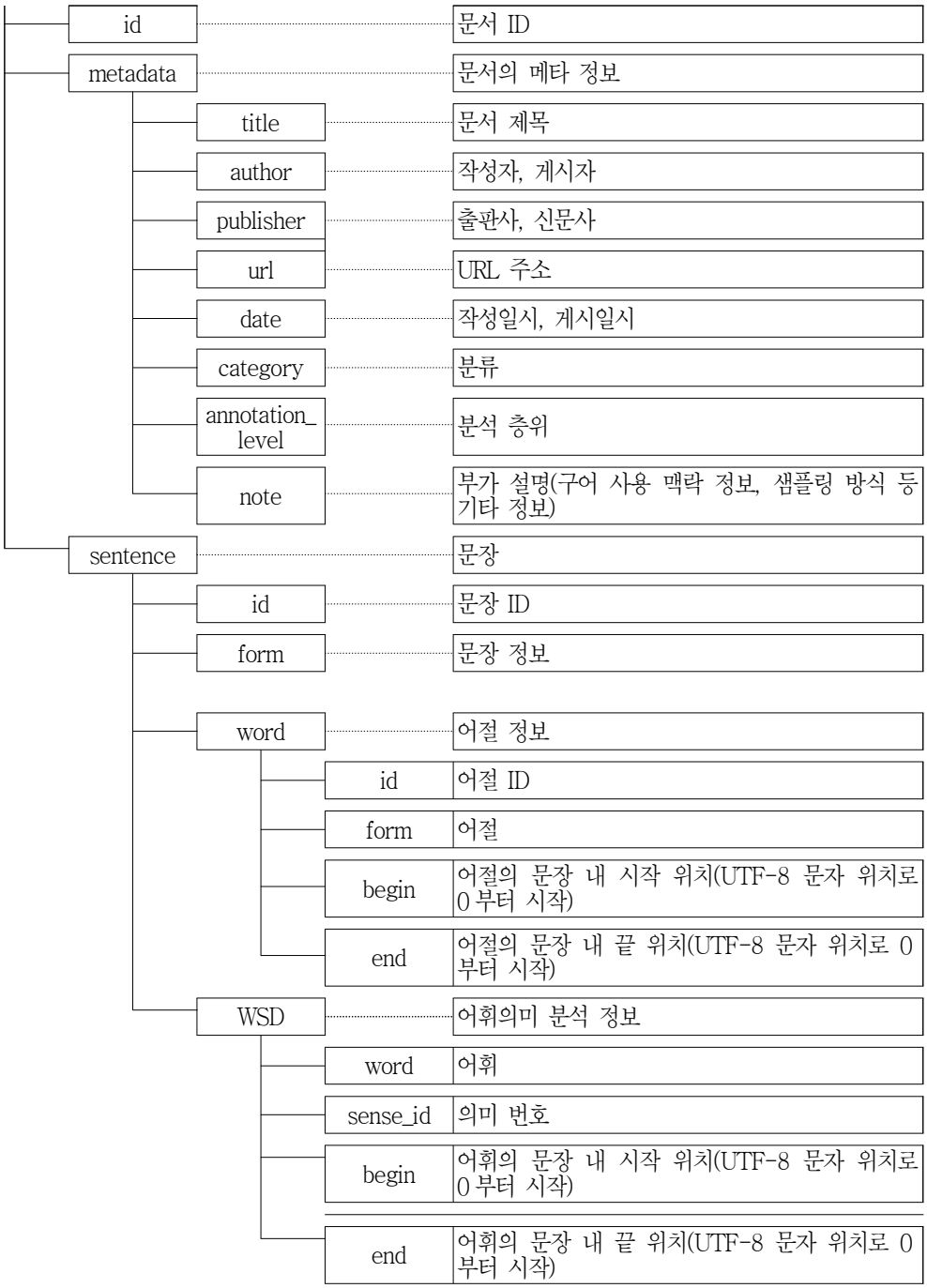
체언 2866개 중 2790개 일치 = 일치율 97.35

전체(용언+체언) 5229개 중 5018개 일치 = 일치율 95.96

2.12. 최종 결과물 산출

어휘의미 분석 말뭉치 구축의 최종 단계는 어휘의미를 분석하고 오류를 수정한 결과를 JSON 형식으로 변환하여 최종 결과물을 산출하는 것이다. <표 8>은 JSON 형식의 기본 구조이고, <표 9>는 JSON 형식으로 전환한 최종 결과물의 일부이다.

1 수준	2 수준	3 수준	4 수준	설명
id				파일 ID
metadata				파일의 메타 정보
	title			파일 제목
	author			작성자, 게시자
	publisher			출판사, 신문사
	year			출판년도
	note			부가 설명(샘플링 방식 등 기타 정보)
document				문서 정보



<표 9> JSON 형식의 기본 구조

```
{
  "id": "MDRW1900000068.1.2.1",
  "form": "그저께부터 name3이 중기이유식 들어가서 하루에 두끼하는데",
  "word": [
    { "id": 1, "form": "그저께부터", "begin": 0, "end": 5 },
    { "id": 2, "form": "name3이", "begin": 6, "end": 12 },
    { "id": 3, "form": "중기이유식", "begin": 13, "end": 18 },
    { "id": 4, "form": "들어가서", "begin": 19, "end": 23 },
    { "id": 5, "form": "하루에", "begin": 24, "end": 27 },
    { "id": 6,  "form": "두끼하는데", "begin": 28, "end": 33 }
  ],
  "morpheme": [
    { "id": 1, "form": "그저께", "label": "NNG", "word_id": 1, "position": 1 },
    { "id": 2, "form": "부터", "label": "JX", "word_id": 1, "position": 2 },
    { "id": 3, "form": "name3", "label": "NNP", "word_id": 2, "position": 1 },
    { "id": 4, "form": "이", "label": "JKS", "word_id": 2, "position": 2 },
    { "id": 5, "form": "중기", "label": "NNG", "word_id": 3, "position": 1 },
    { "id": 6, "form": "이유식", "label": "NNG", "word_id": 3, "position": 2 },
    { "id": 7, "form": "들어가", "label": "VV", "word_id": 4, "position": 1 },
    { "id": 8, "form": "아서", "label": "EC", "word_id": 4, "position": 2 },
    { "id": 9, "form": "하루", "label": "NNG", "word_id": 5, "position": 1 },
    { "id": 10, "form": "에", "label": "JKB", "word_id": 5, "position": 2 },
    { "id": 11, "form": "두", "label": "MMN", "word_id": 6, "position": 1 },
    { "id": 12, "form": "끼", "label": "NNG", "word_id": 6, "position": 2 },
    { "id": 13, "form": "하", "label": "VV", "word_id": 6, "position": 3 },
    { "id": 14, "form": "는데", "label": "EC", "word_id": 6, "position": 4 }
  ],
  "WSD": [
    { "word": "그저께", "sense_id": 1, "pos": "NNG", "begin": 0, "end": 3, "word_id": 1 },
    { "word": "name3", "sense_id": 777, "pos": "NNP", "begin": 6, "end": 11, "word_id": 2 },
    { "word": "중기", "sense_id": 7, "pos": "NNG", "begin": 13, "end": 15, "word_id": 3 },
    { "word": "이유식", "sense_id": 1, "pos": "NNG", "begin": 15, "end": 18, "word_id": 3 },
```

```
      { "word": "들어가", "sense_id": 3, "pos": "VV", "begin": 19, "end": 22, "word_id": 4 },
      { "word": "하루", "sense_id": 1, "pos": "NNG", "begin": 24, "end": 26, "word_id": 5 },
      { "word": "끼", "sense_id": 2, "pos": "NNG", "begin": 29, "end": 30, "word_id": 6 },
      { "word": "하", "sense_id": 4, "pos": "VV", "begin": 30, "end": 31, "word_id": 6 }
    ]
  },
```

<표 10> JSON 형식의 예시

제3장 말뭉치 구축 지침 수립

1. 지침 보완 방향

1.1. 어휘의미 분석 지침 보완 방향

본 사업에서는 2019년도에 마련한 어휘의미 분석 말뭉치 구축 지침에 용언류의 어휘의미 분석을 위한 지침과 메신저 대화에 나타나는 표기적 특징을 처리하는 지침을 추가적으로 마련하였다. 한편으로 2019년도에 마련한 지침을 일부 보완하였는데, 명사류 중심의 예시에 용언류의 예시를 추가하는 등 지침을 보다 명확히 하고 작업의 일관성을 도모하는 데 목적을 두었다.

1.1.1. 용언류의 어휘의미 분석 지침 마련

분석 대상 어휘가 체언류든 용언류든 어휘의미 분석 말뭉치 구축 지침의 기본 원칙와 어휘의미 분석 표지는 동일하다. 다만 용언류의 의미를 분석하는 데 사용될 수 있는 다양한 정보를 세부 지침으로 마련하여 분석 작업의 일관성을 높이고자 한다.

① 갈래뜻 중심의 의미 분석

용언류의 어휘의미 분석 역시 뜻풀이를 중심으로 갈래뜻의 차이를 반영하여 <우리말샘>에 등재된 의미를 최대한 부여하고자 했다. 예를 들어 '기르다'는 <우리말샘>에 "001 【…을】 동식물을 보살펴 자라게 하다", "002 【…을】 아이를 보살펴 키우다", "003 【…을】 사람을 가르쳐 키우다" 등으로 등재되어 있다. 이를 통해 '기르다'를 보살펴 자라게 하는 대상에 따라 구분할 수 있으며, 물리적인 돌봄과 정신적인 가르침 등으로 구

분할 수 있다. <우리말샘>에 나타나는 의미 차를 반영하여 분석 대상 어휘의 의미 번호를 부여하면 예시와 같다.

[예시] 선수단이 직접 닭과 돼지를 기르기도 한다. 기르__001/VV
아이를 낳고 기를 수 있는 사회 환경을 만들어야 한다. 기르__002/VV
광복 이후에는 광주국악원을 만들어 후배를 길렀다. 기르__003/VV

② 예문의 쓰임을 통한 의미 분석

뜻풀이만으로 명확하게 분석되지 않을 경우 예문의 쓰임을 보고 의미 번호를 부여할 수 있다. 예를 들어 '가파르다'는 <우리말샘>에 "001 산이나 길이 몹시 기울어져 있다"로 등재되어 있다. 그러나 몹시 기울어져 있는 대상이 '산'이나 '길'이 아닌 경우에도 해당 의미 번호를 부여할 수 있는지 판단해야 한다. 이때 예문을 참고할 수 있는데, <우리말샘>에 제시된 예문 '이에 따라 휘발유 소비량이 가파른 상승세를 보이고 있습니다.≪MBC 뉴스데스크 1998년 3월≫'를 통해 그래프상의 몹시 기울어져 있는 모습 또한 해당 의미 번호를 부여할 수 있는 것으로 판단된다. 따라서 아래 [예시] 문장에 나타나는 '가파르다'에 '001'을 부여한다.

[예시] 취득세 감면 시한이 다가오면서 가파른 상승곡선을 그렸다. 가파른__001/VA

③ 형태 분석에 근거한 의미 분석

하나의 형태가 동사와 형용사, 보조용언 모두로 쓰일 수 있을 때, 형태 분석에 근거하여 의미 번호를 부여할 수 있다. 예를 들어 '있다'의 경우 <우리말샘>에 동사(001~004), 형용사(005~021), 보조용언(022~023)으로 등재되어 있는데, 형태 분석 결과와 동일한 품사에 해당하는 의미 번호를 분석 대상 어휘에 부여한다. 이 과정에서 형태 분석의 오류를 확인하고 형태 분석이 수정될 수 있도록 피드백을 제공할 수 있다.

[예시] 1시간가량 조용히 있다가 갑자기 일어났다. 있_001/VV

현재 갈등에도 분명 해결책이 있을 것이다. 있_006/VA

서울에 살고 있는 동생 있_023/VX

④ 문장 구조에 의한 의미 분석

분석 대상 어휘가 쓰인 문장의 구조를 파악하여 해당 용언이 요구하는 문장 성분에 따라 어휘의미를 분석할 수 있다. 특히 갈래뜻 사이에 구분이 불분명한 경우 문장 구조에 따라 알맞은 어휘의미 번호를 부여한다. 이때 전후 맥락을 파악하여 생략된 문장 성분을 복원할 수 있다. 예를 들어 '내려오다'는 <우리말샘>에 "001 【…에】 【…으로】 높은 곳에서 낮은 곳으로 또는 위에서 아래로 가다"와 "008 【…을】 높은 곳에서 낮은 곳으로 위치를 옮기다"로 등재되어 있는데, 이 두 갈래뜻 사이에 구분이 불분명하다. 이러한 경우 문장 구조를 통해 어휘의미 번호를 부여할 수 있다. <우리말샘>에 따르면 '001'은 조사 '에'와 '으로'를 논항으로 취하고, '008'의 경우 조사 '을'을 논항으로 취한다. 이를 통해 도착 지점과 함께 쓰이는 '내려오다'는 '001'로, 내려오는 행위가 이루어지는 장소와 함께 나타나는 '내려오다'는 '008'로 분석할 수 있다. 이와 관련하여 도착점을 나타내는 '까지'도 '001'에 포함되고, 행동이 이루어지고 있는 장소와 관련된 '에서'는 '008'에 포함된다.

[예시] 위층 사람들이 아래층에 내려왔다. 내려오_001//VV

산 밑으로 흘러 내려오는 물을 받아 온다. 내려오_001/VV

정상을 거쳐 성판악휴게소까지 내려오는 등산코스 내려오_001/VV

간신히 산을 내려온 이들의 사연을 앞다퉈 보도했다. 내려오_008/VV

집 잘 지어놨으니 산에서 내려오면 들러 달라. 내려오_008/VV

흉기를 든 20대 남성이 계단에서 내려오고 있었다. 내려오_008/VV

⑤ 공기하는 체언의 의미 부류에 따른 의미 분석

문장의 구조로 어휘의미 번호를 확정하기 어려운 경우, 분석 대상 어휘와 공기하는 체언류의 의미 부류에 근거하여 어휘의미를 분석할 수 있다. <우리말샘>에 '없다'는 16개의 의미로 등재되어 있는데, 이중 001과 002는 '없다'가 요구하는 문장 성분 측면에서 동일하기 때문에 문장의 구조로 의미를 분석하기 어렵다. 이러한 경우 공기하는 체언의 의미 부류를 바탕으로 의미를 분석할 수 있다. 다시 말해서 001은 '사람, 동물, 물체' 등 구체적인 대상과 함께 나타나는 데 반해, 002는 '사실이나 현상'처럼 추상적인 것과 주로 어울린다.

[예시] 구속 담당 장관이 없다는 이유로 없_001/VA
처벌 가치가 없는 것으로 판단했다. 없_002/VA

이때 말뭉치에 나타난 분석 대상 용언이 <우리말샘> 뜻풀이에서 제시하고 있는 대상 이외의 체언류와 함께 사용된 경우에는 어휘의미 번호 '888'을 부여한다. <우리말샘>에 '갈아엎다'는 "【…을】 땅을 갈아서 흙을 뒤집어엎다"처럼 흙을 대상으로 풀이되어 있다. 그러나 '판을 갈아엎는'처럼 실제 땅과 흙이 아니라 판, 즉 분위기 정도와 공기하는 '갈아엎다'에는 어휘의미 번호 '888'을 부여한다.

[예시] 판을 완전히 갈아엎는 인사다. 갈아엎_888/VV

다만 뜻풀이에 '따위' 등이 사용되어 대상이 확장될 가능성이 있을 경우에는 해당 의미 번호를 부여한다. 이때 대상의 확장 가능성은 <우리말샘> 어휘지도에 제시된 반의어와 유의어를 통해 확인할 수 있다. <우리말샘>에 등재된 '녹아들다'는 "001 【…에】 다른 물질에 스며들거나 녹아 들어가다"와 "002 【…에】 사상이나 문화 따위가 섞여 어

울리다"로 구분된다. 따라서 얼음이 녹아든 것은 '001'이 되고, 문화가 녹아든 것은 '002'가 되는데, 팀 분위기 등은 '사상이나 문화 따위'에 속하는 것이므로 어휘의미 번호 '002'를 부여할 수 있다.

[예시] 그는 삼성화재의 배구에 녹아들지 못하고 있다. 녹아들_002/VV

⑥ 비유적 표현에 대한 의미 분석

용언류의 어휘의미 분석은 체언류와 달리 관용 표현 등에 나타나는 비유적 의미를 반영하는 방향으로 어휘의미 번호를 부여하였다. 다시 말해서 체언류의 어휘의미 번호는 <우리말샘> 등재 의미를 중심으로 분석했으나, 비유적 표현에 나타나는 용언은 해당 표현의 전체 의미를 고려하여 용언의 어휘의미를 분석하였다. 이때 비유적 표현의 전체 의미는 <우리말샘>의 '속담·관용구'를 참고할 수 있다. 예를 들어 '손을 잡다'는 "서로 힘을 합쳐 협력하다" 정도를 의미하는데, <우리말샘>에 등재된 '잡다'의 다양한 의미 중 "서로 힘을 합쳐 협력하다"의 서술구인 '협력하다'에 해당하는 의미가 없다. 따라서 '잡다'에 어휘의미 번호 '888'을 부여한다.

[예시] 여당과 야당이 손을 잡다. 잡_888/VV

다만 '입을 다물다'처럼 환유적 표현은 <우리말샘>의 해당의미 번호를 부여한다. '입을 다물다'는 "말을 하지 아니하거나 하던 말을 그치다"를 의미하는데, 이는 '입'과 '다물다'의 사전적 의미의 합인 "입을 꼭 맞대다"로, '말을 하지 않는 것'을 의미하는 환유적인 표현이다. 이처럼 환유적인 표현에 해당하는 경우 <우리말샘>에 등재된 어휘의미 번호를 부여한다.

[예시] 그저 입 다물고 사태가 가라앉길 기다리고 있다. [다물_001/VV

1.1.2. 메신저 대화 어휘의미 분석 지침 마련

메신저 대화 어휘의미 분석은 메신지 대화 형태 분석 결과에 근거하기 때문에 형태 분석된 체언류와 용언류에 어휘의미 번호를 부여해야 한다. 이 과정에서 메신저 대화의 특성이 반영된 다양한 표기형들이 체언 혹은 용언으로 분석되는데, 표기 유형별로 어휘의미 분석 지침을 마련하였다.

① 개인정보를 치환한 표지의 경우

개인정보를 보호하기 위해 사용된 표지가 형태 분석 지침에 의해 NNP 또는 NNG로 분석되었기 때문에 어휘의미 분석 말뭉치에서는 이에 대해 어휘의미 번호를 부여해야 한다. 개인정보를 치환한 표지는 발신자에 의한 오표기로 판단할 수 없기 때문에 999를 부여할 수 없다. 따라서 777 또는 888을 부여할 수 있는데, 이들은 <우리말샘>의 검색 결과가 존재하지 않는다는 측면에서 형태 미등재어와 유사하기 때문에 일관되게 어휘의미 번호 777을 부여하였다.

[예시] 저는 affiliation에서 근무해요.	affiliation__777/NNP
아직은 여성 others는 제주에서 저 혼자^^	others__777/NNG

② 초성만으로 표기된 단어의 경우

형태 분석 지침에 의해 체언 혹은 용언으로 분석된 초성 단어는 '777'을 부여하였다. 또한 단어의 일부를 초성으로 표기한 경우도 동일 형태가 사전에 등재되어 있지 않으므로 777을 부여하였다. 다만 1음절 초성 단어의 경우, 'ㄱ__001'과 같이 자모자가 등재되어 있으므로 '888'을 부여하였다.

[예시] ㅇㅈ합니다 ㅇㅈ_777/NNG

나는 ㅋ쟁이거든 ㅋ쟁이_777/NNG

헬스 ㅈ도 모르는게 헬스한다고 가서 ㅈ_888/NNG

③ 접사에 준하는 요소가 결합한 경우

'노', '갓', '충', '개', '핵', '스/쓰' 따위의 접사에 준하는 요소가 결합한 단어는 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여하였다.

[예시] 누가 재미있어해? 궁금쓰 궁금쓰_777/XR

진짜 핵추워 핵춥_777/VA

④ 음절을 첨가하여 장음을 표현한 경우

의미를 강조하기 위해 음절을 첨가하여 장음을 표현한 단어는 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여하였다.

[예시] 수제수제수제버거어 버거어_777/NNG

고오오급 음식이잖아 고오오급_777/NNG

1.1.3. 2019년도 어휘의미 분석 지침 보완

본 사업에서는 분석 대상 어휘의 품사 변경과 말뭉치 유형의 변화에 따른 어휘의미 분석 지침을 마련함과 동시에 분석 대상 어휘 수의 증가로 인한 기존 지침의 미비점을 보완하였다.

① 비표준어형과 오표기 구분의 상세화

2019년도 어휘의미 분석 말뭉치 구축 지침에서 "<우리말샘>에 등재된 표준어에 대한 일반적인 비표준어형으로 판단되는 경우"로 제시된 부분을 상세화하여 오표기로 판단하는 근거와 비표준어형으로 판단하는 근거를 마련하였다.

○ 오표기, 즉 999인 경우
· '고짖, 잇-, 괜찮-'처럼 형태 분석 과정에서 평폐쇄음화와 무관하여 원어절의 표기형 그대로 분석된 단어
· '겨울왕구, 괘낳다'처럼 음운이 탈락한 이유를 설명하기 어려운 단어
· '기뷴, 그짖'처럼 음운이 첨가된 이유를 설명하기 어려운 단어
· '겨울오아국'처럼 자모 배열이 바뀐 단어

○ 비표준어형으로 판단하는 경우, 즉 777 또는 888 등 어휘의미 번호를 부여하는 경우
· '돈까스, 개쫍-'처럼 경음화를 적용한 단어
· '가카'처럼 격음화를 적용한 단어
· '대다나-, 강기'처럼 한국어의 음운 현상으로 설명이 가능한 단어
· '순두부찌게, 잼있-'처럼 표준어형의 발음과 동일한 단어
· '따스허-, 궁금허-'처럼 어휘의 일부가 비표준어형으로 사전에 등재된 단어
  : '따스허다'의 '허다'가 '하다'의 방언으로 등재되어 있으므로 '따스허다' 역시 '따스하다'의 비표준어형으로 판단할 수 있다.
· '이릏-, 선지국'처럼 연관성 있는 어휘가 사전에 등재된 단어
  : "'이렇게'의 방언(전북, 충청)"으로 등재된 '이릏게'를 참고하여 위 예문의 '이릏게'를 '이릏-'의 부사형으로 분석할 수 있으므로 '이릏-'을 비표준어형으로 판단할 수 있다. '선지국' 역시 <우리말샘>에 등재된 '순대국'을 참고하여 해당 어형의 출현을 설명할 수 있으므로 비표준어형으로 판단할 수 있다.

② 분석 대상 어휘 수 증가에 따른 신규 조항

분석 대상 어휘 수가 증가함에 따라 2019년도에 마련한 어휘의미 분석 말뭉치 구축 지침으로는 대응하기 힘든 사례가 나타났다. 이에 관련 지침을 마련함으로써 분석의 일관성을 유지하였다.

○ 외래어인 경우

구어 및 문어 말뭉치에 비해 메신저 대화 말뭉치에는 외래어가 많이 사용되었는데, 대부분 발음에 이끌려 다양한 표기로 나타났다. 예를 들어 '타지마할, 크리스토프'처럼 <우리말샘>의 등재 형태도 있으나 '가니쉬, 룸싸롱'처럼 <우리말샘>의 등재 형태와 다른 표기도 나타났다. 또한 '에픽호이, 키자니아'처럼 <우리말샘>에 미등재된 형태도 등장하는데, 이를 오표기와 비표준어형으로 구분하는 것은 사람마다 차이가 있어 인위적인 기준을 신규 조항을 통해 제시하였다. 즉, <우리말샘>에 등재된 외래어 중에도 규범 표기가 미확정된 것들이 많으므로 외국어의 경우 표준형을 결정하기가 어렵다. 따라서 어휘의미 분석 대상 어휘가 외국어인 경우 음소의 배열 뒤바뀌는 정도가 아니라면 비표준어형으로 판단하여 어휘의미 번호를 부여하였다.

○ 문자 모양의 유사성에 기반한 경우

메신저 대화 말뭉치에는 문자 모양의 유사성에 기반하여 형태를 변형한 단어가 발견되는데, 이 역시 오표기로 판단하지 않고 어휘의미 번호를 부여하였다. 이는 '댕댕이, 머머리' 등이 <우리말샘>에 등재된 것을 고려한 결정으로 문자 모양의 유사성에 기반한 단어들의 추후 등재 가능성을 고려하였다.

○ 참여자 제안 정보인 경우

분석 대상 어휘의 의미가 간혹 <우리말샘>의 '참여자 제안 정보'에 나타나는 경우가 있다. 그러나 본 사업에서는 '전문가 감수 정보'만을 참고하여 의미번호를 부여하였다.

○ 맥락이 불완전한 경우

분석 대상 어휘의 맥락이 존재하지 않아 해당 의미를 분석할 수 없는 경우가 있다. 예를 들어 끝말잇기를 하는 사항에서 나온 단어들은 소리에 이끌린 것이지 특정 의미로 사용된 것이 아니다. 이러한 경우 그 의미 파악이 불가능하므로 기본 의미로 판단되는 001을 부여하였다.

③ 지침의 예시 추가

지침의 예시를 추가함으로써 지침에 담긴 내용을 보다 명확히 하였다.

○ 북한어 및 방언인 경우

메신저 대화 말뭉치에 나타나는 북한어와 방언의 예로 '기_076/NNB('거'의 방언)', '애니_001/VCN('아니다'의 방언)', '뿌수_001/VV('부수다'의 방언)' 등을 추가하였다.

○ 기호가 포함된 경우

메신저 대화 말뭉치에 나타나는 다양한 사례, '당.연_003/NNG', '즐...겁_001/VA', '날씬~하_001/VA'를 추가 제시하였다.

1.2. 형태 분석 지침 보완 방향

본 사업에서는 2019년도에 마련된 형태 분석 말뭉치 구축 지침에 더해 메신저 대화의 형태 분석을 위한 지침을 추가적으로 마련하였다. 한편으로 2019년도에 마련된 지침을 일부 보완하였는데, 설명과 예시를 추가하여 지침에 담긴 내용을 보다 명확히 하고 작업의 일관성을 도모하는 데 목적을 두었다.

1.2.1. 메신저 대화 형태 분석 지침 마련

메신저 대화 자료의 형태 분석 방법은 기본적으로 문어, 구어 자료의 형태 분석 방법과 동일하다. 다만 메신저 대화는 전형적인 문어나 전사된 구어와 구별되는 언어 특성 및 표기 특성을 보여 주므로, 메신저 대화에서 나타나는 특별한 언어 현상을 처리하기 위해, 또 원시 말뭉치에 포함된 특별한 표지를 처리하기 위해 별도의 지침을 마련할 필요가 있다.

본 사업에서는 아래와 같이 세 부분으로 나누어 메신저 대화 형태 분석 지침을 마련하였다. 크게 ①기호의 처리, ②메신저 대화에서 자주 나타나는 언어 현상의 처리(다양한 의성의태어와 감탄사의 출현, 사전 미등재 신조어의 출현 등), ③메신저 대화에서 나타나는 특수한 표기법의 처리(초성만으로 표기한 단어, 표음주의 표기법으로 표기된 경우 등)로 나뉜다.

○ 원시 말뭉치에 포함된 표지 및 기호의 처리
: 줄 바꿈 표지의 처리, 개인정보를 치환한 표지의 처리, 이미지로 된 이모티콘의 처리, 문자나 기호를 사용한 이모티콘의 처리, 이모티콘 외 기호의 처리

○ 메신저 대화에서 자주 나타나는 언어 현상의 처리
: 다양한 의성의태어와 감탄사, 자음이 첨가된 형태, 어미 없이 용언 어간이 단독으로 쓰인 경우, 사전 미등재 어미, 사전에 등재된 요소가 다른 용법으로 사용되는 경우, 미등재어, 외국어, 방언의 처리

○ 메신저 대화에서 나타나는 특수한 표기법의 처리
: 초성만으로 표기한 단어, 음절을 첨가하여 장음을 표시한 경우, 표음주의 표기법을 적용한 경우, 이중모음이 단모음으로 표기되어 형태 분리가 어려운 경우, 오타가 발생한 경우, 탈자로 인해 형태 표지 부여가 어려운 경우, 띄어쓰기 오류로 인해 형태 표지 부여가 어려운 경우, 의미 파악이 어려운 요소, 하나의 형태 내부에 한글 외의

기호가 삽입된 경우, 끊어진 말의 처리

아래에서는 이 세 가지 유형 중 주요 내용에 대한 처리 방법이 어떠한 방향으로 마련되었는지 보이고자 한다.

① 원시 말뭉치에 포함된 표지 및 기호의 처리

- 개인정보를 치환한 표지의 처리

메신저 대화에는 개인의 이름, 계정, 각종 번호, 주소, 소속 등의 개인정보가 노출되는 경우가 있다. 이러한 정보는 개인정보 보호 목적으로 메신저 대화 원시 말뭉치에서 name(이름), account(계정), telnum, cardnum, num(각종 번호), address(주소), affiliation (소속), others(기타)로 치환되어 있다.

이러한 표지에는 본래의 품사를 고려하여 형태 표지를 부여하도록 하였다. 즉 name, account, address, affiliation은 인명, 아이디명, 지명, 회사명 등 본래 고유명사에 해당하는 단어가 치환된 것이므로 NNP(고유명사) 표지를 부여하도록 하였고, telnum, cardnum, num은 숫자가 치환된 것이므로 SN(숫자) 표지를 부여하도록 하였으며, others는 고유명사나 일반명사 등이 두루 치환된 것인데 이 중 NNG(일반명사) 표지를 부여하도록 하였다.

- 이모티콘의 처리

메신저 대화에는 이미지로 된 이모티콘뿐 아니라 문자나 기호를 사용한 다양한 이모티콘이 포함되어 있다. 이러한 이모티콘에는 SW(기타 기호) 표지를 부여하도록 하였다.

[예시] ㅠㅠ/SW, ㅜㅜㅜㅜㅜ/SW, ㅡㅡ/SW, ^ ^/SW, ^ ^*/SW, ㅇㅅㅇ/SW

이모티콘은 다양한 표정과 동작을 묘사하고 있으므로 표정과 동작의 유형에 따라 세부적으로 구분된 태그를 마련하여 구분할 수도 있을 것이나, 본 사업에서는 <21세기 세종계획> 이래로 목록화되어 적용되어 온 형태 표지를 유지함으로써 기구축 말뭉치와의 통일성을 확보하는 것도 중요하다고 판단하여 이모티콘을 위한 세분된 태그를 마련하지 않았다.

② 메신저 대화에서 자주 나타나는 언어 현상의 처리

- 다양한 의성의태어와 감탄사

메신저 대화에는 웃음소리, 울음소리 등의 각종 소리, 그리고 모양을 묘사하는 의성의태어가 다양한 형태로 나타난다. 또한 느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말 등 감탄사도 다양한 형태로 나타난다.

'호호', '하하', '흑흑', '토닥토닥', '덜덜' 등 소리나 모양을 묘사하는 의성의태어는 <우리말샘>에 부사로 등재되어 있으므로, 이에 준하는 미등재어도 부사로 분석하도록 하였다.

또한 '아하', '응', '그래', '아니', '젠장', '빌어먹을', '안녕' 등 느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말은 <우리말샘>에 감탄사로 등재되어 있으므로, 이에 준하는 미등재어도 감탄사로 분석하도록 하였다.

[예시] 흐규흐규/MAG, 확/MAG, 쥬륵/MAG (의성의태어)
[예시] 으아앙 맛있겠다 [으아앙/IC] (감탄사)
왁! 부러워요 [왁/IC+!/SF]

- 자음이 첨가된 형태

메신저 대화 원시 말뭉치에는 '~해용', '~해욤'처럼 주로 어미에 'ㅇ', 'ㅁ'과 같은 자음을 첨가한 형식이 많이 나타난다. 그러한 자음은 앞 형태와 묶어 형태 표지를 부여하도록 하였다.

[예시] 최고징 [최고/NNG+이/VCP+징/EF]
학술대회얌 [학술/NNG+대회/NNG+이/VCP+얌/EF]
아파연(아파요) [아프/VA+아연/EF]

- 어미 없이 용언 어간이 단독으로 쓰인 경우

메신저 대화에는 용언 어간이 어미 없이 단독으로 사용되는 경우가 포함되어 있다. 때로는 용언 어간 중에서도 일부분만이 나타나기도 한다. 한국어의 형태론적 특성에서 벗어나지만, 어절의 뒷부분이 생략되면서 용언 어간만 나타나거나 용언 어간의 일부분만 나타나는 경우가 종종 보이는 것이다. 이러한 현상을 메신저 대화의 한 특성으로 인정하여, 용언 어간만 나타난 경우에는 어미 없이 용언 어간에 형태 표지를 부여하도록 하였고, 용언 어간의 일부만이 나타난 경우에는 그 요소에 XR(어근) 표지를 부여하도록 하였다.

[예시] 고맙. [고맙/VA+./SF] (용언 어간만 나타난 경우)
[예시] 속상.. [속상/XR+../SE] (용언 어간의 일부만이 나타난 경우)
부끄 [부끄/XR]

- 사전 미등재 어미

메신저 대화 원시 말뭉치에는 '그러셈', '뭐하삼', '아니긔' 등에서 볼 수 있는 사전 미등재 어미가 종종 나타난다. 본 분석에서는 이때의 '-셈', '-삼', '-긔' 등을 종결어미로 처리하도록 하였다.

[예시] 아니긔 [아니/VCN+긔/EF]

- 미등재어

메신저 대화 원시 말뭉치에는 준말, 혼성어, 약어 등 <우리말샘>에 등재되지 않은 신조어가 많이 포함되어 있다. 본 분석에서는 아래의 조건에 부합하는 미등재어를 한 단어로 처리하도록 하였다.

첫째, 더 작은 요소로 분리하면 <우리말샘> 미등재어가 도출되는 경우에, 작은 요소로 분리하지 않고 전체를 한 단어로 처리하도록 하였다. 예를 들어 '혼영(혼자 영화)', '당빠(당연 빠따)', '검핑(검정핑크)'의 경우 '혼+영', '당+빠', '검+핑'으로 분리할 경우 <우리말샘> 미등재어가 도출되며 이런 미등재어들은 향후 등재 가능성도 낮다. 이러한 경우 '혼영', '당빠', '검핑'을 더 작은 요소로 분리하지 않고 한 단어로 처리하도록 한 것이다.

[예시] 혼영/NNG(혼자 영화), 혼치킨/NNG(혼자 치킨), 당빠/NNG, 딥빡/NNG, 흠좀무/NNG, 검핑/NNG(검정핑크), 펭하/IC(펭수 하이)

둘째, 전체 단어의 의미가 그것을 구성하는 요소의 의미 합으로 투명하게 설명되지 않는 경우, 더 작은 요소로 분리하지 않고 전체를 한 단어로 처리하도록 하였다. 예를 들어 '곱창떡볶이'는 곱창과 떡볶이를 섞어 만든 음식으로 단순히 '곱창과 떡볶이'가 아니

므로, '곱창+떡볶이'로 분리하지 않고 '곱창떡볶이' 전체를 한 단어로 처리하도록 한 것이다.

[예시] 곱창떡볶이/NNG, 엔빵/NNG, 꿀과목/NNG, 칼약속/NNG

셋째, '노', '갓', '충', '개', '핵', '스/쓰' 따위의 접사에 준하는 요소가 결합한 단어를 한 단어로 처리하도록 하였다. '노', '갓', '충', '개', '핵', '스/쓰'는 본래의 의미에서 다소간 멀어진 채로, 또는 의미가 거의 없는 채로 새로운 단어 형성에 활발히 참여하여 접사에 준하는 요소로 볼 수 있다. 이러한 요소가 결합한 단어를 하나의 파생어처럼 처리하도록 한 것이다.

[예시] 노야근/NNG, 갓겜/NNG, 파밍충/NNG
[예시] 개꿀/NNG, 개빡/NNG, 개빡치/VV+다/EF, 개좋/VA+다/EF
[예시] 핵고통/NNG, 핵춥/VA+다/EF
[예시] 궁금쓰/XR, 동생스/NNG, 대박쓰/NNG

위 예시에서 보듯이 어근이나 명사 뒤에 '스/쓰'가 결합한 형태가 자주 보인다. 이때 어근 뒤에 '스/쓰'가 결합한 말은 그 품사도 어근으로 처리하고, 명사 뒤에 '스/쓰'가 결합한 말은 그 품사도 명사로 처리하도록 하였다.

단, '반갑쓰'처럼 용언 어간 뒤에 '스/쓰'가 결합한 경우에는 '반갑쓰'를 형용사로 처리하기 어렵다는 문제가 생긴다. '반갑쓰'의 문법적 성질에 비추어 볼 때 형용사로 볼 수 없기 때문이다. 따라서 이처럼 용언 어간 뒤에 '스/쓰'가 결합한 경우에는 '스/쓰'를 분석 불능 범주(NA)로 처리하여 '반갑/VA+쓰/NA'와 같이 분석하도록 하였다.

또한 '개', '핵'의 경우 '개 헬'과 같이 뒷말과 공백으로 분리되어 있거나 '핵빵터지다'처럼 한 단어가 아닌 말(빵/MAG+터지/VV+다/EF)과 결합하여 쓰이는 경우가 보인다. 이런

경우에는 '개', '핵'을 부사로 처리하도록 하였다. 결과적으로 '개', '핵'은 한 단어 자격을 갖는 체언, 용언과 붙어서 쓰인 경우에는 접사처럼 취급되고, 그 외의 경우에는 부사로 취급되는 셈이다. 이는 '개', '핵'이 접사에 준하는 요소가 되어 가고 있으나 접사로 지위를 완전히 고정하기는 어려우며 단어와 접사의 중간적인 성격을 지니고 있음을 반영한 것이다.

[예시] 개 헬 [개/MAG]
[예시] 핵빵터짐 [핵/MAG+빵/MAG+터지/VV+ㅁ/ETN]

넷째, 문자 모양의 유사성에 기반하여 변형된 단어는 변형 전 단어의 품사를 고려하여 형태 표지를 부여하도록 하였다. '유래'를 변형한 '윾래', '대한민국'을 변형한 '머한민국' 등이 그러한 예가 되는데, 이때 '윾래'는 '유래'의 품사를 고려하여 일반명사로 처리하고, '머한민국'은 '대한민국'의 품사를 고려하여 고유명사로 처리하도록 한 것이다.

[예시] 윾래/NNG, 머한민국/NNP

메신저 대화에 등장하는 미등재어를 처리하기 위한 더 상세한 지침은 뒤에서 제시될 '형태 분석 말뭉치 구축 지침'에서 확인할 수 있다.

③ 메신저 대화에서 나타나는 특수한 표기법의 처리

- 초성만으로 표기한 단어

메신저 대화에는 단어의 초성만을 표기하는 경우가 자주 나타난다. 이러한 표기도 입력의 경제성을 추구하는 메신저 대화의 특성을 보여 주는 주요 현상으로 인정하여, 본래

단어의 품사에 따라 형태 표지를 부여하도록 하였다.

가령 '흐흐', '키키'를 'ㅎㅎ', 'ㅋㅋ'로 표기한 경우, 소리나 모양을 묘사하는 의성의태어가 <우리말샘>에서 부사로 처리되고 있으므로 그러한 부류의 단어를 초성만으로 표기한 'ㅎㅎ', 'ㅋㅋ'에도 MAG(일반부사) 태그를 부여하도록 한 것이다.

[예시] ㅋㅋ/MAG, ㅎㅎ/MAG, ㄷㄷㄷ/MAG

또한 '아하', '아니', '하이'를 'ㅇㅎ', 'ㅇㄴ', 'ㅎㅇ'로 표기한 경우, 느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말이 <우리말샘>에서 감탄사로 등재되어 있으므로 그러한 부류의 단어를 초성만으로 표기한 'ㅇㅎ', 'ㅇㄴ', 'ㅎㅇ'에도 IC(감탄사) 태그를 부여하도록 하였다.

[예시] ㅇㅎ/IC (아하, 오호), ㅇㅇ/IC (응), ㅇㄴ/IC (아니), ㄴㄴ/IC (노노), ㅎㅇ/IC (하이), ㅂㄱ/IC (방가), ㅂㅂ/IC (바이바이)

이 외에도 초성 단어가 사용되는 경우가 있는데, 모두 본래 단어의 품사에 따라 형태 표지를 부여하도록 하였다.

[예시] ㄹㅇ(레알) 재밌어 [ㄹㅇ/MAG]
[예시] ㅇㅈ(인정) [ㅇㅈ/NNG]
[예시] 두부피만 있어도 ㄱㅊ(괜찮) [ㄱㅊ/VA]

- 음절을 첨가하여 장음을 표시한 경우

메신저 대화는 문자를 통해 이루어지면서도 의사소통 양상은 구어에 가까워, 구어의

음성적 특징을 문자로 담아내려는 노력을 반영하게 된다. 음절을 첨가하여 장음을 표시하는 것도 그러한 노력의 일환으로서 메신저 대화에서 자주 나타난다.

문어 분석을 위한 지침에서는 '그러어엄'처럼 한 어절이 비정상적으로 늘어난 경우 NA(분석 불능 범주)로 처리하도록 하였으나, 메신저 대화 분석을 위한 지침에서는 이러한 장음 표기법을 메신저 대화 표기법의 한 특성으로 인정하고, 장음화된 형태에 장음화되기 전의 형태 표지를 부여하기로 하였다.

[예시] 믓지겠어요오오 [믓지/VA+겠/EP+어요오오/EF]
비켜어어어어어 [비키/VV+어어어어어어/EF]
[예시] 네에~ [네에/IC+~/SO]
오오오오 [오오오오/IC]
아아아아아 [아아아아아/IC]
[예시] 수제버거어 [수제/NNG+버거어/NNG]
[예시] 매에워어(원형: 매워) [매엡/VA+어어/EF]

- 표음주의 표기법을 적용한 경우

메신저 대화에는 형태를 밝혀 적는 규범적 표기법 대신, 소리 나는 대로 적는 표음주의 표기법이 적용된 사례가 많이 포함되어 있다. 이는 메신저 대화의 특성을 반영하는 것으로서 입력의 경제성을 추구하는 현상의 일종이기도 하고 실제 구어 대화와 같은 효과를 주기 위한 것이기도 하다. 그 결과로서 앞말의 끝 자음이 다음 음절의 초성으로 발음되는 연음 현상이 표기에 반영되거나(예: 맞아→마자) 기본형이 아닌 이형태가 표기에 반영된 경우에(예: 싫어→시러), 해당 어절의 형태를 어떤 방식으로 분석할 것인지 결정할 필요가 있다.

그런데 '맞아'와 '마자'는 동일한 언어 기호를 형태주의 표기법으로 적을 것인지 표음

주의 표기법으로 적을 것인지에서 차이를 보인 것일 뿐, 언어 기호 자체가 다른 것은 아니다. '싫어'와 '시러'도 동일한 언어 기호를 형태주의 표기법으로 적을 것인지 표음주의 표기법으로 적을 것인지에서 차이를 보인 것일 뿐 언어 기호 자체가 다르지 않다는 점은 마찬가지이다. 이를 고려하여 본 사업에서는 이처럼 표음주의 표기법이 적용된 경우에도 형태주의 표기법이 적용된 경우와 동일한 방식으로 형태 표지를 부여하기로 하였다.

[예시] 맞아 [맞/VV+아/EF]
마자 [맞/VV+아/EF]
[예시] 싫어 [싫/VA+어/EF]
시러 [싫/VA+어/EF]

다만 분석의 일관성을 위해서는 이러한 처리 방식의 적용 범위를 분명히 제한할 필요가 있다. 이에 표음주의 표기법이 적용된 표기에 대해 형태주의 표기법과 동일하게 형태 표지를 부여하는 것은, 한국어의 필수적인 음운 규칙이 적용된 형태로(경음화 제외) 표기되어 있는 경우에 국한하기로 하였다. 아래와 같은 음운 규칙이 표기형에 반영되어 있을 경우에 음운 규칙 적용 전의 기본형을 밝혀 분석하도록 한 것이다.

첫째, 종성에서의 평폐쇄음화 현상이 표기에 반영된 경우, 평폐쇄음화 이전의 기본형을 밝혀 분석한다.

[예시] 이불덥고(덮고) [덮/VV+고/EC]

둘째, 비음화 현상이 표기에 반영된 경우, 비음화 이전의 기본형을 밝혀 분석한다.

[예시] 수업 끈나써(끝났어) [끝나/VV+았/EP+어/EF]

셋째, 유음화 현상이 표기에 반영된 경우, 유음화 이전의 기본형을 밝혀 분석한다.

[예시] 칼랄(칼날) [칼날/NNG]

넷째, 자음군 단순화 현상이 표기에 반영된 경우, 자음군 단순화 이전의 기본형을 밝혀 분석한다.

[예시] 어이가 업네(없네) [없/VA+네/EF]

다섯째, 용언 어간 말 /ㅎ/ 탈락 현상이 표기에 반영된 경우, /ㅎ/ 탈락 이전의 기본형을 밝혀 분석한다.

[예시] 조아(좋아) [좋/VA+아/EF]

여섯째, 격음화 현상이 표기에 반영된 경우, 격음화 이전의 기본형을 밝혀 분석한다.

[예시] 그러치(그렇지) [그렇/VA+지/EF]

단, 경음화, 구개음화, 첨가 현상이 표기에 반영되어 있는 경우, 또 필수적이지 않은 음운 현상이 표기에 반영되어 있는 경우에는 표기된 형태를 그대로 보존하여 분석한다.

경음화는 유형이 다양하고 그 중에는 규칙적인 경음화도 있지만 예측 불가능한 경우도 많으며, 메신저 대화에서 입력의 비경제성에도 불구하고 경음 표기를 한 데에는 특별한 의도가 있다고도 볼 수 있다. 따라서 경음화 현상이 반영된 표기는 원래 형태로 복원하지 않고 표기형을 그대로 반영하여 분석한다.

[예시] 갈께(갈게)　　[가/VV+ㄹ께/EF]

구개음화가 표기에 반영되는 경우는 '구지(굳이), 가치(같이)' 등 주로 부사 내부에서 나타난다. 이는 표음주의 표기법을 적용한 것이라기보다 '굳다', '같다'와의 의미적 관련성을 인식하지 못한 결과로 볼 가능성이 높으므로, 원래 형태로 복원하지 않고 표기형을 그대로 반영하여 분석한다. 다만, '끝이다'가 'ㄲ치다'로 나타나는 등 형태를 분리해야 하는 부분에서 구개음화가 반영된 경우에는 기본형으로 복원하여 '끝/NNG+이/VCP+다/EF'와 같이 분석한다.

[예시] 가치가자(같이 가자)　　[가치/MAG+가/VV+자/EF]
ㄲ치 없다(끝이 없다)　　[끝/NNG+이/JKS]

사잇소리 첨가나 /ㄴ/ 첨가, /j/ 첨가와 같은 각종 첨가 현상은 필수적이지 않은 음운 현상이므로 표기형을 그대로 반영하여 분석한다. '전화→저놔'에서 볼 수 있는 공명음 사이 /ㅎ/ 탈락 현상, '문법→뭄뻡', '감기→강기'에서 볼 수 있는 양순음화, 연구개음화 현상 등도 필수적이지 않은 음운 현상이다. 이런 음운 현상이 표기에 반영되어 있는 경우에도, 기본형을 복원하지 않고 표기된 형태 그대로를 반영하여 분석한다.

[예시] 다섯시반이여서요(반이어서요)　　[반/NNG+이/VCP+여서/EC+요/JX] (/j/ 첨가)
대다내(대단해)　　[대다나/VA+아/EF] (공명음 사이 /ㅎ/ 탈락)

- 오타가 발생한 경우

메신저 대화에는 의도적이거나 비의도적인 오표기형이 많이 포함되어 있다. 특히 빠르

게 문자를 입력하는 과정에서 나타나는 비의도적 오표기가 빈번하게 나타난다. 본 사업에서는 오표기가 발생했지만 본래 형태가 무엇인지 파악할 수 있는 경우에는, 오표기형을 그대로 두되 본래 형태에 맞는 태그를 부여하기로 하였다.

[예시] 전문가한티(원형: 한테)　　[전문가/NNG+한티/JKB]

그러나 오타로 인해 불완전한 모아쓰기가 이루어진 경우에는, 해당 요소를 포함하여 관련된 형태에 NA(분석 불능 범주)를 부여하도록 하였다. 앞서 본 초성 단어(예: ㅋㅋ)와 같이 의도적으로 초성자만을 사용한 경우가 아니라, 오타로 인해 초성자 또는 모음자만이 남은 경우에 NA 처리를 하도록 한 것이다.

[예시] 핵ㅅ는데(원형: 했는데)　　[하/VV+앜ㅅ/NA+는데/EC]

- 띄어쓰기 오류로 인해 형태 표지 부여가 어려운 경우

띄어쓰기 오류로 인해 형태 표지 부여가 어려워지는 경우에도 각 요소에 NA(분석 불능 범주)를 부여하도록 하였다.

[예시] 보고시 퍼　　[보/VV+고/EC+시/NA, 퍼/NA]

- 하나의 형태 내부에 한글 외의 기호가 삽입된 경우

메신저 대화에는 하나의 형태 내부에 한글 외의 기호가 삽입되는 경우가 종종 포함되어 있다. 이 역시 억양 표시, 표기 변형을 통한 재미 추구 등을 목적으로 하는, 메신저 대화의 장르적 특성에 해당한다.

한글과 기타 기호는 분리하여 분석하는 것이 원칙이지만, 분리할 경우 형태 표지를 부여하기 어려운 요소 또는 어근이 남는 경우가 있다. 그런 경우에는 한글과 기타 기호를 묶은 단위에 형태 표지를 부여하도록 하였다. 가령 아래에 제시한 인사말 '하2'는, 한글과 숫자를 분리할 경우 '하'라는 형태 표지를 부여하기 어려운 요소가 남게 된다. 이런 경우 '하2'를 묶어서 감탄사 표지를 부여하도록 한 것이다.

[예시] 하2 (하이) [하2/IC]

1.2.2. 2019년도 형태 분석 지침 보완

본 사업에서는 메신저 대화의 형태 분석을 위한 지침을 마련하는 한편으로, 2019년도에 마련된 형태 분석 지침을 일부 보완하는 작업도 수행하였다. 2019년도에 마련된 형태 분석 지침은 2019년도에 구축된 형태 분석 말뭉치의 바탕이 되었으므로, 말뭉치 구축 결과물의 일관성을 위해서는 형태 분석 지침의 내용을 되도록 수정하지 않는 편이 바람직하다. 본 사업에서는 기존 형태 분석 지침의 내용을 그대로 유지하되, 형태 분석의 일관성을 도모하기 위해 꼭 필요하다고 생각되는 설명과 예시를 추가함으로써 지침에 담긴 내용을 보다 명확히 하는 방향으로 2019년도 형태 분석 지침을 보완하였다.

보완된 내용의 일부 예를 보이면 다음과 같다.

○ 보조용언(VX) 지침의 주의사항 추가

: <우리말샘>에 등재된 보조용언이 준말 형태로 나타나는 경우, 그 준말 형태도 보조용언으로 분석한다.

[예시] 하고픈 거	[하/VV+고/EC+프/VX+ㄴ/ETM]
가는갑다	[가/VV+는가/EF+ㅂ/VX+다/EF]

○ 접속부사(MAJ) 지침의 주의사항 추가

: 상대방의 말에 맞장구칠 때 쓰이는 '그러니까(요).'도 접속부사로 처리한다. 접속부사로부터 용법의 변화가 발생한 것으로 보이는 사례도 있으나 뒷말이 생략된 것으로 볼 수 있는 사례도 있음을 고려한 것이다.

[예시] A: 날씨가 너무 추워졌어요.
B: 그러니까요. [그러니까/MAJ+요/JX]

○ 감탄사(IC) 지침의 주의사항 추가

: <우리말샘>에 명사로 등재된 단어가 단독으로 쓰여 감탄사와 같은 용법을 보일 때가 있지만 그런 경우에도 <우리말샘>의 품사를 따라 명사로 분석한다.

[예시] 대박! [대/XPN+박/NNG+!/SF]
→ '대박'은 <우리말샘>에 명사로 올라 있으므로, 단독으로 쓰여 감탄사처럼 보이더라도 명사로 분석한다. 이때 '대'가 분리하여 분석해야 할 접두사에 해당하므로 접두사와 명사로 분리하여 분석한다.

<우리말샘>에 감탄사로 올라 있는 단어가 조사나 지정사 앞에서 쓰인 경우, 의미론적인 따옴의 효과가 있는 표현으로 볼 수 있으므로 감탄사로 분석한다. 다만 감탄사로서의 의미와 멀어진 채 조사나 지정사 앞에서 쓰인다면 명사로 분석한다.

[예시] 화이팅이에요. [화이팅/IC+이/VCP+에요/EF+./SF]
→ 감탄사가 지정사 앞에 쓰인 경우이다. 이때에도 감탄사로 분석한다.

[예시] 요즘 컨디션이 메롱이에요. [메롱/NNG+이/VCP+에요/EF+./SF]
→ '메롱'은 <우리말샘>에 놀림의 감탄사로 올라 있으나, 이 예에서는 놀림의 의미에서 멀어진 채 지정사 앞에서 쓰였다. 이런 경우에는 <우리말샘>의 품사와 달리 명사로 분석한다.

○ 동사파생접미사(XSV) 지침의 주의사항 추가

: '말씀드리다', '축하드리다' 등에서 볼 수 있는 '드리다'는 <우리말샘>에 동사 파생 접미사로 올라 있다. 하지만 본 지침에서 분리하여 분석하는 접미사 목록에는 들어 있지 않으므로, 앞에 오는 명사와 묶어서 '말씀드리/VV, 축하드리/VV' 등으로 분석해야 한다. '말씀, 축하, 인사, 감사, 사과' 등 동작을 나타내는 명사 뒤에서 '하다' 대신 공손의 의미를 더하며 쓰이는 '드리다'를 이처럼 앞말과 묶어서 처리함에 유의한다.

이 밖의 자세한 내용은 아래에 제시될 '형태 분석 말뭉치 구축 지침'에서 확인할 수 있다.

2. 어휘의미 분석 말뭉치 구축 지침

이 장에서는 본 사업에서 어휘의미 분석 말뭉치를 구축하기 위하여 적용한 지침의 전모를 보이고자 한다. 이는 2019년도에 마련된 체언류의 어휘의미 분석 지침을 일부 보완하고, 용언류의 어휘의미 분석 지침과 메신저 대화의 어휘의미 분석 지침을 새롭게 추가한 것이다. 지침의 전체 구성은 다음과 같다.

1 기본 원칙

가. 분석대상

어휘의미 분석은 형태 분석 말뭉치의 원문(원어절과 분석 결과 모두)을 가급적 수정하지 않은 채, 체언과 용언을 분석 대상으로 한다. 구체적인 대상 범주는 다음과 같다.

일반명사(NNG), 의존명사(NNB), 고유명사(NNP), 대명사(NP), 수사(NR), 어근(XR),
동사(VV), 형용사(VA), 보조용언(VX), 긍정지정사(VCP), 부정지정사(VCN)

단, 형태 분석 말뭉치에서 발견되는 분석 오류에 대해서는 이를 별도로 기록한 후 형태 분석 말뭉치의 개선에 반영될 수 있도록 한다.

나. 분석원리

어휘의미 분석 말뭉치의 구축 시, 형태 분석 말뭉치를 어휘 분석 말뭉치 수준으로 변환하여 구축하는 것을 원칙으로 한다. 즉, 형태 분석 말뭉치에서 접사 처리(통합) 과정을 거쳐 만들어진 어휘 분석 말뭉치를 대상으로 어휘의미(갈래뜻, sense)를 구별하여 표지(번호)를 부착한다. 단, 통사적 접사를 포함하여 복합어 자체의 의미 특성을 크게 나타내지 않아 어휘의미의 표지를 부여하는 것에 효용성이 떨어지는 접사는 접사 처리(통합) 과정에서 제외한다. 또한 당해의 구축 말뭉치는 향후 활용성을 감안하여 서술성 명사류에 용언을 파생시키는 접미사가 결합된 파생 용언의 경우에는 접사 처리(통합) 과정을 거치지 않고 어근(어기)에 해당하는 부분에 어휘의미의 표지를 부여한다.

[예시] 극은 소극장에서 열리는 콘서트 형식으로 진행된다.
[소/XPN+극장__001/NNG+에서/JKB] (×) [소극장__001/NNG+에서/JKB] (○)
도시형 생활주택에 대한 관심이 높아지고 있다.
[도시__004/NNG+형/XSN] (×) [도시형__001/NNG] (○)
로이터통신의 서울발 기사였다.
[서울발__001/NNG] (×) [서울__002/NNG+발/XSN] (○)
주중에는 공부하고 주말에는 야구하자.

[공부하__001/VV+고/EC] (×)　　[공부__001/NNG+하/XSV+고/EC] (○)

붙임. 접사 처리(통합) 과정에서 제외된 접사 목록

명사파생접미사				동사파생접미사	
가(哥)	김가	들이	1L들이	당하	공격당하다
가량	1시간가량	발(發)	서울발	되	준비되다
간(間)	한 달간	배기*	나이배기	받	강요받다
경(頃)	두 시경	분지(分之)	삼분지 일	시키	운동시키다
계(系)	몽고계	생1(生)	갑자생	하	공부하다
께	10분께	씩	만원씩	형용사파생접미사	
꼴	십 원꼴	어치	만원어치	답	사람답다
끼리	전우끼리	여(餘)	삼십여	되	거짓되다
네	동이네	정(整)	일만 원정	롭	슬기롭다
당(當)	한 사람당	짜리	백 원짜리	스럽	사랑스럽다
대(臺)	억대	째1**	이틀째	하	건강하다
댁(宅)	청주댁	쯤	내일쯤		
들	우리들	하(下)	지배하		

* '-배기'는 '세 살, 다섯 살' 등 수량을 나타내는 구와 결합하는 경우 접사 처리(통합)에서 제외하나, '나이배기, 알배기, 공짜배기, 대짜배기, 진짜배기'는 결합형으로 처리함.

** '-째1'은 접사 처리(통합)에서 제외하나, 의존명사 '번째'는 결합형으로 처리함.

*** 이 밖에 숫자나 문장부호, 띄어쓰기 등으로 어근과 접사가 분리된 경우 접사 처리(통합)에서 제외함. (예. 제2차, 고(高)지대를, 4분의 1 등)

다. 분석원칙

분석 대상 후보가 되는 모든 어형에 어휘의미 번호를 부여한다.

2. 어휘의미 분석 표지

가. 어휘의미의 목록과 체계는 국립국어원의 <우리말샘>을 기준으로 한다.

나. 말뭉치 정보 부착 방법은 해당 어절의 분석 결과에 '__'과 세 자리의 어휘의미 번호 '○○○'를 부착하는 방식을 사용한다. 의미 번호는 국립국어원 <우리말샘>에 나타난 표제어 등재 번호와 일치한다.

[예시] 돈이 꽃이 되면 기적 같은 일을 일으킨다. [기적__003/NNG]
서울 종로구청 앞에 첫 번째 소녀상을 세웠다. [번째__001/NNB]
나는 지금 하는 일을 좋아서 하고 있는가. [나__003/NP+는/JX]
원나라의 몰락과 홍건적의 침입을 겪으며 [겪__001/VV+으며/EC]
어떠한 정보도 제공하지 않고 있다. [어떠하__001/VA+ㄴ/ETM]
공간을 이동하는 '순간이동' 마술이다. [마술__002/NNG+이__004/VCP+다/EF+./SF]

<우리말샘>에 제시된 갈래뜻 사이에 구분이 불분명할 경우에는 <우리말샘>의 예문을 최대한 검토하여 확정한다. 이때 한쪽이 예문이 없는 것이라면 예문이 있는 쪽의 갈래뜻을 선택한다.

[예시] 후보가 되면 심층 인성평가를 거친다. [인성__002/NNG+평가__001/NNG+를/JKO]
※<우리말샘>에 등재된 '인성'의 다양한 의미 중 "001 사람의 성품"과 "002 각 개인이 가지는 사고와 태도 및 행동 특성"은 뜻 사이에 구분이 불분명하다. 이때 전자의 예문 '인성 교육, 인성이 착하다, 올바른 인성을 기르다'와 후자의 예문 '인성 개발, 인성 검사'을 검토하여 위 문장에 제시된 '인성'에 어휘의미 번호 '002'를 부여한다.

제주도가 세계환경수도 조성에 잰걸음을 내딛었다. [내딛__002/VV+었/EP+다/EF+./SF]
※<우리말샘>에 등재된 '내딛다'는 "001 【…을】 '내디디다'의 준말"과 "002 【…을】 '내디디다'의 준말"로 동일하게 풀이되어 있다. 이때 전자의 예문 '길이 험하여 발을 내딛기가 힘들다.'를 통하여 001은 장소의 이동과 관련된 것으로 파악할 수 있고, 후자의 예문 '새로운 삶에 첫발을 내딛는 결혼.'을 통하여 002는 새로운 범위 안에 처음 들어섬으로 분석할 수 있다. 이러한 예문 검토를 통하여 제시된 '내딛다'에 어휘의미 번호 '002'를 부여한다.

중국과 얼어붙은 관계를 녹이려는 분명한 신호를 보냈다. [녹이__005/VV+려는/ETM]

※<우리말샘>에 등재된 '녹이다'의 의미 중 "005 【…을】 감정을 누그러지게 하다. '녹다'의 사동사"와 "006 【…을】 감정을 누그러지게 하다. '녹다'의 사동사"은 동일하게 풀이되어 있다. 이때 전자의 예문 '미움을 녹일 수 있는 것은 서로의 처지를 이해하고 아끼며 보살펴 주는 마음밖에는 없다.'를 통하여 '녹이다005'는 "감정이 누그러지다"에 해당하는 '녹다004'의 사동사로 파악할 수 있고, 후자의 예문 '요염한 여자가 남자를 녹이다.'를 통하여 '녹이다006'은 "어떤 대상에 몹시 반하거나 홀리다"에 해당하는 '녹다011'의 사동사로 분석할 수 있다. 이러한 예문 검토를 통하여 위 문장에 제시된 '녹이다'에 어휘의미 번호 '005'를 부여한다.

매서운 강풍이 문풍지를 뒤흔들었다. [강풍__006/NNG+이/JKS]

※<우리말샘>에 등재된 '강풍'의 다양한 의미 중 '003'과 '006'은 "세게 부는 바람"으로 의미가 동일하다. 그러나 '003'에는 예문이 없고 '006'에만 예문이 있기 때문에 위 문장에 제시된 '강풍'에 어휘의미 번호 '006'을 부여한다.

단, 양쪽 모두 예문이 없다면 포괄적인 갈래뜻을 선택한다.

[예시] 창밖으로 시원한 소나무숲이 들어온다. [소나무__001/NNG+숲__001/NNG+이/JKS]

※<우리말샘>에 등재된 '소나무'의 다양한 의미 중 '001'은 "소나뭇과의 모든 식물을 통틀어 이르는 말"로, 학명이 함께 제시된 '002'의 의미를 포함하고 있다. 이때 예문을 검토하여 의미를 확정하는 것이 원칙이나, 두 갈래뜻 모두 예문을 제시하지 않아 한쪽을 선택하기 어렵다. 이 경우 포괄적인 의미를 나타내는 '001'을 위 문장에 제시된 '소나무'에 어휘의미 번호로 부여한다.

근처 중국식당에서 피로연을 즐겼다. [중국__001/NNP+식당__002/NNG+에서/JKB]

※<우리말샘>에 등재된 '중국'의 다양한 의미 중 '001'은 삼황오제 시대부터 지금의 중화 인민 공화국을 가리키는 "아시아 동부에 있는 나라"로, "아시아 동북부에 있는 인민 공화국"을 의미하는 '002'를 포함하고 있다. 두 갈래뜻 모두에 예문이 제시되지 않아 한쪽을 선택하기 어려우므로 포괄적인 의미를 나타내는 '001'을 위 문장에 제시된 '중국'에 어휘의미 번호로 부여한다.

다. <우리말샘>에 동일한 한글 배열로 이루어진 형태가 등재되지 않은 어휘의 경우, 어휘의미 번호는 '777'로 한다.

[예시] 나는 된장을 끓을 때, 홍고추도 넣고 청고추도 넣는다. [청고추__777/NNG+도/JX]
전공에 관련된 텍스트북을 구비하였다. [텍스트북__777/NNG+을/JKO]
요즘처럼 날씨가 더운 때에는 공캉스가 최고야. [공캉스__777/NNG+가/JKS]
※공캉스 : 공항에서 보내는 바캉스의 의미
군데군데 시멘트가 덧발려 있었고, [덧발리__777/VV+어/EC]
팽창기가 달콤고소한 튀밥만 안겨준 것은 아니었다. [달콤고소하__777/VA+ㄴ/ETM]

말뭉치에 나타난 준말에 대응하는 본말이 <우리말샘>에 등재되어 있더라도 준말의 형태는 미등재어이기 때문에 어휘의미 번호 '777'을 부여한다.

[예시] 오송역에서 교원대로 향했다. [교원대__777/NNP+로/JKB]
금융투자협회 회장 선거가 여의도 금투협회에서 열렸다. [금투협회__777/NNP+에서/JKB]
※'교원대'와 '금투협회'의 본말인 '한국^교원대'와 '한국^금융^투자^협회'가 <우리말샘>에 등재되어 있더라도 준말의 형태는 미등재어에 해당하기 때문에 어휘의미 번호 '777'을 부여한다.

라. <우리말샘>에 동일한 한글 배열로 이루어진 형태가 등재되어 있으나 해당 어휘의 의미가 제시되어 있지 않은 경우, 어휘의미 번호는 '888'로 한다.

[예시] 집을 구하려면 부동산을 찾아가 보는 것이 편리하다. [부동산__888/NNG+을/JKO]
※<우리말샘>에 '부동산'이라는 표제어가 등재되어 있는데, 현재 사전상의 의미는 '동산'(動産)의 반의어인 '부동산'(不動産)이다. 그런데 위 문장에 쓰인 '부동산'의 의미는 "움직여 옮길 수 없는 재산"이 아니라 "공인 중개 사무소"이다. 따라서 그 형태는 등재되어 있지만 해당 의미가 제시되어 있지 않은 '부동산'에 어휘의미 번호 '888'을 부여한다.

코리안 드림을 꿈꾸다. [드림__888/NNG+을/JKO]
※<우리말샘>에 "매달아서 길게 늘이는 물건" 등의 의미로 '드림'이 등재되어 있으나 "꿈"에 해당하는 의미가 제시되어 있지 않으므로 '드림(dream)'에 어휘의미 번호 '888'을 부여한다.

금융감독원의 제재를 받았다. [금융__001/NNG+감독원__888/NNG+의/JKG]
※<우리말샘>에 "감독하는 직무를 맡은 사람"을 의미하는 '감독원(監督員)'이 등재되어 있으나 "감독 기관"에 해당하는 의미가 제시되어 있지 않으므로 '감독원(監督院)'에 어휘의

미 번호 ‘888’을 부여한다.

그는 신작에 다양한 제주 설화를 끌어들였다. [끌어들이__888/VV+었/EP+다/EF+./SF]
※<우리말샘>에 따르면 ‘끌어들이다’는 “남을 권하거나 꾀어서 자기편이 되게 하다”를 의미하나, 위 문장에 사용된 ‘끌어들이다’는 “인용하다”에 가까운 의미로 판단된다. 따라서 <우리말샘>에 등재되지 않은 의미가 사용된 ‘끌어들이다’에 어휘의미 번호 ‘888’을 부여한다.

우완 투수들이 공을 던질 때 왼쪽 어깨가 닫혀 있어야 한다는 게 원칙이지만,
[닫히__888/VV+어/EC]
※<우리말샘>에 ‘닫히다’는 “001 열린 문짝, 뚜껑, 서랍 따위가 도로 제자리로 가 막히다. ‘닫다’의 피동사”, “002 하루의 영업이 끝나다. ‘닫다’의 피동사”, “003 굳게 다물어지다. ‘닫다’의 피동사”의 의미로 등재되어 있다. 그러나 위 문장은 투수가 공을 던질 때 타자에게 가슴이 보이지 않게끔 어깨가 벌어지지 않아야 한다는 것을 의미하는데, 이러한 의미의 ‘닫히다’는 <우리말샘>에 등재되어 있지 않다. 따라서 위 문장에 사용된 ‘닫히다’에 어휘의미 번호 ‘888’을 부여한다.

특정 분야의 전문 용어로 <우리말샘>에 등재된 어휘가 다른 분야에서 쓰이거나 일상 용어로 사용된 경우 어휘의미 번호 ‘888’을 부여한다.

[예시] 대부분이 여행사의 마라톤 패키지 상품으로 홍콩을 찾았다. [패키지__888/NNG]
※<우리말샘>에 제시된 ‘패키지’는『영상』분야에 사용되는 전문 용어로 “시나리오 작가와 인기 배우, 인기 소설과 인기 배우 등의 결합”을 의미한다. 위 문장에 쓰인 ‘패키지’ 역시 “일괄 상품 혹은 상품의 결합” 정도의 의미로 <우리말샘>의 등재 어휘와 의미적 관련성을 찾을 수 있으나『서비스업』이라는 다른 분야에서 사용되는 용어이기 때문에 해당 어휘의 의미가 등재되지 않은 것으로 판단하여 어휘의미 번호 ‘888’을 부여한다.

골프채의 헤드와 샤프트 모두 나무를 깎아 만들기 시작했다. [샤프트__004/NNG]
※<우리말샘>에 제시된 ‘샤프트’는『체육』분야에 사용되는 전문 용어로 “배드민턴 라켓의 긴 막대 부분”을 의미한다. 위 문장에 쓰인 ‘샤프트’는 “골프채의 긴 막대 부분”을 가리키기 때문에 <우리말샘>의 등재 어휘와 의미적 관련성을 찾을 수 있다. 비록 구체적인 운동 종목은 다르나『체육』이라는 동일 분야에서 사용되는 용어이기 때문에 <우리말샘>에 등재된 어휘의미 번호 ‘004’를 부여한다.

<우리말샘>에 등재된 일반명사가 고유명사로 쓰이거나, 고유명사가 일반명사로 사용된 경우 고유명사의 대상성을 고려하여 어휘의미 번호 '888'을 부여한다.

[예시] 한겨레를 펼쳐 보다. [한겨레__888/NNP+를/JKO]

※매체명 '한겨레'의 경우 <우리말샘>에 등재된 '한겨레'와 의미적 유사성을 가질 수도 있으나 그러한지를 일일이 판단하고 확인하는 것이 불가능하기 때문에 고유명사의 대상성을 고려하여 어휘의미 번호 '888'을 부여한다.

영화 '변호인'을 보기 위해 극장으로 몰려갔다. ['/SS+변호인__888/NNP+'/SS+을/JKO]

※일반명사에서 파생된 고유명사의 경우 <우리말샘>에 등재되어 있더라도 대상성을 고려하여 어휘의미 번호 '888'을 부여한다.

파격적 구도의 고려 수월관음도가 일본에서 발견되었다. [수월관음도__888/NNG+가/JKS]

※<우리말샘>에 등재된 '수월관음도'는 용인대학교에 소장된 보물 제1286호를 가리킨다. 그러나 위 문장처럼 문화재명 '수월관음도'가 맥락상 일반명사로 쓰일 경우 어휘의미 번호 '888'을 부여한다.

마. 말뭉치 원어절의 오타, 탈자 등의 오류로 의미 분석이 불가능한 경우, 어휘의미 번호는 '999'로 한다.

[예시] 고짖을 부리다. [고짖__999/NNG+을/JKO]

인도네시아 산업부 관리들이 묻던 1961호에 별생각 없이 들어갔다. [묻__999/VV+던/ETM]

아미산이 어디 잇는 거예여? [잇__999/VA+는/ETM]

좀비랜드2 보세요 재밋네요 [재밋__999/VA+네/EF+요/JX]

야채를 복았는데 [복__999/VV+았/EP+는데/EC]

잘 본것 같아요 [같__999/VA+아요/EF]

엽떢은 [엽떢__999/NNP+은/JX]

괟찮아 [괟찮__999/VA+아/EF]

꺼내기가귀찮네 [꺼내__001/VV+기/ETN+가/JKS+귀찮__999/VA+네/EF]

※형태 분석 과정에서 평폐쇄음화와 무관하여 원어절의 표기형 그대로 분석된 단어는 오표기로 판단하여 '999'를 부여한다.

이정도면 괜낞다 [괜낞__999/VA+다/EF]

나 겨울왕구 볼거야 [겨울왕구__999/NNP]
※'괜찮다'에서 /ㅊ/가 탈락한 '괘낞다', '겨울왕국'에서 /ㄱ/가 탈락한 '겨울왕구' 등은 해당 음운이 탈락한 이유를 설명하기 어려움으로 '999'를 부여한다.

순삭되는기뷴? [순삭__005/NNG+되/XSV+는/ETM+기뷴__999/NNG+?/SF]
별그짖같은 제목이야 [별/MMA+그짖__999/NNG+같__003/VA+은/ETM]
금욜일이라 힘이 나네요 [금욜일__999/NNG+이/VCP+라/EC]
※'기분'에 /j/이 첨가된 '기뷴', '그지'에 /ㅈ/가 첨가된 '그짓', '금요일'에 /ㄹ/가 첨가된 '금욜일' 등은 음운이 첨가된 이유를 설명하기 어려움으로 '999'를 부여한다.

나 겨울오아국 안 봤어 [겨울오아국__999/NNP]
※'겨울왕국'에서 자모 배열이 바뀐 '겨울오아국'은 입력 과정의 단순 실수로 판단되므로 '999'를 부여한다.

단, 아래에 해당하는 경우 <우리말샘>에 등재된 표준어에 대한 비표준어형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

ㄱ. 경음화를 적용한 경우

입력의 비경제성에도 불구하고 의도적으로 경음 표기를 사용한 단어는 비표준어형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다. 다만 표준형의 경음이 원 어절에서 평음으로 나타날 경우, '999'를 부여한다.

[예시] 다섯씨쯤 끝낭 [다섯__001/NR+씨__888/NNB+쯤/XSN]
점심으로 돈까스 먹었엉 [돈까스__777/NNG]
삐싸기두하고 [삐싸__777/VA+기/ETN+두/JX+하__042/VX+고/EC]
경기도 싸나입니다 [싸나이__777/NNG+이__004/VCP+ㅂ니다/EF]
게껍찔까지 씹어먹고 ㅋㅋㅋㅋㅋ [게__001/NNG+껍찔__777/NNG+까지/JX]
화장실이나 부얼 개쫍아서 [개쫍__777/VA+아서/EC]
※위 문장처럼 경음화를 적용한 단어는 일반적인 비표준어형으로 판단한다.

떡볶이 먹을꺼야 [먹__002/VV+을/ETM+꺼__001/NNB+이__888/VCP+야/EF]
짤라서 보니까 [짜르__001/VV+아서/EC]
※위 문장에 사용된 '꺼'와 '짜르-'는 <우리말샘>에 등재되어 있으므로 해당 의미번호를 부여한다.

그지골이 되고 [그지골__999/NNG+이/JKC]
※'그지꼴'이 원어절에서 '그지골'로 나타난 경우로, '999'를 부여한다.

ㄴ. 격음화를 적용한 경우

형태소 내부에서 발생한 격음화 현상이 표기에 반영된 단어는 비표준어형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 가카, 그만 내려오시죠 [가카__777/NNG+,/SP]

ㄷ. 기타 한국어의 음운 현상을 적용한 경우

형태 분석 지침에서 표기된 형태를 그대로 보존한 음운 현상(형태 분석 지침 93~95쪽 참고)이 적용된 경우 비표준어형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 대다내 [대다나__777/VA+아/EF]
※'대다내'는 '대단해'에서 /ㅎ/가 탈락한 것으로, 일반적인 비표준어형으로 판단한다.

잘하고싶어가꼬 [잘/MAG+하/XSV+고/EC+싶__001/VX+어/EC+가__888/VX+꼬/EC]
※'가꼬'는 '갖고[갇꼬]'에서 /ㄷ/가 탈락한 것으로, 일반적인 비표준어형으로 판단한다.

꾸꾹이 해주더라 [꾸꾹이__777/NNG]
※'꾸꾹이'는 '꾹꾹이'에서 /ㄱ/가 탈락한 것으로, 일반적인 비표준어형으로 판단한다.

출근길은 너무 빡셔요 [빡시__888/VA+어요/EF]
※'빡시다'는 '빡세다'에서 /ㅔ/가 /ㅣ/로 변하는 고모음화를 적용한 표기로, 일반적인 비표준어형으로 판단한다.

강기 [강기__888/NNG]
※'강기'는 '감기'에 연구개음화를 적용한 표기로, 일반적인 비표준어형으로 판단한다.

ㄹ. 표준어형의 발음과 동일한 경우

표준어형의 현실 발음과 동일한 단어는 비표준어형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 순두부찌게 먹었어요 [순두부찌게__777/NNG]
김치찌게를 가장한 김치국이요... [김치찌게__777/NNG+를/JKO]
※현실 발음에서 /ㅐ/와 /ㅔ/의 구분이 불분명하여 표준어형인 '찌개'와 비표준어형인 '찌게'의 현실 발음이 동일하므로 '순두부찌게'와 '김치찌게'를 일반적인 비표준어형으로 판단한다.

여행 다니면서 잼있게 살자 [잼있__777/VA+게/EC]
※'잼있-'의 발음과 <우리말샘>에 등재된 '재밌다'의 현실 발음이 동일하므로 '잼있-'을 일반적인 비표준어형으로 판단한다.

ㅁ. 어휘의 일부가 비표준어형으로 사전에 등재된 경우

분석 대상 어휘의 일부가 비표준어형으로 <우리말샘>에 등재된 단어는 그 전체를 비표준어형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 영등포주민들 따스허네ㅠㅠㅠ [따스허__777/VA+네/EF+ㅠㅠㅠ/SW]
잘 하고 있는가 궁금허네 [궁금허__777/VA+네/EF]
※'허다'는 '하다'의 비표준형으로 <우리말샘>에 등재되어 있으므로 이를 포함한 '따스허-'와 '궁금허-'를 일반적인 비표준어형으로 판단한다.

ㅂ. 연관성 있는 단어가 사전에 등재된 경우

분석 대상 어휘의 형성 가능성을 유추할 수 있는 표현이 <우리말샘>에 등재된 단어는 비표준형으로 판단하여 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 근데 이릏게 길게 얘기해도 돼냐? [이릏__777/VA+게/EC]
※<우리말샘>에 "'이렇게'의 방언(전북, 충청)"으로 등재된 '이릏게'를 참고하여 위 예문의 '이릏게'를 '이릏-'의 부사형으로 분석 가능하므로 '이릏-'을 일반적인 비표준어형으로 판단한다.

선지국도 좋아하고 [선지국__777/NNG+도/JX]
※<우리말샘>에 '순댓국'의 비표준어형으로 등재된 '순대국'을 참고하여 '선지국'을 일반적인 비표준어형으로 판단한다.

3 세부 지침

체언류

가. 일반명사(NNG)

수를 셀 수 있는 단위사로 쓰일 수 있는 명사들은 기본 뜻풀이와 셀 때의 뜻풀이가 서로 상이하므로, 그에 맞는 어휘의미 번호를 주어야 한다.

[예시] 골목은 폭이 좁아 마주 오는 사람과 스칠 듯했다. [사람__001/NNG+과/JKB]
동갑내기인 두 사람은 중학교 시절부터 친구 사이였다. [사람__010/NNG+은/JX]

정부와 국가기관 관련 어휘는 뜻풀이에 '우리나라'로 한정되어 있더라도 다른 나라에 유사한 기관이 존재하는 것이 인정되므로 동일한 어휘의미 번호를 부여한다.

[예시] 연방 결혼법이 미국 대법원에서 위헌 판결이 났다. [대법원__001/NNG+에서/JKB]
※<우리말샘>에 등재된 '대법원'은 "우리나라의 최고 법원"으로 정의되어 있으나, 미국 등 다른 나라의 최고 법원 역시 대법원으로 번역되므로 <우리말샘> 등재어 '대법원'의 포함되는 포함되는 의미로 다룰 수 있다.

나. 고유명사(NNP)

인명의 경우 <우리말샘>을 참조하여 최대한 어휘의미를 부여하는 것을 원칙으로 하되, 사전에 등재되어 있지 않은 것(외국인의 성, 이름 포함)은 어휘의미 번호 '777'을 부여하고, 한자 또는 영문 표기의 일치 여부와 관계없이 해당 표기가 등재되어 있는 것은 어휘의미 번호 '888'을 부여한다.

[예시] 이순신__002/NNP : 충무공　　김옥균__001/NNP : 갑신정변
아인슈타인__001/NNP : 상대성이론　　비틀즈__001/NNP : 록 그룹
김종서__001/NNP : 조선 전기의 충신　　김종서__888/NNP : 록 가수
김대중__777/NNP　　노무현__777/NNP
이명박__777/NNP　　박근혜__777/NNP

'성'(性)의 경우에는 '우리나라 성씨 순위, 성씨별 인구밀도 순위' 등의 자료를 참고하여 사전에 등재되어 있는 같은 '성'(性) 중 대표적인 성의 의미 번호로 통일한다.

[예시] 정 시인은 주로 자연의 정취를 주제로 서정시를 쓴다.　　[정__054/NNP]
※<우리말샘>에는 나주, 창원 등 10여 본이 현존하는 '정(丁)'과 경주, 동래 등 120여 본이 현존하는 '정(鄭)'이 등재되어 있는데, 성씨별 인구수 등을 고려하여 '정(鄭)'의 의미 번호 '054'를 부여한다.

붙임. 둘 이상 등재된 성(性)의 대표 의미 번호

강__004	도__015	빈__003	수__008*	옹__006	장__028	채__026
경__036	돈__012	사__011	순__007	우__015	전__007	천__002
공__012	마__006	서__013	승__009	운__005	정__054	초__018
구__013	모__029	석__007	시__009	원__003	조__023	추__001
국__008	반__019	선__012	신__003	위__033	종__020	편__003
기__010	방__003	설__013	심__015	유__022	주__012	하__007
노__013	범__006	성__006	양__017	윤__001	지__010	호__016
뇌__005	변__003	소__025	여__018	이__023	진__026	
단__017	봉__010	송__001	연__003	임__010	창__007	

(국가통계포털 "성씨·본관별 인구(5인 이상)-전국"을 참고함.)
*인구조사표에는 '수'(隋)의 인구가 많으나 <우리말샘>에 미등재되어 있어 그 다음 인구수를 보이는 '수'(水)의 의미 번호를 부여함.

지명의 경우 <우리말샘>을 참조하여 최대한 어휘의미를 부여하는 것을 원칙으로 하되, 한자 또는 영문 표기의 일치 여부와 관계없이 해당 표기가 사전에 등재되어 있지 않은 것은 어휘의미 번호 '777'을 부여하고, 등재되어 있는 것은 어휘의미 번호 '888'을 부여한다.

[예시] 서울__002/NNP　　송파구__777/NNP
캘리포니아__777/NNP　　런던__003/NNP
런던__888/NNP : 캐나다 온타리오주　　분당__888/NNP : 경기 성남
가수리__002/NNP : 경북 영천　　가수리__003/NNP : 경남 의령
가수리__888/NNP : 강원 정선　　가수리__888/NNP : 전남 화순

다. 수사(NR)

수사는 때로 수사와 수관형사, 수사와 명사의 구별이 애매한 경우가 있다. 이에 형태 분석에서는 특이한 형식을 가진 예만을 수관형사로 취급하고, 그 밖의 것들은 모두 수사로 분석하였다. 이에 어휘의미 분석에서도 수사로 처리된 어휘에 대해 수사의 의미 번호를 부여하여 품사와 어휘의미 번호를 일치하게 한다.

[예시] 우리는 하나로 뭉쳤다.　　[하나__001/NR+로/JKB]
만점자가 주위에 한둘이 아니다.　　[한둘__001/NR+이/JKC]
상대에게 몇 차례 결정적 기회를 내줬다.　　[몇__001/NR]
지금도 몇몇 분야는 두각을 나타내고 있다.　　[몇몇__001/NR]
농사지은 것으로 일년 내내 밥상을 차렸다.　　[일__018/NR+년002/NNB]

용언류

가. 동사(VV) 및 형용사(VA)

비유적 의미보다는 <우리말샘>의 등재 의미 중심으로 분석한 체언과 달리 용언은 관용 표현 등에 나타나는 비유적 의미를 나타내는 방향으로 어휘의미 번호를 부여하는 것을 기본 원칙으로 한다. 이때 관용 표현의 비유적 의미는 <우리말샘>의 '속담·관용구'를 참고한다. 이와 관련하여 용언의 의미 번호를 부여하는 기준은 다음과 같다.

우선 뜻풀이를 중심으로 갈래뜻의 차이를 반영하여 <우리말샘>에 등재된 의미를 최대한 부여한다. 다만 뜻풀이만으로 명확하지 않은 경우 예문의 쓰임을 보고 의미 번호 부여 가능성을 판단한다.

[예시] 선수단이 직접 닭과 돼지를 기르기도 한다. [기르__001/VV+기/ETN+도/JX]
아이를 낳고 기를 수 있는 사회 환경을 만들어야 한다. [기르__002/VV+ㄹ/ETM]
광복 이후에는 광주국악원을 만들어 후배를 길렀다. [기르__003/VV+었/EP+다/EF+./SF]
※<우리말샘>에 "001 【…을】 동식물을 보살펴 자라게 하다", "002 【…을】 아이를 보살펴 키우다", "003 【…을】 사람을 가르쳐 키우다" 등으로 등재되어 있는 '기르다'는 보살펴 자라게 하는 대상에 따라 구분되며, 물리적인 돌봄과 정신적인 가르침 등으로 구분된다. 이처럼 <우리말샘>에 나타나는 의미 차를 반영하여 분석 대상 어휘의 의미 번호를 부여한다.

체체파리는 젖을 먹여 새끼를 뱃속에서 기르는데 평생 그 수는 6마리에 그친다.
[기르__002/VV+는데/EC]
※위에서 살핀 바와 같이 '기르다'는 대상이 동식물인 경우(001)와 아이인 경우(002)로 구분된다. 위 문장에 나타나는 보살핌의 대상은 체체파리의 새끼, 즉 동식물에 속하지만 주체인 체체파리 입장에서는 '아이'에 해당하므로 <우리말샘>에 등재된 의미 " 【…을】 아이를 보살펴 키우다"의 어휘의미 번호 '002'를 부여한다.

취득세 감면 시한이 다가오면서 가파른 상승곡선을 그렸다. [가파른__001/VA+ㄴ/ETM]
※<우리말샘>에 따르면 '가파르다'는 "001 산이나 길이 몹시 기울어져 있다"를 의미한다. 위 문장의 경우 '산'이나 '길'이 아닌 '주택 거래량'의 상태를 가리키는데, 뜻풀이만으로는 거래량의 모습과 산의 모습을 나타내는 '가파르다'가 동일한 의미인지 파악하기 어렵다. 하지만 '001'의 예문 '이에 따라 휘발유 소비량이 가파른 상승세를 보이고 있습니다.≪MBC 뉴스데스크 1998년 3월≫'를 통해 거래량이 몹시 기울어진 모습도 '001'에 해당하는 것으로 판단할 수 있다. 따라서 위 문장에 사용된 '가파르다'에 어휘의미 번호 '001'을 부여한다.

하나의 형태가 동사와 형용사, 보조용언 모두로 쓰일 수 있을 때, 형태 분석에 근거하여 의미 번호를 부여한다. 단, 형태 분석에 오류가 있을 경우 이를 별도로 기록한 후 형태 분석 말뭉치의 개선에 반영한다.

[예시] 1시간가량 조용히 있다가 갑자기 일어났다. [있__001/VV+다가/EC]
현재 갈등에도 분명 해결책이 있을 것이다. [있__006/VA+을/ETM]
서울에 살고 있는 동생 [있__023/VX+는/ETM]
※<우리말샘>에 '있다'는 동사(001~004), 형용사(005~021), 보조용언(022~023)으로 등재되어 있다. 이 경우 형태 분석 결과와 동일한 품사에 해당하는 의미 번호를 분석 대상 어휘에 부여한다.

분석 대상 어휘가 쓰인 문장의 구조를 파악하여 해당 용언이 요구하는 문장 성분에 따라 어휘의미 번호를 부여한다. 특히 갈래뜻 사이에 구분이 불분명한 경우 문장 구조에 따라 알맞은 어휘의미 번호를 부여할 수 있다. 이때 전후 맥락을 파악하여 생략된 문장 성분을 복원하여 분석한다.

[예시] 이 팀은 종합예술을 만든다는 게 강점이다. [만들__011/VV+ㄴ다는/ETM]
그간 정부가 밝혀온 원칙을 무색하게 만들었다. [만들__013/VV+었/EP+다/EF+./SF]
게시판의 열독자로 만들 수 있다. [만들__012/VV+ㄹ/ETM]
※<우리말샘>에 13개의 의미로 등재되어 있는 '만들다'는 3개의 문장 유형으로 구분된다. 대다수는 목적격 조사 {을}을 요구하는 문장으로, 위의 첫 문장이 이에 해당한다. 따라서 동일한 문장 유형의 의미 번호(001~011) 중 '영화나 드라마 따위를 제작하다'를 의미하는 011을 부여한다. 선행 용언에 연결어미 {-게}가 결합한 두 번째 문장 역시 동일한 문장 구조를 가지는 013을, {으로}가 결합된 부사어가 나타난 세 번째 문장은 {을}과 {으로}를 동시에 요구하는 012를 각각 부여한다. 이때 전후 맥락을 통하여 세 번째 문장에 생략된 목적어 "특정인을"을 복원하여 분석한다.

위층 사람들이 아래층에 내려왔다. [내려오__001//VV+았/EP+다/EF+./SF]
산 밑으로 흘러 내려오는 물을 받아 온다. [내려오__001/VV+는/ETM]
정상을 거쳐 성판악휴게소까지 내려오는 등산코스 [내려오__001/VV+는/ETM]
간신히 산을 내려온 이들의 사연을 앞다퉈 보도했다. [내려오__008/VV+ㄴ/ETM]
집 잘 지어놨으니 산에서 내려오면 들러 달라. [내려오__008/VV+면/EC]
흉기를 든 20대 남성이 계단에서 내려오고 있었다. [내려오__008/VV+고/EC]
※<우리말샘>에 등재된 '내려오다'의 의미 중 "001 【…에】 【…으로】 높은 곳에서 낮은 곳으로 또는 위에서 아래로 가다"와 "008 【…을】 높은 곳에서 낮은 곳으로 위치를 옮기다"는 갈래뜻 사이에 구분이 불분명하다. 이러한 경우 문장 구조를 통해 어휘의미 번호를 부여할 수 있는데, '001'은 주로 도착 지점과 함께 쓰이고, '008'은 주로 내려오는 행위가

이루어지는 장소와 함께 나타난다. <우리말샘>에서는 '001'과 관련하여 조사 '에'와 '으로'를 제시하고 있으나 도착점과 관련된 '까지'도 '001'에 포함될 수 있다. '008'의 경우 조사 '을'뿐만 아니라 행동이 이루어지고 있는 장소와 관련된 '에서'도 포함될 수 있다.

문장의 구조로 어휘의미 번호를 확정하기 어려운 경우, 분석 대상 어휘와 공기하는 체언류의 의미 부류에 근거하여 어휘의미를 분석한다.

[예시] 마술사는 아무 일도 없었다는 듯이 무대 위로 걸어 올라왔다. [없__003/VA+었/EP+다는/ETM]
나라간 이견이 큰 현안에서는 진전이 별로 없었다고 해석했다. [없__004/VA+었/EP+다고/EC]
구속 담당 장관이 없다는 이유로 [없__001/VA+다는/ETM]
처벌 가치가 없는 것으로 판단했다. [없__002/VA+는/ETM]

※<우리말샘>에 '없다'는 16개의 의미로 등재되어 있다. 이중 15개가 형용사와 관련되는데 특별한 부사격 조사를 취하지 않는 유형과 {에}나 {에게} 등을 요구하는 유형으로 구분된다. 위의 네 문장은 요구하는 문장 성분 측면에서 동일하기 때문에 문장의 구조로 의미를 분석하기 어렵다. 이러한 경우 <우리말샘>에 뜻풀이를 최대한 활용하여 어휘의미를 분석한다. 즉 첫 번째 문장은 마술사가 아무 일도 생겨 나타나지 않은 듯이 행동하고 있으므로 "003 어떤 일이나 현상이나 증상 따위가 생겨 나타나지 않은 상태이다"로 분석 가능하고, 두 번째 문장은 '별로'라는 부사를 통해 "004 어떤 것이 많지 않은 상태이다"로 분석할 수 있다. 그러나 무언가가 존재하지 않은 상태를 나타내는 세 번째 문장과 네 번째 문장은 '없다'와 공기하는 체언류의 의미 부류에 근거하여 의미 번호를 부여한다. <우리말샘>에 따르면 001은 '사람, 동물, 물체' 등 구체적인 대상과 함께 나타나는 데 반해, 002는 '사실이나 현상'처럼 추상적인 것과 주로 어울린다. 이를 바탕으로 위의 세 번째 문장과 네 번째 문장에 사용된 '없다'의 의미를 구별할 수 있다.

이때 말뭉치에 나타난 분석 대상 용언이 <우리말샘> 뜻풀이에서 제시하고 있는 대상 이외의 체언류와 함께 사용된 경우, 어휘의미 번호 '888'을 부여한다.

[예시] 구시가지를 갈아엎고 고층 아파트를 올렸다. [갈아엎__001/VV+고/EC]
판을 완전히 갈아엎는 인사다. [갈아엎__888/VV+는/ETM]
※<우리말샘>에 등재된 '갈아엎다'는 "【…을】 땅을 갈아서 흙을 뒤집어엎다"처럼 흙을 대상으로 풀이되어 있다. '구시가지를 갈아엎고'의 경우 특정 지역의 땅을 갈아서 흙을 뒤집어엎은 것과 관련되나, '판을 갈아엎는'의 경우 실제 땅과 흙이 아니라 판, 즉 분위기 정도를 뒤집어엎은 것으로 분석된다. 이처럼 대상이 땅, 흙과 관련된 '갈아엎다'에는 어휘의미

번호 ‘001’을 부여하고, 그 밖에 대상과 함께 쓰인 ‘갈아엎다’에는 어휘의미 번호 ‘888’을 부여한다.

대체 투입된 버스로 갈아타 인명피해는 없었다. [갈아타__001/VV+아/EC]
저금리 대출로 갈아탈 수 있다. [갈아타__888/VV+ㄹ/ETM]
※<우리말샘>에 등재된 ‘갈아타다’는 “【…으로】 【…을】 타고 가던 것에서 내려 다른 것으로 바꾸어 타다”로 등재되어 있는데, 예문을 살펴보면 ‘버스, 지하철, 비행기, 말’ 등 탈것을 대상으로 한다. 그러나 실제 쓰임에서는 ‘대출, 정당, 펀드’ 등 탈것과 무관한 것을 대상으로 하며 “옮기다” 정도를 의미한다. 이처럼 <우리말샘>에 등재되지 않은 ‘갈아타다’에는 어휘의미 번호 ‘888’을 부여한다.

다만 뜻풀이에 ‘따위’ 등이 사용되어 대상이 확장될 가능성이 있을 경우에는 해당 의미 번호를 부여한다. 이때 대상의 확장 가능성은 <우리말샘> 어휘지도에 제시된 반의어와 유의어를 통해 확인할 수 있다.

[예시] 산 정상 부근의 얼음이 급속히 녹아든 것이 원인이다. [녹아들__001/VV+ㄴ/ETM]
현대 클래식의 영향이 두루 녹아들었다. [녹아들__002/VV+었/EP+다/EF+./SF]
그는 삼성화재의 배구에 녹아들지 못하고 있다. [녹아들__002/VV+지/EC]
※<우리말샘>에 등재된 ‘녹아들다’는 “001 【…에】 다른 물질에 스며들거나 녹아 들어가다”와 “002 【…에】 사상이나 문화 따위가 섞여 어울리다”로 구분된다. 따라서 얼음이 녹아든 것은 ‘001’이 되고, 영향이 녹아든 것은 ‘002’가 된다. ‘삼성화재의 배구’의 경우 팀 분위기 혹은 경기 방식을 가리키는데, 이는 ‘사상이나 문화 따위’에 속하는 것으로 해당 분석 대상 어휘에 어휘의미 번호 ‘002’를 부여한다.

시나몬 향이 청량하면서도 은은한 향기를 자아낸다. [자아내__003/VV+ㄴ다/EF+./SF]
※위 문장의 ‘향기를 자아내다’는 “향기가 저절로 생겨난다” 정도의 의미인데, 이는 <우리말샘>의 등재 의미 “003 【…을】 어떤 감정이나 생각, 웃음, 눈물 따위가 저절로 생기거나 나오도록 일으켜 내다”와 관련된 것으로 판단된다. <우리말샘>에 따르면 ‘자아내다003’의 대상은 ‘감정이나 생각, 웃음, 눈물 따위’인데 여기에 ‘향기’가 포함될 가능이 있기 때문에 위 문장의 ‘자아내다’에 어휘의미 번호 ‘003’을 부여한다.

인터넷세대 당기는 복간본 시집의 매력 [당기__888/VV+는/ETM]
※위 문장에 사용된 ‘당기다’는 <우리말샘>의 등재 의미 “003 【…을】 물건 따위를 힘을 주어 자기 쪽이나 일정한 방향으로 가까이 오게 하다”와의 관련성을 생각해 볼 수 있다. 이

때 뜻풀이에 제시된 대상의 범주인 '물건 따위'에 '인터넷세대'가 포함될 수 있는가를 판단해야 한다. 여기서 활용할 수 있는 것이 '당기다003'의 반의어 '밀다'이다. 위 문장에 사용된 '당기다'는 '밀다'와 반의어 관계를 형성하지 않으므로 '003'보다는 등재되지 않은 새로운 의미로 판단된다. 실제로 위 문장의 '당기다'는 "유혹하다" 정도의 의미로 판단되므로 어휘의미 번호 '888'을 부여한다.

새로운 민원이 5건 넘게 들어와 방법을 모색하는 중이다. [들어오__888/VV+아/EC]
※위 문장에 사용된 '들어오다'는 <우리말샘>의 등재 의미 "002 【…에】 【…으로】 수입 따위가 생기다"와의 관련성을 생각해 볼 수 있다. 이때 뜻풀이에 제시된 대상의 범주인 '수입 따위'에 '민원'이 포함될 수 있는가를 판단해야 한다. 여기서 활용할 수 있는 것이 '들어오다002'의 반의어 '나가다'이다. 위 문장에 사용된 '들어오다'는 '나가다'와 반의어 관계를 형성하지 않으므로 '002'보다는 등재되지 않은 새로운 의미로 판단된다. 실제로 위 문장의 '들어오다'는 "접수되다" 정도의 의미로 판단되므로 어휘의미 번호 '888'을 부여한다

분석 대상 용언이 선행하는 체언구와 함께 어우러져 비유적 표현의 일부로 사용된 경우, 비유적 표현 전체 의미의 서술어 부분에 해당하는 의미가 등재되지 않은 경우 '888'을 부여한다. 이때 비유적 표현의 전체 의미는 <우리말샘> '속담·관용구'를 참고한다.

[예시] 여당과 야당이 손을 잡다. [잡__888/VV+다/EF+./SF]
※위 문장에 쓰인 '잡다'는 명사구 '손을'과 결합하여 "서로 힘을 합쳐 협력하다" 정도를 의미한다. 그러나 <우리말샘>에 등재된 '잡다'의 다양한 의미 중 "서로 힘을 합쳐 협력하다"의 서술구인 '협력하다'에 해당하는 의미가 없으므로 '잡다'에 어휘의미 번호 '888'을 부여한다.

부모들은 억장이 무너지는 기분이라고 입을 모았다. [모으__888/VV+았/EP+다/EF+./SF]
※위 문장에 쓰인 '모으다'는 흔히 '입'과 함께 쓰여 "여러 사람이 같은 의견을 말하다"를 의미하는데, <우리말샘>에 등재된 '모으다'의 다양한 의미 중 "여러 사람이 같은 의견을 말하다"의 서술구인 '의견을 말하다' 정도에 해당하는 의미가 없으므로 '모으다'에 어휘의미 번호 '888'을 부여한다.

사업 아이템이 줄줄이 열매를 맺었다. [맺__888/VV+었/EP+다/EF+./SF]
※위 문장에 쓰인 '맺다'는 '열매'와 함께 쓰여 "노력한 일의 성과가 나타나다", 즉 "성과를 내다"를 의미하는데, <우리말샘>에 등재된 '맺다'의 다양한 의미 중 "노력한 일의 성과를 내다"의 서술구인 '성과를 내다' 정도에 해당하는 의미가 없으므로 '맺다'에 어휘의미 번호

‘888’을 부여한다.

동계스포츠까지 손을 뻗쳐 사익을 노렸다. [뻗치__888/VV+어/EC]
※위 문장에 쓰인 ‘뻗치다’는 흔히 ‘손’과 함께 쓰여 “이제까지 하지 아니하던 일까지 활동 범위를 넓히다”를 의미한다. 그러나 <우리말샘>에 등재된 ‘뻗치다’의 다양한 의미 중 ‘활동 범위를 넓히다’에 해당하는 의미가 없으므로 ‘뻗치다’에 어휘의미 번호 ‘888’을 부여한다.

손이 큰 어머니는 언제나 음식을 푸짐하게 차리신다. [크__003/VA+ㄴ/ETM]
※위 문장에 쓰인 ‘크다’는 흔히 ‘손’과 함께 쓰여 “씀씀이가 후하고 크다”를 의미하는데, 서술구에 해당하는 ‘후하고 크다’는 ‘크다003’, 즉 “일의 규모, 범위, 정도, 힘 따위가 대단하거나 강하다”에 해당하는 것이므로 어휘의미 번호 ‘003’을 부여한다.

기업가의 돈을 먹고서도 양심 있는 사람이라고 할 수 있는가. [먹__009/VV+고서/EC+도/JX]
※위 문장에 쓰인 ‘먹다’는 흔히 ‘돈’과 함께 쓰여 “(속되게) 뇌물을 받다”를 의미하는데, 서술구에 해당하는 ‘뇌물을 받다’는 ‘먹다009’, 즉 “(속되게) 뇌물을 받아 가지다”에 해당하는 것이므로 어휘의미 번호 ‘009’를 부여한다.

머리를 맞대고 대책을 강구하다. [맞대__001/VV+고/EC]
※위 문장에 쓰인 ‘맞대다’는 흔히 ‘머리’과 함께 쓰여 “어떤 일을 의논하거나 결정하기 위하여 서로 마주 대하다”를 의미하는데, 서술구인 ‘서로 마주 대하다’는 ‘맞대다001’, 즉 “서로 가깝게 마주 대하다”에 해당하는 것이므로 어휘의미 번호 ‘001’를 부여한다.

숨 가쁘게 돌아가는 뉴스의 세계 [가쁘__003/VA+게/EC]
※위 문장에 쓰인 ‘가쁘다’는 흔히 ‘숨’과 함께 쓰여 “어떤 일이 몹시 힘에 겹거나 급박하다”를 의미하는데, 서술구에 해당하는 ‘힘에 겹거나 급박하다’는 ‘가쁘게003’, 즉 “((주로 ‘가쁘게’ 꼴로 쓰여)) 몹시 급하거나 빠르다”에 해당하는 것이므로 어휘의미 번호 ‘003’를 부여한다.

환유적 표현에 해당하는 경우 <우리말샘>의 해당 어휘의미 번호를 부여한다.

[예시] 그저 입 다물고 사태가 가라앉길 기다리고 있다. [다물__001/VV+고/EC]
※위 문장에 쓰인 ‘입을 다물다’는 “말을 하지 아니하거나 하던 말을 그치다”를 의미하는데, 이는 ‘입’과 ‘다물다’의 사전적 의미의 합인 “입을 꼭 맞대다”로, ‘말을 하지 않는 것’을 의

미하는 환유적인 표현이다. 이처럼 환유적인 표현에 해당하는 경우 <우리말샘>에 등재된 어휘의미 번호를 부여한다. 따라서 위 문장에 사용된 '다물다'에 <우리말샘> 등재 의미인 "【…을】 입술이나 그처럼 두 쪽으로 마주 보는 물건을 꼭 맞대다"에 해당하는 어휘의미 번호 '001'을 부여한다.

나. 긍정지정사(VCP) 및 부정지정사(VCN)

긍정지정사(VCP) '-이다'와 부정지정사(VCN) '아니다'는 각각 <우리말샘>에 조사 '이다'와 형용사 '아니다'로 등재된 어휘의미 번호를 부여한다.

[예시] 지역 라이벌이라 불리지만 [라이벌__001/NNG+이__004/VCP+라/EC]
세계적인 스타들이 펼치는 축구의 향연 [세계적__001/NNG+이__005/VCP+ㄴ/ETM]
그의 강점을 살린 영리한 선택이었다. [선택__001/NNG+이__006/VCP+었/EP+다/EF+./SF]
불법 감금 기간이 상당히 오래이고 [오래/MAG+이__007/VCP+고/EC]
이사를 생각한 것은 집이 좁게 느껴져서다.
[느끼__006/VV+어/EC+지__016/VX+어서/EC+이__008/VCP+다/EF+./SF]
봄이 와서 꽃이 피는 것이 아니라 [아니__001/VCN+라/EC]
선점 효과를 막으려 하는 게 아니냐는 얘기다. [아니__002/VCN+냐는/ETM]

작품명 뒤에 또는 종결어미 뒤에 쓰인 '-이다'의 경우, 선행하는 작품명과 문장을 명사구로 판단하여 어휘의미 번호를 부여한다.

[예시] 오늘부터 우리는이었어요. [우리__003/NP+는/JX+이__004/VCP+었/EP+어요/EF+./SF]
잘 알지도 못하면서라는 영화거든요 [못하__004/VX+면서/EC+이__004/VCP+라는/ETM]
다 다른데 요번 잃어버린 어~ 시간을 찾아서 라는 플스투의 소설을 이 그림하고
[이__004/VCP+라는/ETM]
초딩 때는 초등학생일 때는 인데 줄인 겁니다. [이__004/VCP+ㄴ데/EC]
※위 문장들은 각각 '오늘부터 우리는'(노래명), '잘 알지도 못하면서'(영화명), '잃어버린 시간을 찾아서'(소설명) 뒤에 '-이다'가 사용된 경우이다. 이 경우 작품명 전체를 하나의 명사구로 판단하여 어휘의미 번호 '004'(주어가 지시하는 대상의 속성이나 부류를 지정하는 뜻을 나타내는 서술격 조사)를 부여한다. '초딩 때는'~'초등학생일 때는'처럼 줄어든 말의 본딧말을 밝히는 경우에도 어휘의미 번호 '004'를 부여한다.

이제 아 우리에 대통령으로 뽑을 만한 사람인가라는 생각을 하면서
[사람__001/NNG+이__004/VCP+ㄴ가/EF+이__004/VCP+라는/ETM]
부도덕할뿐만 아니라 무능하다라는 게
[무능__001/NNG+하/XSA+다/EF+이__004/VCP+라는/ETM]
그게 어떻게 이익이랑 연결이 된다라는 거야 그게.
[되__022/VV+ㄴ다/EF+이__004/VCP+라는/ETM]
그게 지금 안 되고 있다라는 게 [있__023/VX+다/EF+이__004/VCP+라는/ETM]
또는 누구와 가깝다라는 것에 [가깝__002/VA+다/EF+이__004/VCP+라는/ETM]
늙어 보인다라는 악플을 단 거예요. [보이__002/VV+ㄴ다/EF+이__004/VCP+라는/ETM]
제가 얻었다라는 생각을 합니다. [얻__001/VV+었/EP+다/EF+이__004/VCP+라는/ETM]
※위 문장들은 종결어미 '-ㄴ가, -ㄴ다, -다' 등과 결합한 '-이다'가 사용된 경우이다. 이 경우 선행 문장을 명사구로 판단하여 어휘의미 번호 '004'(주어가 지시하는 대상의 속성이나 부류를 지정하는 뜻을 나타내는 서술격 조사)를 부여한다.

기타

가. 북한어 및 방언인 경우

말뭉치의 분석 대상 어휘가 <우리말샘>에 북한어, 방언 등으로 등재되어 있으면 해당 어휘의미 번호를 부여한다.

[예시] 며늘아, 차례 음식은 대행에 맡기렴. [며늘__001/NNG+아/JKV+,/SP]
※며늘001 : '며느리'의 방언(경북, 중국 흑룡강성).
어딨는겨? [어디__001/NP+있__001/VA+는/ETM+기__076/NNB+이__888/VCP+여/EF+?/SF]
※기076 : '거'의 방언(경상, 충청).
애님니께~~~~ [애니__001/VCN+ㅂ니께/EF+~/SO+~/SO+~/SO+~/SO]
※애니다001 : '아니다'의 방언(경북).
아빤 뭐 뿌수고 [뿌수__001/VV+고/EC]
※뿌수다001 : '부수다'의 방언(강원, 경기, 경상, 전라, 제주, 충청).
국수 한그륵 후루룩 [한/MMN+그륵__001/NNG]
※그륵001 : '그릇'의 방언(강원, 경상, 전남, 충남).

멀미땜시 이따 다시올게요 [멀미__001/NNG+땜시__001/NNB]

※“때문에(전북)”의 방언형으로 등재된 ‘땜시’는 품사 정보가 제시되지 않아 어떤 품사로든 연결이 가능하다. 따라서 위 문장에 사용된 ‘땜시’에 어휘의미 번호 ‘001’을 부여한다.

나. 외래어인 경우

<우리말샘>에 등재된 외래어 중에는 규범 표기가 미확정된 것들이 많다. 이처럼 외국어로 된 단어의 경우 표준형을 결정하기가 어렵다. 따라서 어휘의미 분석 대상 어휘가 외국어로 된 단어인 경우 해당 어휘의 사전 등재 유무에 따라 어휘의미 번호를 부여한다.

[예시] 그냥 타지마할 사진 찍었다고 해도 믿겠어요 [타지마할__001/NNP]

키자니아 엄청자주가던데 [키자니아__777/NNP]

에픽호이 타블라 [에픽호이__777/NNP]

난 크리스토프도 좋던뎅...친구같아서 [크리스토프__888/NNP+도/JX]

넹 가니쉬랑 다 들어있어서 요리하기 간편해요~~~ [가니쉬__777/NNG+랑/JC]

※<우리말샘>에 ‘가니시’가 등재되어 있으나 규범 표기가 미확정이기 때문에 이를 표준형으로 보기 어렵다. 이에 ‘가니쉬’ 역시 가능한 표기로 판단하여 어휘의미 번호 ‘777’을 부여한다.

룸싸롱못가서 그렇답니다 [룸싸롱__777/NNG+못/MAG+가__001/VV+아서/EC]

※<우리말샘>에 ‘룸살롱’이 규범적 표기로 등재되어 있으나 다른 외래어와 동일하게 다양한 발음의 가능성을 인정하여 어휘의미 번호 ‘777’을 부여한다.

다. 숫자나 알파벳이 포함된 경우

숫자나 알파벳이 포함된 단어는 <우리말샘>에 해당 숫자나 알파벳을 한글로 옮겨 쓴 형태가 등재되어 있으면 해당 어휘의미 번호를 부여한다.

[예시] 모두 한국신기록을 세우며 대회 3관왕이 됐다. [3관왕__002/NNG+이/JKC]

U턴 문제 등 우려하는 문제가 불거지지 않도록 하겠다. [U턴__001/NNG]

라. 기호가 포함된 경우

기호가 포함된 단어는 <우리말샘>에 기호를 생략한 형태가 등재되어 있으면 해당 어휘의미

번호를 부여한다.

[예시] 건물의 냉·난방비를 줄여준다. [냉·난방비__001/NNG+를/JKO]
국·공립대 중 1인당 장학금 수혜율 3위를 기록하고 있다. [국공립대_001/NNG]
뭐.. 당.연.ㅋㅋㅋ [당.연__003/NNG+./SF+ㅋㅋㅋ/MAG]
즐...거운 월요일...이지 [즐...겁__001/VA+ㄴ/ETM]
날씬~하구요 [날씬~하__001/VA+구/EC+요/JX]
뜨끈~한 보상주는 [뜨끈~하__001/VA+ㄴ/ETM]

마. 문자 모양의 유사성에 기반한 경우

문자 모양의 유사성에 기반하여 형태를 변형한 단어는 오표기로 판단하지 않고 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 댕댕이는 산책 안 시킴 힘들어지죠 [댕댕이__003/NNG+는/JX]
※'댕댕이'는 '멍멍이'를 문자 모양의 유사성에 기반하여 변형한 단어이다. <우리말샘>에 '댕댕이'가 등재되어 있으므로 해당 의미번호인 '003'을 부여한다.
팔도네넴띤 이것도 좋아요 [팔도네넴띤__777/NNP]
※'팔도네넴띤'는 '팔도비빔면'을 문자 모양의 유사성에 기반하여 변형한 단어이다. <우리말샘>에 해당 형태가 등재되어 있지 않으므로 어휘의미 번호 '777'을 부여한다.

바. 구 등재어인 경우

어휘의미 분석은 형태 분석 말뭉치의 분석 결과를 바탕으로 한다. 즉 형태 분석 지침에 의거하여 구성 요소가 분리되어 분석된 경우에는 각 구성 요소에 어휘의미 번호를 부여하고, 하나로 분석된 어휘에는 하나의 어휘의미 번호를 부여한다. 이때 말뭉치에서 둘 이상으로 분리하여 형태 분석된 단어가 합쳐져서 <우리말샘>에 구로 등재되어 있더라도 분석된 어휘 각각에 해당하는 어휘의미 번호를 부여한다. 단, 하나로 형태 분석된 어휘가 <우리말샘>에 구로 등재되어 있으면 해당 어휘에 구 등재어의 어휘의미 번호를 부여한다.

[예시] 국사편찬위원회 위원장으로써 올해의 계획이 또 있으실 것 같거든요?
[국사__004/NNG+편찬__001/NNG+위원회__001/NNG]
※'국사편찬위원회'는 형태 분석 지침에 따라 일반명사 '국사/NNG+편찬/NNG+위원회/NNG'가 결합된 구성으로 분석되기 때문에 <우리말샘>에 구로 등재된 '국사^편찬^위원회'가 있더라도 '국사__004/NNG+편찬__001/NNG+위원회__001/NNG'처럼 구성 요소 각각에 어

휘의미 번호를 부여한다.

국가인권위원회는 학습권 보장의 필요성을 제기했다. [국가인권위원회__001/NNP+는/JX]

※'국가인원위원회'는 형태 분석 지침에 따라 하나의 고유명사로 분석되기 때문에 <우리말샘>에 구로 등재된 '국가^인권^위원회'의 어휘의미 번호를 '국가인권위원회__001/NNP'처럼 분석 대상 어휘 전체에 하나의 어휘의미 번호를 부여한다.

이 건물은 평소 템플스테이 숙소로 쓰여 왔다. [템플스테이__001/NNG]

뜻밖에 템플 스테이 [템플__888/NNG, 스테이__888/NNG]

※원시말뭉치에 붙여쓰기된 '템플스테이'와 띄어쓰기된 '템플 스테이'는 형태 분석 지침에 따라 일반명사 하나로 분석되기도 하고 일반명사 둘의 결합으로 분석되기도 한다. 어휘의미 분석은 형태 분석 말뭉치의 분석 결과를 바탕으로 하기 때문에 하나로 분석된 '템플스테이'에는 <우리말샘>의 구 등재어 '템플^스테이'의 어휘의미 번호를 부여하고, 둘로 분석된 '템플 스테이'는 구 등재어 여부와 상관없이 '템플'과 '스테이' 각각에 어휘의미 번호를 부여한다.

사. 참여자 제안 정보인 경우

<우리말샘> '참여자 제안 정보'에 제시된 단어는 어휘의미 분석에 활용하지 않는다.

[예시] 얼죽아는 진리죠 [얼죽아__777/NNG+는/JX]

※'얼죽아'가 <우리말샘> '참여자 제안 정보'에 "아무리 추워도 찬 음료를 마심. 또는 그런 사람을 이르는 말"로 제안되었으나, 이를 고려하지 않고 어휘의미 번호 '777'을 부여한다.

안먹는게 중요하죠 행님 [행님__777/NNG]

※'행님'이 <우리말샘> '참여자 제안 정보'에 "깡패집단 사이에서 '형님'을 부르는 말"로 제안되었으나, 이를 고려하지 않고 어휘의미 번호 '777'을 부여한다.

아. 맥락이 불완전한 경우

맥락이 불완전하여 의미 분석이 곤란한 단어는 어휘의미 번호 '001'을 부여한다.

[예시] 고물 [고물__001/NNG]

※위의 '고물'은 끝말잇기 상황에서 출현한 단어로 맥락상 그 의미 파악이 불가능하다. 이 경우 기본 의미로 판단되는 '001' 부여한다.

메신저 대화

가. 개인정보를 치환한 표지의 경우

형태 분석 지침에 의해 NNP, NNG로 분석된 개인정보 치환 표지에는 어휘의미 번호 '777'을 부여한다.

[예시] 김name3 [김name3__777/NNP]
제 카트라이더 아이디랑 비슷하네요 account [account__777/NNP]
address으로 이사와. [address__777/NNP+으로/JKB]
저는 affiliation에서 근무해요. [affiliation__777/NNP+에서/JKB]
아직은 여성 others는 제주에서 저 혼자^^ [others__777/NNG+는/JX]

나. 초성만으로 표기된 단어의 경우

형태 분석 지침에 의해 분석된 초성 단어는 '777'을 부여한다. 다만 1음절 초성 단어의 경우, 'ㄱ__001'과 같이 자모자가 등재되어 있으므로 '888'을 부여한다.

[예시] ㄱㅅ 기차탓오요 [ㄱㅅ__777/NNG]
당면 두부피만 잇어도 ㄱㅊ [ㄱㅊ__777/VA]
원인을 못찾는데 ㅂㅅ [ㅂㅅ__777/NNG]
겨울왕국 안본거에서 까인거임 ㅅㄱ [ㅅㄱ__777/NNG]
자기소개 ㄹㅇ ㅆㅇㅈ [ㅆㅇㅈ__777/NNG]
ㅇㅈ합니다 [ㅇㅈ__777/NNG+하/XSV+ㅂ니다/EF]
ㅅㅂ년아 [ㅅㅂ년__777/NNG+아/JKV]
헬스 ㅈ도 모르는게 헬스한다고 가서 [ㅈ__888/NNG+도/JX]
나는 ㅋ쟁이거든 [ㅋ쟁이__777/NNG+이__004/VCP+거든/EF]
※위 문장은 단어의 일부를 초성으로 표기한 경우로, 동일 형태가 사전에 등재되어 있지 않으므로 '777'을 부여한다.

다. 접사에 준하는 요소가 결합한 경우

'노', '갓', '충', '개', '핵', '스/쓰' 따위의 접사에 준하는 요소가 결합한 단어는 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 노야근이란 없다 [노야근__777/NNG+이란/JX]
파밍충이얌 [파밍충__777/NNG+이__004/VCP+얌/EF]
나도 런던!!대박스 [런던__003/NNP+!/SF+!/SF+대박스__777/NNG]
오호 겨울왕국쓰~ [겨울왕국쓰__777/NNP+~/SO]
누가 재미있어해? 궁금쓰 [궁금쓰__777/XR]
진짜 핵추워 [핵춥__777/VA+어/EF]
갓치... [갓치__777/NNG+.../SE]
개꿀이네 [개꿀__888/NNG+이__004/VCP+네/EF]
확실히 갓겜이긴함 [갓겜__001/NNG+이__004/VCP+기/ETN+ㄴ/JX+하/VX+ㅁ/ETN]
※위 문장에 사용된 '노야근', '파밍충', '궁금쓰', '대박스', '핵춥다', '갓치' 등은 의미를 강조하기 위해 의도적으로 사용된 표현으로, <우리말샘>의 등재 여부를 판단하여 형태가 미등재된 경우 '777'을, '개꿀'처럼 형태는 등재되어 있으나 해당 의미가 없는 경우 '888'을, '갓겜'처럼 <우리말샘> 등재어의 경우 해당 의미 번호를 부여한다.

라. 음절을 첨가하여 장음을 표현한 경우

의미를 강조하기 위해 음절을 첨가하여 장음을 표현한 단어는 <우리말샘> 등재 유무에 따라 해당 어휘의미 번호를 부여한다.

[예시] 봄여어어어어어어름갈겨어어어어어울이에요
[봄__001/NNG+여어어어어어어름__777/NNG+갈__002/NNG+겨어어어어어울__777/NNG+이__004/VCP+에요/EF]
수제수제수제버거어 [수제__001/NNG+수제__001/NNG+수제__001/NNG+버거어__777/NNG]
고오오급 음식이잖아 [고오오급__777/NNG]
고오오오맙다 [고오오오맙__777/VA+다/EF]
매에워어 [매엡__777/VA+어어/EF]
다이어트 하려고....고구마...파프리카아..
[하__001/VV+려고/EC+..../SE+고구마__002/NNG+.../SE+파프리카아__777/NNG+../SE]

과암??부러워..　　[과암__777/NNP+?/SF+?/SF+부럽__001/VA+어/EF+../SE]
줄오오오오오지게　　[줄__002/NNG+오오오오오지__777/VA+게/EC]
※위 문장에 사용된 '여어어어어어름', '겨어어어어울', '버거어', '고오오급', '고오오오맙-', '매엡-', '파프리카아', '오오오오오지-' 등은 의미를 강조하기 위해 의도적으로 음절을 첨가하여 장음으로 표시한 경우로, <우리말샘>에 형태가 등재되지 않았음으로 '777'을 부여한다.

3. 형태 분석 말뭉치 구축 지침

이 장에서는 본 사업에서 형태 분석 말뭉치를 구축하기 위하여 적용한 지침의 전모를 보이고자 한다. 이는 2019년도에 마련된 문어·구어 분석을 위한 형태 분석 지침을 일부 보완하고, 메신저 대화 분석을 위한 지침을 새롭게 추가한 것이다. 지침의 전체 구성은 다음과 같다.

1 기본 원칙

가. 분석대상

형태 분석은 하나의 어절을 분석 대상으로 한다.

나. 분석원리

본 분석은 '형태소' 차원이 아닌 '형태' 차원의 분석이므로 이형태를 최대한 반영한다.

다. 분석원칙

형태분석은 분석 대상인 원시 말뭉치를 가급적 훼손하지 않는다. 본 분석은 국립국어원의 <우리말샘>의 표제어를 기준으로 한다.

2 어절 분석 표지

가. 이 어절 분석표지는(이하 세종 형태 표지) 21세기 세종계획 국어기초자료 구축 분과에서 '형태 분석 말뭉치(morpheme tagged corpus)'를 구축하기 위해 마련된 것을 토대로 작성된 것이다.

나. 이 분석 표지는 큰 틀은 21세기 세종계획의 어절 분석 표지를 따르고, 품사 태그의 경우는 TTA의 분석 표지를 참고하였다. 그리고 세종 말뭉치의 문어, 구어 분석 표지를 통합한 것이다.

다. 이 형태 표지는 단계적인 분석을 할 수 있도록 고려하였다.

대분류	소분류	세분류	태그
체언	명사	일반명사	NNG
		고유명사	NNP
		의존명사	NNB
	대명사	대명사	NP
	수사	수사	NR
용언	동사	동사	VV
	형용사	형용사	VA
	보조용언	보조용언	VX
	지정사	긍정지정사	VCP
		부정지정사	VCN
수식언	관형사	성상 관형사	MMA
		지시 관형사	MMD
		수 관형사	MMN
	부사	일반부사	MAG
		접속부사	MAJ
독립언	감탄사	감탄사	IC
관계언	격조사	주격조사	JKS
		보격조사	JKC
		관형격조사	JKG
		목적격조사	JKO

		부사격조사	JKB
		호격조사	JKV
		인용격조사	JKQ
	보조사	보조사	JX
	접속조사	접속조사	JC
의존형태	어미	선어말어미	EP
		종결어미	EF
		연결어미	EC
		명사형전성어미	ETN
		관형형전성어미	ETM
	접두사	체언접두사	XPN
	접미사	명사파생접미사	XSN
		동사파생접미사	XSV
		형용사파생접미사	XSA
	어근	어근	XR
기호	일반기호	마침표, 물음표, 느낌표	SF
		쉼표, 가운뎃점, 콜론, 빗금	SP
		따옴표, 괄호표, 줄표	SS
		줄임표	SE
		붙임표(물결)	SO
		기타 기호	SW
	외국어	외국어	SL
	한자	한자	SH
	숫자	숫자	SN
	분석불능범주	분석불능범주	NA
		명사추정범주	NF
		용언추정범주	NV
기타		개인정보처리필요요소	NAP

3 표지별 분류 기준 및 세부 지침

가 체언

체언은 명사, 대명사, 수사를 포괄하는 대범주로서, 조사와 결합하거나 그 자체로 다른 체언이나 용언과 어울려 하나의 문장성분이 될 수 있다.

1) 명사(NN)

명사는 사물의 이름을 나타내는 품사이다. 본 표지에서는 명사를 일반명사, 고유명사, 의존명사로 세분한다.

가) 일반명사(NNG)

사물의 이름을 나타내는 단어로서 <우리말샘>에 명사로 등재된 표제어(고유명사와 의존명사를 제외한 모든 명사)와 독립된 음절(한자어), 약어, 고사성어 등 사전 표제어는 아니나 다른 품사로 분석될 수 없는 단위들을 포함한다.

(1) 일반명사로 분석할 수 있는 단어

(가) <우리말샘>의 명사 표제어

[예시] 국어/NNG, 연구/NNG

(나) 1음절 한자어가 독립된 단위로 사용되는 경우

[예시] 서울초등학교 졸	[졸/NNG]

(다) 한자성어

[예시] 백척간두(百尺竿頭)	[백척간두/NNG+(/SS+百尺竿頭/SH+)/SS]

(라) 기타 다른 품사로 분석될 수 없는 단위

표기상 한글과 기타 문자(부호나 숫자, 외국 문자)가 섞여 있고, 한글과 기타 문자를 분리했을 때 '절', '루수', '관왕', '유'같이 다른 품사로 분석될 수 없는 단위가 도출되는 경우에는 분리하지 않고 통합하여 분석한다.

[예시] 3.1절(국경일) [3.1절/NNG]
1루수 [1루수/NNG]
5관왕 [5관왕/NNG]
벙커C유 [벙커C유/NNG]

기타 문자가 포함된 단위에서 접사를 분리할 수 있을 것으로 생각되더라도, 접사 분리 시 단어의 의미 구조와 맞지 않게 된다면 접사 분리를 하지 않는다.

[예시] 제3자 [제3자/NNG]
→ '제-'는 본 지침의 분석 대상 접사이고, '-자'는 분석 대상이 아닌 접사이다. 이때 '제-'를 '3자'로부터 분리하여 [제/XPN+3자/NNG]로 분석하면, 이 단어의 의미 구조가 '제3의 사람', 즉 '제3-자'인 것과 맞지 않게 된다. 따라서 접사 '제-'를 분리하지 않고 전체 단어를 일반명사로 처리한다.
[예시] 제3국 [제3국/NNG]
→ 위와 마찬가지로 '제-'는 본 지침의 분석 대상 접시아고 '-국'은 분석 대상이 아닌 접사이다. 역시 '제-'를 '3국'으로부터 분리하여 [제/XPN+3국/NNG]로 분석하면, 이 단어의 의미 구조가 '제3의 나라', 즉 '제3-국'인 것과 맞지 않게 된다. 따라서 접사 '제-'를 분리하지 않고 전체 단어를 일반명사로 처리한다.

기타 문자를 분리하고 남는 단위가 '의존명사+비분석 접사'인 경우에는, '의존명사+비분석 접사'를 합하여 의존명사로 처리한다.

[예시] 16개교 [16/SN+개교/NNB]
→ '개'는 의존명사이고 '교'는 사전에 등재되어 있지 않으나 접사에 준하는 요소로 파악된다. 또한 이 '-교'는 본 지침에서 분석하지 않는 접사이다. 이때 비분석 접사인 '-교'를 의존명사 '개'에 합하여 의존명사 '개교'를 설정하여 분석한다.
[예시] 16강전 [16/SN+강전/NNB]
→ 위의 경우와 마찬가지로 의존명사 '강'에 비분석 접사 '-전'이 결합한 '강전'을 의존명사로 설정하여 분석한다. <우리말샘>에 '16강전'이 '십육-강전'이 아닌 '십육강-전'으로 올라 있어 본 지침의 처리와 달라지기는 하나, '개교'와의 구조적 유사성, '강전'이 다양한 수사와 어울려 쓰이며 의존명사와 같은 행태를 보임에 주목하는 것이다.
[참고] 십육강전 [십육강전/NNG]
→ '십육강전'이 기타 문자 없이 한글만으로 쓰인 경우에는 사전의 처리를 따라 '십육강전/NNG'으로 분석한다. 기타 문자를 이용해서 표기했는지 한글만으로 표기했는지에 따라

달리 처리하게 되지만, 기타 문자는 분리하는 것이 원칙이라는 점을 고려한 것이다.

단, 숫자나 외국어로만 표기된 경우에는 모두 각각을 분석한다.

[예시] 6.25 [6/SN+./SP+25/SN]

(2) 명사 상당어의 분석

(가) 동사의 활용형이 따옴표 없이 문장 속에서 명사처럼 기능하는 경우는 원래 품사대로 분석한다.

[예시] 어디 가느냐가 그의 물음이었다. [가/VV+느냐/EF+가/JKS]

(나) 따옴표를 가진 성분이나 요소도 명사처럼 기능할 수 있으나, 원래 품사대로 분석한다.

[예시] 그것은 "는"이 아니라 "를"이다. ["/SS+는/JX+"/SS+이/JKC]
[예시] A: 미않. B: 미안이지 않이 뭐니? [않/NA+이/JKS]
→ 따옴표는 없지만 위 예시와 같은 경우로, 상대방의 오타를 지적하며 오타 부분에 해당하는 '않'을 따와 명사처럼 사용한 경우이다. 이때의 '않'은 그 자체로 품사를 가지지 않는 말이므로 분석 불능 범주(NA)를 부여한다.

(다) 부사 뒤에 격조사가 쓰이는 것도 의미론적인 따옴의 효과에 의하여 부사가 명사적인 용법을 가지는 것이므로 분석은 '부사'로 한다.

[예시] 기름을 꽉 채우려면 가득을 누르세요. [가득/MAG+을/JKO]

나) 고유명사(NNP)

고유명사는 특정한 사물에 붙여진 이름으로, 기본적으로 최하의어에 속하는 대상을 서로 변별하기 위하여 붙인 이름이며, 원칙적으로 지시 대상만 가질 뿐 의미 내용은 가지지 않는다. 고유명사의 분석 기준은 매우 다양하므로, 본 지침에서는 아래에 제시하는 것만을 고유명사로 인정한다.

아래에 제시한 고유명사 부류에 속하는 말이 두 어절 이상에 걸쳐 나오는 경우가 있다. 이때에는 전체가 외국어로 구성된 말인지, 하나라도 고유어/한자어를 포함하고 있는지에 따라 달리

처리한다.

전체 어절이 외국어로 구성된 경우: 각 어절 모두를 NNP로 처리한다.
하나라도 고유어/한자어를 포함하고 있는 경우: 각각의 어절에 포함된 말이 무엇인지를 살펴 적절한 형태표지를 부여한다.

[예시] 블루 이즈 더 워미스트 컬러 (영화 제목)
[블루/NNP, 이즈/NNP, 더/NNP, 워미스트/NNP, 컬러/NNP]
[예시] 바람과 함께 사라지다 (영화 제목)
[바람/NNG+과/JKB, 함께/MAG, 사라지/VV+다/EF]

아래 지침은 주로 고유어/한자어를 포함하고 있는 고유명사 부류에 대한 설명임에 유의한다. 간혹 필요에 따라 전체가 외국어 구성인 경우에 대한 예시와 설명도 포함하였다.

(1) 인명, 종족명

(가) '씨(氏), 공(公), 군(君), 양(孃), 옹(翁), 대왕(大王)' 등 성 또는 이름 뒤에 같이 쓰이는 호칭어나 직책명은 분리해서 분석한다.

[예시] 남수/NNP+군/NNB, 김/NNP+씨/NNB, 최치원/NNP+옹/NNB, 케네디/NNP+씨/NNB
정/NNP+과장/NNG, 최/NNP+선생/NNG, 세종/NNP+대왕/NNG, 광개토/NNP+대왕/NNG

(나) 성과 이름, 호가 함께 쓰이면 하나의 단위로 분석한다.

[예시] 김철수/NNP, 이태백/NNP, 마르코폴로/NNP

(다) '씨, 군' 등과 달리 '가(哥)'는 접미사이므로, '김가(金哥), 이가(李哥)'는 파생어이다.

[예시] 김/NNP+가/XSN

(라) 사람 이름의 뒤에 '이'가 붙는 경우는 이름과 함께 하나의 단위로 분석한다.

[예시] 진현이/NNP+가/JKS

(마) 특정한 종족의 이름은 고유명사가 된다.

[예시] 알타이족/NNP, 피그미족/NNP, 돌궐족/NNP, 한족/NNP, 유대인/NNP

(바) 특정 동물에게 붙여진 이름도 인명에 준하여 고유명사가 된다.

[예시] 코코/NNP, 똘이/NNP

(사) 소설, 애니메이션 등 허구의 세계에서 쓰인 인명이나 동물명도 고유명사가 된다.

(2) 지명

<우리말샘>에 『지명』으로 올라 있는 단어 부류를 참고하여 지명 여부를 판단하고, 지명에 해당하는 부분과 지역의 종류를 나타내는 말을 묶어 전체를 고유명사로 처리한다.

(가) 내륙, 대륙, 지대, 주, 평원, 만, 늪, 습지, 분지, 사막, 유전, 탄전, 군락지
섬, 제도, 열도
바다, 해변, 포구, 강, 유역, 나루, 호수, 계곡, 연못, 갯벌, 폭포, 삼각주, 빙하, 피오르
길, 거리, 수로, 루트, 로드
산, 산맥, 산지, 화산, 동굴, 숲, 고개, 언덕, 오름, 구릉, 고원, 광산, 절리, 화산대 등의 이름

[예시] 카스피해/NNP, 템즈강/NNP, 태백산맥/NNP, 미시시피호/NNP, 갈라파고스제도/NNP, 갈론계곡/NNP, 감지해변/NNP, 강경포구/NNP, 강계분지/NNP, 강주연못/NNP, 거창분지/NNP, 고수동굴/NNP, 관동팔경/NNP, 그레이트빅토리아사막/NNP

(나) 도(道), 시(市), 읍(邑), 면(面), 리(里), 군(郡), 구(區), 동(洞), 골, 촌, 마을, 자치구, 연구단지, 관광단지, 공업지대, 지역, 지방, 지구 등의 이름은 그 구역의 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 서울특별시/NNP, 성북구/NNP, 강진군/NNP, 인창동/NNP, 빨래골/NNP, 해방촌/NNP, 고포마을/NNP, 네바다주/NNP, 경기지역/NNP, 경인공업지대/NNP, 관서공업지역/NNP, 광시좡족자치구/NNP, 가자지구/NNP, 서안지구/NNP

(3) 국가명 또는 왕조명

(가) 국가의 명칭, 또는 왕조의 명칭은 고유명사로 분석한다.

[예시] 대한민국/NNP, 조선/NNP, 코리아/NNP, 러시아연방/NNP, 미얀마연방공화국/NNP

(나) 다른 형태가 붙어 국가나 왕조의 존립 기간을 나타내는 경우 일반명사로 분석한다.

[예시] 대한제국기/NNG, 조선조/NNG

(다) '남한'과 '북한'을 의미하는 '남, 북, 남북'은 모두 일반명사와 고유명사를 구별한다. 남한을 뜻하는 '남'과 북한을 뜻하는 '북'을 고유명사로 분석한다.

[예시] 남/NNP+과/JC 북/NNP+의/JKG 의견/NNG 차이/NNG
남북/NNP 적십자/NNP+회담/NNG
[참고] 북미/NNP 회담/NNG
→ '북미' 자체는 <우리말샘> 등재어가 아니지만 구 표제어인 '북미^제네바^기본^합의서' 속에서 한 단어로 처리되고 있음을 참고하여 한 단어로 처리한다.

(라) 어떤 국가의 국민을 나타내는 '국가+인(人)'은 통합하여 일반명사로 분석한다.

[예시] 이집트인/NNG, 아제르바이젠인/NNG, 이스라엘인/NNG, 조선인/NNG

(마) 어떤 국가의 군대를 나타내는 '국가+군(軍)'은 통합하여 일반명사로 분석한다.

[예시] 미군/NNG, 북한군/NNG, 영국군/NNG, 일본군/NNG

(4) 건축물이나 시설물 혹은 구조물의 이름

<우리말샘>에 『지명』으로 올라 있는 단어 부류를 참고하여 고유명사 여부를 판단하고, 구조물명, 시설물명에 해당하는 부분과 구조물, 시설물의 종류를 나타내는 말을 묶어 전체를 고유명사로 처리한다.

(가) 도로, 항만, 항구, 터널, 대교, 철교, 뱃길, 운하, 댐, 공항, 터미널, 철도, 전철, 지하철 및 그 명칭과 함께 쓰이는 부대시설은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 부산항/NNP, 대전역/NNP, 서울지하철/NNP, 인천공항/NNP, 테헤란로/NNP, 경부고속도로/NNP, 분당선/NNP, 경춘선/NNP, 정우터널/NNP, 강화대교/NNP, 경인아라뱃길/NNP

단, 어느 지역의 지하철에나 존재하는 '1호선, 2호선' 등은 특정성이 낮으므로 고유명사로 보지 않는다.

[예시] 1호선 [1/SN+호선/NNB]
→ 의존명사 '호'에 비분석 접사 '-선'이 결합한 구성으로, '-선'을 앞말에 붙여 '호선/NNB'로 처리한다.

(나) 해수욕장, 공원, 광장, 정원, 목장, 유원지, 유적지, 절터, 관광지, 테마파크, 전망대, 온천, 시장, 장터, 저수지, 기지, 묘지 등의 시설물도 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다.

[예시] 속초해수욕장/NNP, 올림픽공원/NNP, 설악산국립공원/NNP, 한강시민공원/NNP, 광화문광장/NNP, 화정역광장/NNP, 황복사터/NNP, 가락시장/NNP, 노량진수산시장/NNP, 남극세종기지/NNP, 도라전망대/NNP, 현충원/NNP, 국립서울현충원/NNP

단, '농산물도매시장', '생활체육공원'과 같이 특정 시장이나 공원의 이름이 아니라 시장이나 공원의 유형을 나타내기 위해 쓰인 말은 고유명사로 보지 않는다.
'문화예술공원'은 서울특별시 서초구에 있는 특정 공원의 이름으로 쓰이기도 하고 공원의 유형을 나타내기 위한 말로 쓰이기도 하는데, 맥락을 구분하여 전자의 경우에는 고유명사로, 후자의 경우에는 일반명사로 분석한다.

[예시] 서초구 문화예술공원 (특정 공원의 이름일 때)
[문화예술공원/NNP]
우리 구에 문화예술공원을 설립합시다. (공원의 종류를 나타낼 때)
[문화/NNG+예술/NNG+공원/NNG]

(다) 배, 비행기와 같은 건조물의 이름은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다. [참고]와 같이 특정 집단의 우두머리 이름을 따 배에 빗대어 표현하는 경우에도 고유명사로 처리한다.

[예시] 최영함/NNP, 나로호/NNP
[참고] 신태용호/NNP

(라) 빌딩, 박물관, 극장 등 건물명은 그 종류를 나타내는 말과 함께 전체가 고유명사가 된다. 그리고 <우리말샘>에 구로 등재된 단위라고 하여도 그 통합형을 고유명사로 처리한다.

[예시] 서울역사/NNP, 세종문화회관/NNP, 개나리유치원/NNP, 청와대/NNP, 국회의사당/NNP
국립중앙박물관/NNP, 국립민속박물관/NNP, 구텐베르크박물관/NNP
신라호텔/NNP, 미도파백화점/NNP, 동궁예식장/NNP, 명보극장/NNP, 고대병원/NNP,
인천타워/NNP, 고마신사/NNP

(5) 회사, 상품, 학교, 정당, 기관이나 단체의 이름

(가) 특정 회사, 학교, 학회, 협회, 재단, 정당의 이름은 고유명사로 분석한다.

[예시] 삼성/NNP, 삼성그룹/NNP, 엘지전자/NNP, LG전자/NNP, 현대자동차서비스/NNP
코스닥/NNP, 나스닥/NNP, 코스피/NNP
고려대학교/NNP, 잠실고등학교/NNP, 송파중학교/NNP
국어학회/NNP, 영어영문학회/NNP, 한국사학회/NNP
대한축구협회/NNP, 승마협회/NNP, 프로야구선수협회/NNP
한나라당/NNP, 자유민주주의연합/NNP

회사, 학교, 학회, 협회, 정당의 이름이라고 해도 아래와 같이 한글 없이 기타 문자로만 표기된 경우에는 고유명사가 아니라 기호로 처리한다. 인명 등도 마찬가지이다.

[예시] LG에서 [LG/SL+에서/JKB]

(나) 특정 회사의 상품명과 브랜드명은 전체를 묶어 고유명사로 처리한다. 회사명과 상품명이 한 어절에 나왔을 때나 상품명과 상품의 종류를 나타내는 말이 한 어절에 나왔을 때에도 전체를 묶어 고유명사로 처리한다.

[예시] 초코하임/NNP, 코카콜라/NNP, 갤럭시S8/NNP, e편한세상/NNP, 쉐보레/NNP, 티맵/NNP
[예시] 농심새우깡/NNP, 칠성사이다/NNP, 신한SOL/NNP (회사명+상품명)
[예시] e편한세상아파트/NNP (브랜드명+종류명)
[예시] 하나에스라인적금/NNP (회사명+상품명+종류명)

(다) 상호명은 그 통합형을 고유명사로 처리한다. 상호명과 업종명이 한 어절에 나타난 경우에는 전체를 통합하여 고유명사로 분석하고, 상호명과 업종명이 두 어절로 분리되어 나타난

경우에는 상호명에 해당하는 부분을 통합하여 고유명사로 분석한다.

[예시] 한마음약국/NNP, 동일카센터/NNP
[예시] 다나은 이비인후과 [다나은/NNP, 이비인후과/NNG]

(라) 정부기관의 명칭 중 **지명 등의 고유명사를 포함하고 있어** 특정성이 높은 것만을 통합하여 고유명사로 처리한다. <우리말샘>에 구로 등재된 단위라고 하여도 그 통합형을 고유명사로 처리한다.

단, 특정 기관의 '지청, 지원, 지부', 특정 정당 소속의 '시당' 등은 그것이 지명과 함께 나타났더라도 모든 기관에 존재할 수 있어 특정성이 낮으므로 고유명사로 처리하지 않고 [예시]와 같이 단어별로 분리하여 일반명사로 분석한다.

하지만 '지방법원'의 준말인 '지법'은 지명과 함께 쓰일 경우 특정 기관을 가리키므로(예: 서울지법, 부산지법), '서울지법/NNP, 부산지법/NNP'로 처리한다.

고유명사를 포함하지 않아 고유명사에 들지 않는 정부기관 명칭은 단어별로 분리하여 처리한다.

[예시] 서울고등법원/NNP, 서울시경찰서/NNP, 서대문구치소/NNP
[예시] 여주/NNP+지청/NNG, 충북/NNP+지부/NNG, 서울시/NNP+당/NNG(서울시+당[명사])
[예시] 헌법/NNG+재판소/NNG, 국립/NNG+국어원/NNG, 여성/NNG+가족부/NNG

어떤 단어가 정부기관의 명칭인지, 건물명인지 혼동되는 경우가 있다. 이때에는 <우리말샘>의 뜻풀이를 참고하여, 주된 뜻풀이에 '기관'이라 되어 있으면 정부기관으로 판단하고, '건물'이라 되어 있으면 건물명으로 판단한다.

[예시] 청와대 (사전: 서울 경복궁 뒤 북악산 기슭에 있는 우리나라 대통령 **관저**)
→ 사전의 뜻풀이에서 건물명으로 처리되고 있으므로 [청와대/NNP]로 처리한다.
[예시] 경찰청 (사전: 안전 행정부 소속하에 설치되어 경찰 업무를 관장하는 정부 행정 **기관**)
→ 사전의 뜻풀이에서 정부기관명으로 처리되고 있으며 고유명사가 포함되어 있지 않으므로 [경찰청/NNG]로 처리한다.

(마) 연구소, 위원회, 협의회, 본부, 종교단체의 경우에는 **고유명사나 '국가, 전국, 국제, 세계' 및 이와 유사한 단어를 포함하고 있어** 특정성이 높은 것만을 통합하여 고유명사로 처리한다.

고유명사에 들지 않는 연구소, 위원회, 협의회, 본부, 종교단체명은 단어별로 나누어 처리한다.

[예시] 한국전자통신연구소/NNP, 대한예수교장로회/NNP, 대한불교조계종/NNP, 서울자사고교장협의회/NNP, 한국야구위원회/NNP, 국가인권위원회/NNP, 국제통화기금/NNP

[예시] 조계종/NNP, 천태종/NNP (인명, 지명을 포함하고 있는 말임)

[참고] 감리교/NNG, 장로교/NNG (고유명사를 포함한 말이 아님)

[예시] 통일/NNG+연구소/NNG, 생활/NNG+체육/NNG+연구소/NNG,
통사/NNG+론/XSN+연구회/NNG, 입주자/NNG+대표/NNG+협의회/NNG,
수니파/NNG, 방송/NNG+통신/NNG+위원회/NNG
→ 이처럼 고유명사나 '국가, 전국, 국제, 세계'를 포함하지 않은 단체명은 단어별로 나누어 처리한다.

(바) 특정 부대, 스포츠팀, 그룹, 조직, 클럽에 붙여진 이름은 종류를 나타내는 말과 함께 묶어 고유명사로 처리한다. 이 밖에도 이 부류와 유사한 특정 집단에 고유하게 붙여진 이름을 고유명사로 처리한다.

[예시] 백마부대/NNP (부대명) cf) '육군', '해군', '공군', '해병대'는 일반명사임.
인디언스/NNP, 기아/NNP, 타이거즈/NNP, 서울FC/NNP (스포츠팀명)
트와이스/NNP, 송골매/NNP (그룹명)
서방파/NNP (폭력조직명)
위너스클럽/NNP (클럽명)
티파티/NNP (미국 정부의 건전한 재정 운용을 위해 세금 감시 운동을 펼치는 시민 중심의 신생 보수 단체.)

(사) 약어나 준말의 처리

고유명사가 축약된 형태(준말)로 쓰일 경우 본디말과 함께 준말도 인정하여 축약된 형태 그대로를 고유명사로 분석한다.

[예시] 육사/NNP, 고대/NNP, 자민련/NNP, 서울고법/NNP

(아) 단체장명의 처리

지명/단체명에 '-장(長)'과 같이 접미사가 결합하거나 '수(守)'와 같이 접사에 준하는 요소가 결합하여 단체장명이 만들어진 경우, 해당 요소를 바로 앞말에 붙여 분석한다. (가령 '헌법재판소'는 본 지침에서 '헌법/NNG+재판소/NNG'가 되므로, 여기에 '-장'이 결합한 경우 '헌법/NNG+재판소장/NNG'으로 처리한다.) 또한 단체장명은 일반명사임에 유의하여 형태 표지를 부여한다.

[예시] 가평군수(가평군/NNP+수) [가평군수/NNG]
서울시장(서울시/NNP+장) [서울시장/NNG]
헌법재판소장(헌법/NNG+재판소/NNG+장) [헌법/NNG+재판소장/NNG]
편찬위원회장(편찬/NNG+위원회/NNG+장) [편찬/NNG+위원회장/NNG]
영등포노인종합복지관장 [영등포노인종합복지관장/NNG]
인사부장 [인사부장/NNG]
상황실장 [상황실장/NNG]
비대위원장(비대위/NNG+원+장) [비대위원장/NNG]
→ 이 단어는 '비대위'에 구성원을 나타내는 접사 '-원', 그리고 우두머리를 나타내는 접사 '-장'이 차례로 결합한 것으로 파악된다. 따라서 '비대위'에 '-원'이 결합하여 '비대위원'이 되고, 여기에 '-장'이 결합하여 '비대위원장'이 된 것으로 보아 전체를 하나의 명사로 분석한다. '영진위원장' 등도 마찬가지이다.

접사가 아니라 '소장, 원장, 지사'와 같은 명사가 결합하여 단체장명이 만들어지는 경우가 있다. 그런 경우에는 아래와 같이 분석한다.

[예시] 헌법재판소소장 [헌법/NNG+재판소/NNG+소장/NNG]
민족문화연구원원장 [민족/NNG+문화/NNG+연구원/NNG+원장/NNG]
서울시의원 [서울시/NNP+의원/NNG]
제주도지사 [제주/NNP+도지사/NNG]
→ '제주도+지사'로 분석할 수도 있으나, 합성어 주의사항 ⑥에 따라 뒤쪽에 더 많은 음절수가 남는 '제주+도지사'를 선택함.

(6) 책, 연극, 영화, 음악 등 창작물의 제목

창작물의 제목이 한 어절에 나타난 경우 통합하여 고유명사로 분석한다.

[예시] 삼국사기/NNP, 손자병법/NNP, 고래사냥/NNP, 동이/NNP, 봉오동전투/NNP(영화 제목), 신과함께/NNP, 토지/NNP

(7) 신문, 잡지, 방송 채널, 웹사이트 등 매체의 이름

매체명이 한 어절에 나타난 경우 통합하여 고유명사로 분석한다. 두 어절 이상으로 분리되어 나온 경우에는 각 어절에 맞는 형태표지를 부여한다. 단, 외국어로 구성된 매체명이 여러 어절로 나타났을 때는 아래 다)의 외국어 처리 지침에 따라 모든 어절을 고유명사로 처리한다.

[예시] 조선일보/NNP, 여성동아/NNP, 폭스뉴스/NNP, 티브이엔/NNP, 보람TV/NNP(개인 채널명)
[예시] 유튜브/NNP, 트위터/NNP, 네이버/NNP, 페이스북/NNP
[예시] 스포츠 서울 [스포츠/NNG, 서울/NNP]
→ 전체 외국어 구성이 아님: 분리된 각 단어에 적합한 형태표지를 부여함.
[예시] 뉴스 위클리 [뉴스/NNP, 위클리/NNP]
스포츠 투데이 [스포츠/NNP, 투데이/NNP]
뉴욕 타임즈 [뉴욕/NNP, 타임즈/NNP]
→ 전체 외국어 구성인 경우: 외국어 지침에 따름

(8) 언어명

언어명의 경우 '-어'의 형태만을 통합하여 고유명사로 인정한다. '한국말'과 같은 경우는 일반 명사로 분석한다.

[예시] 한국어/NNP, 일본어/NNP, 영어/NNP, 알타이어/NNP, 네덜란드어/NNP

(9) 유물명, 식물명, 동물명

유물명과 아래와 같은 식물명, 동물명은 전체를 묶어 일반명사로 취급한다.

[예시] 청자상감국화무늬긴목병/NNG
북부점박이올빼미/NNG

(10) 위에서 명시되지 않은 부류는 모두 고유명사로 인정하지 않는다.

[예시] 임진왜란 (사건명) [임진왜란/NNG]
노벨평화상 [노벨/NNP+평화상/NNG]
네안데르탈인 [네안데르탈인/NNG]
메이지 (연호) [메이지/NNG]
고양국제꽃박람회 (행사명) [고양/NNP+국제/NNG+꽃/NNG+박람회/NNG]
한국시리즈 (대회명) [한국/NNP+시리즈/NNG]

다) 외국어의 처리

외국어는 아래의 방식에 따라 처리한다.

(1) <우리말샘>에 한 단어로 등재된 외국어

전체를 묶어 <우리말샘>에 등재된 품사로 처리한다. 그 외의 경우(<우리말샘>에 등재되지 않았거나 구로 등재된 것, <우리말샘>에 한 단어로 등재되었지만 원문에서 여러 어절로 분리되어 나타난 것)에 대한 처리는 아래의 (2)에 따른다.

[예시] 가든파티 (사전: 가든-파티) [가든파티/NNG]
[예시] 마추픽추 (사전: 마추픽추) [마추픽추/NNP]

(2) <우리말샘>에 등재되지 않았거나 구로 등재된 외국어

(가) 한 어절로 나타났든 두 어절 이상으로 나타났든 **전체 외국어 표현이 본 지침의 고유명사 부류에 든다면,** 그 고유명사를 이루는 **각 어절 모두를 (어절 내부 분석 없이) NNP로** 처리한다.

[예시] 레드제플린 (록 그룹 이름) [레드제플린/NNP]
레드 제플린 [레드/NNP, 제플린/NNP]
[예시] 카미노데산티아고 (지명, 순례길 이름) [카미노데산티아고/NNP]
카미노 데 산티아고 [카미노/NNP, 데/NNP, 산티아고/NNP]
[예시] 로버트다우니주니어 (인명) [로버트다우니주니어/NNP]
로버트 다우니 주니어 [로버트/NNP, 다우니/NNP, 주니어/NNP]
페르디낭 드 소쉬르 [페르디낭/NNP, 드/NNP, 소쉬르/NNP]
cf) 루이9세 [루이/NNP+9/SN+세/NNB]
루이 9세 [루이/NNP, 9/SN+세/NNB]
[예시] 웻 앤 와일드 워터월드 (시설물명) [웻/NNP, 앤/NNP, 와일드/NNP, 워터월드/NNP]
[예시] 블루 이즈 더 워미스트 컬러 (영화제목)
[블루/NNP, 이즈/NNP, 더/NNP, 워미스트/NNP, 컬러/NNP]

(나) 위의 경우가 아니라면, 외국어를 포함하고 있는 각각의 어절에 대하여 다음과 같은 절차를 적용하여 처리한다.

(나-1) **각각의 어절 속에 포함된 외국어 요소가 한 단어라면,** <우리말샘> 등재 여부와 무관하게 그 단어를 의미에 따라 **고유명사 또는 일반명사로** 분석한다.

[예시] 마이너하다 [마이너/NNG+하/XSA+다/EF]

→ 미등재어인 '마이너'는 '마이너 감성', '메이저와 마이너의 경계' 등에서 볼 수 있듯이 다른 말과도 결합하여 쓰이므로 어근보다는 명사의 성격을 띠는 단어로 볼 수 있다. 외국어의 이런 특성을 고려하여, 한 단어에 해당하는 미등재 외국어를 어근(XR)이 아닌 체언류로 처리한다.

[예시] 마이너 리그 [마이너/NNG, 리그/NNG]

[예시] 라 리가 (축구 리그) [라/NNG, 리가/NNG]

(나-2) **한 어절 속에 포함된 외국어 요소가 둘 이상의 단어일 때에는** 그 외국어 요소를 **단어 단위로 분리한다.** 단어 단위를 판단할 때에는 해당 외국어에서의 표기법(띄어쓰기 여부)을 참고한다. 그 후 아래의 지침을 따른다.

[예시] 배팅글러브 [배팅, 글러브]

[예시] 아시안게임 [아시안, 게임]

[예시] 리우올림픽 [리우, 올림픽]

[예시] 보이그룹 [보이, 그룹]

[예시] 라리가 (축구 리그) [라, 리가]

(나-3) 분리된 각 단어가 본 지침의 고유명사 부류에 들거나 <우리말샘>에 단독 일반명사로 등재되어 있어서 **각각의 단어를 따로 처리할 수 있는 상황이라면, 각 단어를 분리하여 형태표지를 부여**한다.

[예시] 배팅글러브 ('배팅' 등재, '글러브' 등재) [배팅/NNG+글러브/NNG]

[예시] 아시안게임 ('아시안' 고유명사, '게임' 등재) [아시안/NNP+게임/NNG]

→ '아시안', '아메리칸', '브리티시' 등 고유명사의 형용사형을 모두 고유명사로 처리한다.

[예시] 리우올림픽 ('리우' 고유명사, '올림픽' 등재) [리우/NNP+올림픽/NNG]

(나-4) 분리된 각 단어 중 어느 하나라도 위의 방식에 따라 고유명사로 또는 일반명사로 **처리할 수 없다면,** 각 단어를 분리하지 않고 **전체를 묶어서 의미에 따라 고유명사 또는 일반명사로** 분석한다.

[예시] 보이, 그룹: '그룹'은 팀의 의미로 단독으로 등재되어 있으나, '보이'는 '소년'의 의미로서는 단독으로 명사로 등재되지 않음. '보이'의 처리가 어려우므로 전체를 묶어 일반명사로 분석함. [보이그룹/NNG]

[예시] 라, 리가: '라'와 '리가' 모두 고유명사 또는 일반명사로 처리하기 어려움. 대회명은 본 지침에서 고유명사에 들지 않으므로 전체를 묶어 일반명사로 분석함. [라리가/NNG]

(나-5) 위에 제시한 절차를 외국어를 포함하고 있는 모든 어절에 각각 적용한다. 예시는 다음과 같다.

[예시] 피지컬 트레이닝	[피지컬/NNG, 트레이닝/NNG]
[예시] 워터해저드	[워터해저드/NNG]
[예시] 골든 커리어 그랜드슬램 (기록명)	[골든/NNG, 커리어/NNG, 그랜드슬램/NNG]
[예시] 글로벌 북카페 (신문 코너명)	[글로벌/NNG, 북카페/NNG]
글로벌 북 카페	[글로벌/NNG, 북/NNG, 카페/NNG]

(다) 아래와 같이 **외국어의 '한 문장'이 한글로 전사되어 나타난 경우, 각 어절을 내부 분석 없이,** 그리고 각 단어의 <우리말샘> 등재 여부와 무관하게 **NA로** 처리한다.

[예시] 렛츠고	[렛츠고/NA]
[예시] 익스큐즈 미	[익스큐즈/NA, 미/NA]
[예시] 아이 러브 유	[아이/NA, 러브/NA, 유/NA]
[예시] 굿!	[굿/NA+!/SF]
[예시] 곤니치와	[곤니치와/NA]
[예시] 니하오	[니하오/NA]
[예시] 해피버스데이 투 유	[해피버스데이/NA, 투/NA, 유/NA]

라) 의존명사(NNB)

의존명사는 자립해서 쓰일 수 없는 명사로, 수식 성분을 반드시 동반해야 한다. 의존명사는 비단위성 의존명사와 단위성 의존명사로 나뉠 수 있으나, 본 분석에서는 이를 세분하지 않는다. 의존명사와 일반명사의 구분은 <우리말샘>에 따른다.

(1) 의존명사와 일반명사의 구분

(가) '연대, 연도'는 '년대, 년도'와 달리 일반명사이다.

[예시] 연도별로 정리된 자료	[연도/NNG]
몇 년도에 일어난 일	[년도/NNB]

(나) '월, 연, 일, 주, 달러, 원' 등은 독립되어 쓰일 경우 모두 일반명사의 자격을 가지므로 일반명사로 분석해야 한다.

[예시] 나는 월 30만원을 받는다.	[월/NNG]
달러의 가치는	[달러/NNG]
시간당 만원을 받는다.	[시간/NNG+당/XSN]

주의사항

'원'은 <우리말샘>에 의존명사로만 올라 있지만, 독립되어 쓰인 경우에는 일반명사로 분석한다. 다른 유사 경우도 이에 따라 처리한다.

[예시] 1달러를 원으로 환산하면	[원/NNG+으로/JKB]

골프에서 '2타를 줄였다' 등으로 쓰이는 '타'가 있는데, 이 말이 사전에 올라 있지 않다. 본 분석에서는 이와 같은 '타'를 일반명사로 분석한다.

[예시] 2타를 줄여	[2/SN+타/NNG+를/JKO]

(2) 단위를 나타내는 표현

(가) 길이, 무게, 수효, 시간 따위의 수량을 수치로 나타내는 단위들 중 '미터, 그램, 리터' 등은 의존명사(NNB)로, 외국어로 된 'm, g, l' 등은 기호(SW)로 분석한다.

(나) 일반명사가 단위적인 용법으로 쓰인 경우에는 의존명사가 아니므로 주의한다.

[예시] 사람, 그릇…	
한 사람이 교실로 들어왔다.	[사람/NNG+이/JKS]
자장면 한 그릇만 주세요.	[그릇/NNG+만/JX]

(3) '것'과 구어형 '거'의 분석

'거'의 형태를 그대로 인정하여 분석한다.

[예시] 공부할 거를 준비해 왔니? [거/NNB+를/JKO]
공부할 걸 가져왔니? [거/NNB+ㄹ/JKO]
연습할 건 있니? [거/NNB+ㄴ/JX]
먹을 게 모자라다. [거/NNB+이/JKS]

2) 대명사(NP)

대명사는 그 자체로는 자신의 본유적 지시물을 가지지 않은 채, 다만 사람이나 사물 등 어떤 대상을 간접적으로 지시하는 품사이다. 단, 동일한 대명사가 방언이나 고어의 이형태를 가진 경우에는 이들도 대명사로 같이 분석한다.

가) 인칭 대명사

(1) 1인칭 대명사

[예시] 나, 내, 우리, 저, 제, 저희

(2) 2인칭 대명사

[예시] 너, 네, 그대, 당신, 댁

(3) 기타 대명사

[예시] 이이, 이분, 그이, 그분, 저이, 저분, 아무, 아무개, 누구, 무엇, 뭐, 뭐시기, 어디, 언제, 자기, 걔, 쟤, 얘, 이것, 저것, 그것, 이거, 저거, 그거, 여기, 저기, 거기, 이곳, 그곳, 저곳, 어디, 모(某), 모모(某某)

나) 대명사와 관형사의 두 가지 분석이 가능한 단어

(1) '모(某)'는 관형사와 대명사로 분석될 수 있으므로 주의를 요한다.

[예시] 모 기업체 [모/MMD]
김 모씨 [모/NP+씨/NNB]

(2) ‘모모(某某)’도 위와 같이 분석될 수 있다.

[예시] 모모가 말했다	[모모/NP+가/JKS]
모모 기관의 조사를 마쳤다	[모모/MMD]

다) 대명사의 이형태 분석

(1) ‘이것, 그것, 저것; 이거, 그거, 저거’는 분석하지 않고 대명사로 인정한다. ‘~거’의 경우, ‘~거’의 형태를 그대로 인정하여 분석한다.

[예시] 난 저거를 먹을래.	[저거/NP+를/JKO]
나는 여태 그걸 믿어 왔단다.	[그거/NP+ㄹ/JKO]

(2) 다음과 같이 원형을 밝힐 수 있는 대명사는 원형대로 분석한다.

[예시] 내	이제부터는 내 명령을 따라라.	[나/NP+의/JKG]
내게	내게 전자우편으로 알려 다오.	[나/NP+에게/JKB]
네게	어제 네게 보낸 선물이 잘못되었다.	[너/NP+에게/JKB]
제게	문제가 있다면 제게 말씀해 주세요.	[저/NP+에게/JKB]
누가	누가 전화를 하는지 보고해라.	[누구/NP+가/JKS]
뉘	뉘 집 애기가 이렇게 울고 있는 거야?	[누구/NP+의/JKG]
뭐가	도대체 뭐가 문제라는 거야?	[뭐/NP+가/JKS]

(3) ‘제’의 경우, ‘제/NP+가/JKS’를 제외하고는 모두 ‘저/NP+의/JKG’로 분석한다.

[예시] 제가 갈 것입니다.	[제/NP+가/JKS]
철수는 제 잘못을 안다.	[저/NP+의/JKG]
제 무게를 못 견디다.	[저/NP+의/JKG]

3) 수사(NR)

수사는 사물의 수량이나 차례를 나타내는 품사를 말한다.

가) 수사의 종류

(1) 양수사

[예시] 하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉, 열, 스물, 서른, 마흔, 쉰, 예순, 일흔, 여든, 아흔, 백
한둘, 두서넛, 서넛, 너덧, 네다섯, 네댓, 대여섯, 예닐곱, 일여덟, 일고여덟, 열두서넛, 열대여섯, 열일고여덟, 스물두서넛
일, 이, 삼, 사, 오, 육, 칠, 팔 구, 십, 백, 천, 만, 억, 조
기십, 기백, 기천,
수십, 수백, 수천, 수만, 수억, 수십만, 수백만, 수천만

(2) 서수사

[예시] 첫째, 둘째, 셋째, 넷째, …, 열째, 열한째,…, 스물한째,…,
아흔아홉째, 백째, 백한째, …

주의사항

① 값이 없는 수를 나타내는 '영', '공'은 사전에 명사로 올라 있으나 수사로 분석한다.

② 외국어로 된 말인 '제로, 원, 투, 쓰리' 등은 우리말 단위 명사와 잘 결합하지 않는 등 일반적인 수사와 다른 성격을 지니므로 일반명사로 분석한다.

③ 복수의 수사가 한 어절 내에 나타날 때에는 전체를 통합해서 분석한다.

[예시] 백만오천삼십사 [백만오천삼십사/NR]
공일공일이삼사오륙칠팔 [공일공일이삼사오륙칠팔/NR]
[참고] '일일구', '일일이' 등 <우리말샘>에 명사로 올라 있는 것은 명사로 분석한다.

④ '하나'는 <우리말샘>에 그 품사가 명사와 수사로 되어 있지만 본 지침에서는 수사로 분석한다.

[예시] 광에 가서 물건 하나만 가져오렴. [하나/NR+만/JX]
우리는 하나로 뭉쳤다. [하나/NR+로/JKB]

⑤ 때로 수사와 수관형사의 구별이 애매한 경우가 있다. 이 분석에서는 **다음과 같이 특이한 형식을 가진 예만을 수관형사로 취급하고**, 그 밖의 것들은 모두 수사로 분석한다. 사전에 수사와 관형사 동일 형태로 등재된 '몇'과 '몇몇' 역시 모두 수사로 분석한다. 이는 <우리말 샘>의 품사 처리와는 다른 방식임에 유의한다.

[예시] 한, 한두, 한두어, 두, 두어, 두세, 두서너, 세, 석, 서, 서너, 네, 너, 넉

⑥ 순서를 나타내는 '제일, 제이' 등은 접두사 '제-'와 수사의 결합으로 분석한다.

[예시] 제일, 제이, 제삼, 제사, 제오 … 제구십구, 제백… [제/XPN+일/NR], [제/XPN+이/NR], …

⑦ 순서를 나타내는 '첫째, 둘째' 등에 포함된 접미사 '-째'는 분석하지 않는다. '첫 번째', '두 번째' 등 의존명사 '번째'에 포함된 접미사 '-째'는 분석한다.

[예시] 첫째, 둘째, 셋째, 넷째, 다섯째, …, 아흔아홉째, … [첫째/NR], [둘째/NR], …
[예시] 첫 번째 [번/NNB+째/XSN]

나 용언

용언은 동사, 형용사, 지정사를 가리킨다. 용언 범주에서는 분석 대상이 본용언일 경우에만 동사와 형용사로 구분하여 표시하고, 보조용언의 경우에는 보조동사와 보조형용사를 구분하지 않고 'VX'라는 하나의 표지만을 준다. 또한 학교 문법에서 서술격조사로 다루는 '이다'는 조사의 범주에 넣지 않고 '지정사'라는 용언의 하위범주에 넣기로 한다. 지정사는 다시 긍정 지정사(VCP)와 부정 지정사(VCN)로 세분된다.

1) 동사(VV)

동사는 사물의 움직임이나 작용을 나타내는 용언을 말한다. 동사는 일반적으로 목적어의 필요성 여부에 따라 자동사, 타동사로 나누기도 하지만, 본 분석에서는 그것을 위한 별도의 표지를 세분하지 않고 모두 'VV'로 표시한다.

주의사항

'있다'는 동사 용법과 형용사 용법을 모두 가지고 있다. '있다'는 대개의 경우 형용사로 쓰이는 것으로 보아, '있다'가 동사로 쓰였다는 적극적인 증거가 있을 때에만 동사로 분석하고 나머지 경우는 형용사로 분석한다.

'-고 있다', '-어 있다'형으로 쓰여 앞의 사태가 진행/지속되고 있음을 나타내거나 앞의 사태가 끝나고 그 결과가 유지되고 있음을 나타낸다면 그때의 '있다'는 보조용언(VX)임에 유의한다.

형용사 '있다'의 특징

① '존재하다'의 뜻을 갖는 것은 형용사이다.

[예시] 신이 있다 / 날지 못하는 새도 있다 / 기회가 있다 / 증거가 있다
짜임새 있다 / 쓸모 있다 / 진정성 있다 / 경쟁력 있다 / 가능성 있다 / 필요 있다

② '수 있다', '바 있다', '적이 있다' 구성의 '있다'도 모두 형용사이다.

③ 종결어미 '-다'와 바로 결합하여 '있다.'형으로 쓰이면 형용사이다.

[예시] 이런 경우도 있다. / 그는 서울에 있다.

④ '~에 있어서' 구성의 '있다'도 형용사이다.

[예시] 인간에게 있어서 중요한 것은 사랑이다.

⑤ '누가 어떤 자격으로 있다' 구성의 '있다'도 형용사이다.

[예시] 그는 지금 대기업의 과장으로 있다.
그는 대기 선수로 있다가 출전권을 얻었다.

⑥ 내포문에서 '있다'가 쓰였을 때에는, 종결형으로 바꾸었을 때 '있다.'를 사용하여 표현할 수 있는 경우면 모두 형용사로 판단한다.

[예시] 서울광장에서 있었던 콘서트가 그 예이다.
→ '서울광장에서 콘서트가 있다.'로 바꾸어도 문장이 성립하므로, 형용사로 판단한다.
친구와 둘만 있는 상황이 되면...
→ '(나는 지금) 친구와 둘만 있다.'로 바꾸어도 문장이 성립하므로 형용사로 판단한다.

※ '머물다'의 뜻을 갖는다고 해서 모두 동사로 판단하지 않는다. '머물다'의 의미는 아래와 같이 사전에 동사로도, 형용사로도 기술되어 있다. 따라서 위 ⑥에서 언급했듯이 '있다.'형

으로 바꾸어도 문장이 성립되면 '머물다'의 뜻이어도 모두 형용사로 판단하기로 한다. '있다.'형으로 바꿀 수 없거나 '-는다', '-어라', '-자'처럼 동사와 결합하는 어미와 함께 나타났을 때에만 동사로 판단한다.

있다1 [I] 동사
「1」 사람이나 동물이 어느 곳에서 떠나거나 벗어나지 아니하고 머물다.
예) 그는 내일 집에 있는다고 했다.

있다1 [II] 형용사
「2」 사람이나 동물이 어느 곳에 머무르거나 사는 상태이다.
예) 그는 한동안 이 집에 있었다.

⑦ '있다'가 기간을 나타내는 부사어와 함께 쓰일 때에는 종결형 '있다.'를 사용하여 표현하는 것이 어색하다. 이런 경우 동사로 판단한다.

[예시] 그는 노쇠해서 이 자리에 오래 있기 힘들다.
→ '그는 이 자리에 오래 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.
1시간가량 조용히 있다가 갑자기 일어나 총을 꺼내 들었다.
→ '그는 1시간가량 조용히 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.
오늘은 덕수궁 지하도에 더 있다 갈게요.
→ '그는 덕수궁 지하도에 더 있다.'로 바꾸면 문장이 어색하므로 동사로 판단한다.

동사 '있다'의 특징

① '-는다'(평서), '-어라', '-자', '-읍시다' 등(명령, 청유)이 결합한 것은 동사이다.
② '얼마의 시간이 경과하다'의 뜻일 때에는 동사이다.

[예시] 10분 있다 만나자. / 얼마 안 있어 기다리던 시간이 왔다.

③ 아래 예시와 같이 '잘'과 결합한 '있다'는 동사로 판단한다. '잘 계세요'로 치환이 가능하다는 점에서 동사의 행태를 보이기 때문이다. 단, '내 그물이 잘 있나.'에서처럼 주어가 사람이 아닌 경우에는 동사로 판단하지 않는다.

[예시] (헤어질 때) 잘 있어요. / 아버지는 잘 있느냐. / 잘 있었니.
cf) 잘 계세요. / 아버지는 잘 계시느냐. / 잘 계셨습니까.

④ '-고 싶다', '-려(고) 하다'는 주로 동사와 결합하므로, 이와 결합한 '있다'는 동사로 판단한다.

[예시] 나도 그 자리에 있고 싶다.
나도 여기 있으려고 한다.

⑤ '가만히 있-', '마냥 있-'은, '가만있다'가 동사임을 참고하여, 또 '철수가 가만히 있다.', '철수가 마냥 있다.' 같은 표현이 빈번히 쓰이지 않는 것을 고려하여 동사로 판단한다.
단, 종결어미 '-다'가 바로 결합하여 '가만히 있다.', '마냥 있다.'로 쓰였다면 형용사로 판단한다.

⑥ '있은 지', '있은 후' 등에서 나타나는 '있은'의 '있-'은 동사로 판단한다. '-은'이 결합하여 과거를 나타내는 것이 동사의 특성이기도 하고, '-ㄴ 지', '-ㄴ 후' 등도 주로 동사와 결합하여 쓰이기 때문이다.

[예시] 그 일이 있은 지 수일이 지났다.

2) 형용사(VA)

형용사는 사물의 성질이나 상태를 나타내는 용언을 가리킨다.

주의사항

① 사전에 형용사로 등재된 단어가 동사와 같은 활용을 보일 때가 있다. 그러나 그럴 때에도 사전을 따라 형용사 형태표지를 부여한다.

[예시] 현실과 동떨어지는 문제가 있다. [동떨어지/VA+는/ETM]
→ '동떨어지다'는 <우리말샘>에 형용사로 등재되어 있다. 위의 예에서는 관형형 어미 '-는'과 결합하여 동사와 같은 활용 양상을 보여 주고 있으나, 본 지침에서는 이러한 경우에도 <우리말샘>의 품사를 따라 형용사로 분석한다.

② '못하다'는 <우리말샘>에 보조용언, 형용사, 동사 모두로 등재되어 있다. 따라서 용법에 맞게 품사를 구별하여 분석해야 한다.

[예시] 노래를 못한다.	[못/MAG+하/XSV+ㄴ다/EF+./SF]
[예시] 맛이 예전만 못하다.	[못/MAG+하/XSA+다/EF+./SF]
못해도 열 명은 올 것이다.	[못/MAG+하/XSA+아도/EC]
[예시] 밥을 먹지 못한다.	[못하/VX+ㄴ다/EF+./SF]
옳지 못하다.	[못하/VX+다/EF+./SF]
보다 못해 간섭을 했다.	[못하/VX+아/EC]

3) 보조용언(VX)

이 분석에서는 보조용언을 보조동사와 보조형용사로 하위 구분하지 않는다.

가) 보조용언 분석 원칙

(1) 보조용언의 후보는 <우리말샘>에 그 쓰임이 제시되어 있어야 한다.

(2) 보조용언 앞에는 반드시 다른 용언이 위치해 있어야 한다.

(3) 보조용언이 동시에 두 개 이상이 연결되어 나타날 수도 있다.

(4) 본용언과 보조용언의 결합형이 <우리말샘>에 하나의 어휘로 등재되어 있으면 보조용언을 따로 분석하지 않고 전체를 하나의 용언으로 처리한다. 특히 '-어하다', '-어지다' 결합형이 사전에 하나의 어휘로 올라 있는 경우가 많으므로 유의한다.

[예시] 아이를 예뻐하고	[예뻐하/VV+고/EC]
[예시] 눈이 동그래졌다	[동그래지/VV+었/EP+다/EF]

나) 보조용언의 예

보조용언의 예시는 다음과 같다. 이 목록은 <우리말샘>을 참고한 것이다.

가다	책을 다 읽어 간다.	[가/VX+ㄴ다/EF+./SF]

가지다	일을 그렇게 해 가지고는 기일을 맞출 수 없다.	[가지/VX+고는/EC]
계시다	손님께서 와 계십니다.	[계시/VX+ㅂ니다/EF+./SF]
나가다	정책을 추진해 나가는 과정에서 문제가 생겼다.	[나가/VX+는/ETM]
나다	일을 마치고 나니 상쾌하다.	[나/VX+니/EC]
내다	힘들겠지만 잘 견뎌 내야 한다.	[내/VX+아야/EC]
놓다	약속을 잡아 놓고 출장을 가다니	[놓/VX+고/EC]
달다	이번 시험 문제의 정답을 알려 다오.	[달/VX+오/EF+./SF]
대다	자꾸 졸라 대는 통에 그만 허락해 주고 말았다.	[대/VX+는/ETM]
두다	남겨 둔 돈도 이제 바닥이 났다.	[두/VX+ㄴ/ETM]
드리다	염려를 끼쳐 드려 송구하옵니다.	[드리/VX+어/EC]
들다	도무지 내 말은 믿으려 들지 않는다.	[들/VX+지/EC]
말다	어렵더라도 희망을 잃지 말아야 한다.	[말/VX+아야/EC]
먹다	나는 오늘도 약속을 잊어 먹었다.	[먹/VX+었/EP+다/EF+./SF]
못하다	그 참상을 차마 보지는 못할 것이다.	[못하/VX+ㄹ/ETM]
버리다	음식이 다 타 버렸다.	[버리/VX+었/EP+다/EF+./SF]
보다	이제는 새벽이 오는가 보다.	[보/VX+다/EF+./SF]
빠지다	썩어 빠진 생선을 사오다니	[빠지/VX+ㄴ/ETM]
싶다	너를 보고 싶다.	[싶/VX+다/EF+./SF]
쌓다	꼬치꼬치 물어 쌓는 통에 정신이 없었다.	[쌓/VX+는/ETM]
아니하다	일이 순리대로 풀리지 아니했다.	[아니하/VX+았/EP+다/EF+./SF]
않다	시간이 지나도 기차는 오지 않았다.	[않/VX+았/EP+다/EF+./SF]
오다	날이 밝아 온다.	[오/VX+ㄴ다/EF+./SF]
있다	그녀는 검정 옷을 입고 있었다.	[있/VX+었/EP+다/EF+./SF]
주다	아버지는 아기에게 동화책을 읽어 주었다.	[주/VX+었/EP+다/EF+./SF]
지다	평소보다 깨끗해진 내 방이 너무 좋다.	[깨끗하/VA+아/EC+지/VX+ㄴ/ETM]
치우다	다섯 명이 10인분의 식사를 먹어 치웠다.	[치우/VX+었/EP+다/EF+./SF]
터지다	끓인 지 오래 되어서 라면이 불어 터졌다.	[터지/VX+었/EP+다/EF+./SF]
하다	나귀를 쉬게 하는 것이 좋겠다.	[하/VX+는/ETM]

주의사항

① 다음과 같은 어절은 <우리말샘>에서 보조용언으로 취급되고 있으나, 여기서는 '의존명사+접사' 또는 '의존명사+보조용언'으로 분석한다. 이들 앞에는 항상 관형어가 온다는 분포적인 특성을 중시한 것이다.

[예시] 양하다/체하다/척하다/듯하다/법하다/뻔하다	[양/NNB+하/XSV+다/EF]
듯싶다	[듯/NNB+싶/VX+다/EF]

② <우리말샘>에서 보조용언으로 취급되는 '버릇하다'의 경우, 일반명사 '버릇'과 크게 구별되지 않으므로 '버릇'은 명사로 분석한다.

[예시] 자꾸 울어 버릇한다.	[버릇/NNG+하/XSV+ㄴ다/EF+./SF]

③ <우리말샘>에 등재된 보조용언이 준말 형태로 나타나는 경우, 그 준말 형태도 보조용언으로 분석한다.

[예시] 하고픈 거	[하/VV+고/EC+프/VX+ㄴ/ETM]
가는갑다	[가/VV+는가/EF+ㅂ/VX+다/EF]

4) 지정사(VC)

지정사는 학교 문법의 서술격 조사에 대응되는 것인데, 용언과 같이 활용한다는 특성을 중시한 술어이다. 여기서는 학교 문법의 '이다'를 긍정 지정사로, '아니다'를 부정 지정사로 하위 구분한다. 일반적으로 '아니다'는 형용사로 다루어지기도 하나, 여기서는 '아니다'가 '이다'의 부정 형이라는 점을 중시하여 '부정지정사'로 다룬다.

[예시] 철수는 매우 우수한 학생이다.	[학생/NNG+이/VCP+다/EF+./SF]
철수는 모범적인 학생이 아니다.	[아니/VCN+다/EF+./SF]

가) 지정사 '이/VCP'를 복원해야 하는 경우

(1) 체언에 어미가 직접 연결된 경우

[예시] 철수는 훌륭한 교사다. [교사/NNG+이/VCP+다/EF+./SF]

(2) 조사에 어미가 직접 연결된 경우

[예시] 우리가 그를 본 것은 서울에서다. [서울/NNP+에서/JKB+이/VCP+다/EF+./SF]

(3) '-였다'

[예시] 그 당시 나는 아이였다. [아이/NNG+이/VCP+었/EP+다/EF+./SF]

(4) 어미 '-라고, -라는, -라도, -라며, -라면서, -라서'

[예시] 나는 그에게 절교라고 말했다. [절교/NNG+이/VCP+라고/EC]
나는 친구라는 말이 좋다. [친구/NNG+이/VCP+라는/ETM]
"집에 간다"라는 말에 놀랐다. [가/VV+ㄴ다/EF+"/SS+이/VCP+라는/ETM]
→ '"집에 갈걸"이라는 말'을 [가/VV+ㄹ걸/EF+"/SS+이/VCP+라는/ETM]으로 분석하게 됨을 참고하여, '"집에 간다"라는 말'에서도 '이/VCP'를 복원한다. 다른 유사 경우도 마찬가지로 처리한다.
집에 간다라는 말에 놀랐다. [가/VV+ㄴ다/EF+이/VCP+라는/ETM]
나이가 어린 자라도 존중해 주어야 한다. [자/NNB+이/VCP+라도/EC]
그는 최고라며 나를 추켜 주었다. [최고/NNG+이/VCP+라며/EC]
"바보"라며 놀렸다. ["/SS+바보/NNG+"/SS+이/VCP+라며/EC]
그는 실수라면서 얼버무렸다. [실수/NNG+이/VCP+라면서/EC]
"밥을 먹고"라면서 화를 냈다. [먹/VV+고/EC+"/SS+이/VCP+라면서/EC]
너는 부자라서 우릴 이해하지 못할 것이다. [부자/NNG+이/VCP+라서/EC]

(5) 참고로, 아래와 같이 인용문 뒤에서 '하-'가 생략된 채 쓰인 '-며', '-는' 등은 '하-'의 복원 없이 형태 표지를 부여한다.

[예시] 얼마나 친절하냐?"며 [친절/NNG+하/XSA+냐/EF+?/SF+"/SS+며/EC]
얼마나 친절하냐?"는 [친절/NNG+하/XSA+냐/EF+?/SF+"/SS+는/ETM]

다 수식언

1) 관형사(MM)

관형사는 체언 앞에서 그것을 꾸미는 품사를 말한다. 관형사는 지시관형사(MMD), 수관형사(MMN), 성상관형사(MMA)로 세분하여 분석한다.

[예시] 한	한 가정	[한/MMN]
그까짓	그까짓 일	[그까짓/MMD]
그	그 문제	[그/MMD]
이	이 사람	[이/MMD]

가) 지시관형사(MMD)

'이, 그, 저'와 같이 발화 현장이나 문장 밖에 존재하는 대상을 가리키는 관형사를 지시관형사로 분석한다. '어느, 무슨, 웬'과 같이 정해지지 않은 것을 대신하는 관형사, '이내 신세'의 '이내'와 같이 인칭 의미를 나타내는 관형사도 지시관형사로 분석한다. 이 밖에 '귀(貴)'와 '본(本)'은 청자 측과 화자 측을 지시하고 '동(同)'은 공간을, '현(現)'과 '전(前)'은 시간을 지시한다는 점에서 지시관형사로 볼 수 있다.

[예시] 이, 그, 저, 요, 고, 조, 이런, 그런, 저런, 다른, 타(他), 어느, 무슨, 웬, 어떤, 아무, 아무런, 귀(貴), 본(本), 동(同), 현(現), 전(前), 모(某), 그까짓, 각(各), 매(每), 오른, 왼

나) 수관형사(MMN)

(1) 체언 앞에서 사물의 수량이나 차례를 나타내는 관형사를 수관형사로 분석한다. 단 '다섯, 여섯' 등 수사와 수관형사의 형태가 동일한 경우에는 수사로 분석한다.

[예시] 한, 두, 세/서/석, 네/너/넉, 닷, 엿, 스무, 한두, 두세, 서너, 두서너, 일이, 이삼, 삼사, 여러, 모든, 온, 온갖, 갖은, 전(全), 첫, 양(兩)

(2) 복수의 수사와 수관형사가 한 어절 내에 나타날 때에는 전체를 통합해서 수관형사로 분석한다.

[예시] 스물한	[스물한/MMN]
십수	[십수/MMN]

(3) '한'은 다음과 같이 수관형사 또는 성상관형사로 분석한다.

[예시] 책 한 권	[한/MMN] ('하나'의 뜻)
한 마을에 효자가 살고 있었다.	[한/MMN] ('어떤'의 뜻)
동생과 나는 한 이불을 덮고 잔다.	[한/MMN] ('같은'의 뜻)
한 20분쯤 걸었다.	[한/MMA] ('대략'의 뜻)

다) 성상관형사(MMA)

체언의 성질이나 상태를 나타내는 관형사를 성상관형사로 분석한다.

[예시] 새, 헌, 옛, 순(純), 구(舊), 주(主), 약(約), 양대(兩大), 만(滿) 10세, 단(單), 총(總)

주의사항

① '지시, 성상, 수' 중 어느 한 쪽으로 보기 힘든 관형사는 모두 '성상 관형사'의 테두리에 포함시킨다.

② 관형사는 때로 문맥에 따라 다른 품사로 분석될 가능성이 있으니 문맥을 잘 살펴서 분석해야 한다.

[예시] 관형사, 명사 통용	
전 학기에 장학금을 받았다.	[전/MMD]
그 사람을 전에 본 적이 있다.	[전/NNG]
[예시] 관형사, 부사 통용	
단 세 명이서 그 일을 꾸몄다.	[단/MMA]
단, 그 일은 해서는 안 된다.	[단/MAJ]
[예시] 관형사, 명사, 부사 통용	
이내 마음을 어찌 알리요.	[이내/MMD]
아침 들판에 이내가 끼었다.	[이내/NNG]
그는 이내 떠나갔다.	[이내/MAG]

③ 수사가 명사를 단독으로 수식하는 경우 그것을 관형사로 분석하기 쉬우나, '수'를 나타내는 말 가운데서 앞서 언급한 수관형사를 제외하고는 수사는 오로지 수사로만 분석한다. 즉, 수사와 관형사의 품사 통용을 인정하지 않는 것이다. 따라서 다음과 같이 '다섯'은 모든 환경에서 중의성 없이 '수사'로만 분석된다.

[예시] 다섯이 먹기에 충분하다. [다섯/NR+이/JKS]
다섯 명이 앉아 있었다. [다섯/NR]

2) 부사(MA)

부사는 주로 용언을 꾸며서 그 뜻을 더 세밀하고 분명하게 해 주는 품사를 말한다. 여기서는 부사를 세분하지 않고, 접속부사와 일반부사로만 나누기로 한다.

가) 접속부사(MAJ)

<우리말샘>에 등재된 접속부사만을 대상으로 접속부사 표지를 부여한다.

주의사항

① 접속부사는 종종 용언의 활용형으로도 쓰일 수 있으므로 주의한다.

[예시] 그래서 마지막에는 조심하라고 했지? [그래서/MAJ]
상황이 그래서 영희가 결석을 했구나. [그렇/VA+어서/EC]

② '그리고나서'의 분석

[예시] 그리고나서 [그리/MAG+하/XSV+고/EC+나/VX+아서/EC]

③ '그래도'는 용언의 활용형일 수도 있고 접속부사일 수도 있다. 두 용법을 구별하여 표지를 부여해야 하며, 용언의 활용형일 때는 아래와 같이 분석한다. 동사 '그러다'의 활용형인지 형용사 '그렇다'의 활용형인지가 불분명하고 두 가지 해석이 모두 가능할 때는 형용사 '그렇다'의 활용형으로 판단한다.

[예시] (누가) 그래도 [그러/VV+어도/EC]
(상황이) 그래도 [그렇/VA+어도/EC]

④ '그런데도'는 [그렇/VA+ㄴ데/EC+도/JX]로 분석한다.

⑤ 상대방의 말에 맞장구칠 때 쓰이는 '그러니까(요).'도 접속부사로 처리한다. 접속부사로부터 용법의 변화가 발생한 것으로 보이는 사례도 있으나 뒷말이 생략된 것으로 볼 수 있는 사례도 있음을 고려한 것이다.

[예시] A: 날씨가 너무 추워졌어요.
B: 그러니까요. [그러니까/MAJ+요/JX]

나) 일반부사(MAG)

주의사항

① 일반부사는 종종 일반명사와 동일형태를 띠고 있어 구분이 어려운 경우가 있다. 이들은 뒤에 조사가 결합하느냐의 여부와, 문맥에서 후행 명사를 수식하느냐의 여부에 따라 부사와 명사로 분석될 수 있다.

[예시] 너의 진짜 속셈이 무엇인지 말해 봐라. [진짜/NNG]
그 수학 문제는 진짜 어려웠다. [진짜/MAG]
지금이 공부하기 딱 좋은 때이다. [지금/NNG+이/JKS]
나는 지금 막 집에 도착했다. [지금/MAG]

② 부사적인 용법을 가졌음에도 불구하고 일반부사가 아닌 일반명사로만 <우리말샘>에 등재되어 있는 단어는 오로지 일반명사로만 분석한다.

[예시] 구석구석, 여기저기, 오랫동안, 이곳저곳, 좌우간, 처음, 최근

③ 일반부사로 분석하기 쉬운 활용상의 불완전동사인 '덩달아, 더불어'는 모두 동사로 옳게 분석해야 함에 주의한다.

[예시] 너는 덩달아 왜 난리니? [덩달/VV+아/EC]
우리 함께 더불어 살아가자. [더불/VV+어/EC]

④ '명사+없이'는 원칙적으로 '일반명사+없이/MAG'로 태깅하지만, 아래와 같이 하나의 단어로 굳어져 사전에 등재된 경우는 '없이' 통합형 자체를 하나의 일반부사로 분석한다.

[예시] 관계없이, 그지없이, 꾸밈없이, 끊임없이, 난데없이, 남김없이 등

라 독립언

1) 감탄사(IC)

감탄사는 화자의 부름이나 느낌, 놀람이나 대답을 직접적으로 나타내는 품사를 말한다.

[예시] 그럼, 야호, 어머, 앗, 아, 예, 그래, 아니(요), 글쎄, 참, 참나, 참내, 아이구, 와아, 오호, 세상에

주의사항

① 사람이 입으로 직접 내는 소리를 대상으로 하되, 흉내를 내는 의도가 없는 것과 본능적인 놀람이나 느낌을 나타내는 것을 대상으로 한다. 또한 감탄사와 혼동되는 부사로서 음성상징어류의 부사어가 있는데, 이는 감탄사가 아닌 일반부사로 분석한다.

[예시] 야호! 드디어 정상이다. [야호/IC+!/SF]
쿨럭쿨럭 기침을 했다. [쿨럭쿨럭/MAG]

② 동물의 울음소리 등은 감탄사가 아니라 일반부사로 분석한다.

[예시] 검둥이는 멍멍 짖으며 수풀 속으로 뛰어들어갔다. [멍멍/MAG]

③ <우리말샘>에 명사로 등재된 단어가 단독으로 쓰여 감탄사와 같은 용법을 보일 때가 있지만 그런 경우에도 <우리말샘>의 품사를 따라 명사로 분석한다.

[예시] 대박! [대/XPN+박/NNG+!/SF]
→ '대박'은 <우리말샘>에 명사로 올라 있으므로, 단독으로 쓰여 감탄사처럼 보이더라도 명사로 분석한다. 이때 '대'가 분리하여 분석해야 할 접두사에 해당하므로 접두사와 명사로 분리하여 분석한다.

④ 욕이나 욕설을 나타내는 말은 전체를 감탄사로 분석한다.

[예시] 빌어먹을! [빌어먹을/IC+!/SF]

⑤ <우리말샘>에 감탄사로 올라 있는 단어가 조사나 지정사 앞에서 쓰인 경우, 의미론적인 따옴의 효과가 있는 표현으로 볼 수 있으므로 감탄사로 분석한다. 다만 감탄사로서의 의미와 멀어진 채 조사나 지정사 앞에서 쓰인다면 명사로 분석한다.

[예시] 화이팅이에요. [화이팅/IC+이/VCP+에요/EF+./SF]
→ 감탄사가 지정사 앞에 쓰인 경우이다. 이때에도 감탄사로 분석한다.
[예시] 요즘 컨디션이 메롱이에요. [메롱/NNG+이/VCP+에요/EF+./SF]
→ '메롱'은 <우리말샘>에 놀림의 감탄사로 올라 있으나, 이 예에서는 놀림의 의미에서 멀어진 채 지정사 앞에서 쓰였다. 이런 경우에는 <우리말샘>의 품사와 달리 명사로 분석한다.

⑥ '뭐'는 문맥에 따라 대명사와 감탄사의 두 가지 쓰임이 있다.

[예시] 뭔지도 모른 채 [뭐/NP+이/VCP+ㄴ지/EF+도/JX]
신문에 뭐 대단한 특종이라도 실렸습니까? [뭐/IC]

⑦ 한 어절이 비정상적으로 늘어나거나 비정상적으로 늘어난 것에 다른 기호가 개입되었을 경우 분석불능 범주(NA)로 분석한다.

[예시] 그러어엄/NA, 으~어~이/NA

⑧ 구어에서 나타나는 담화 표지는 <우리말샘>을 참고로 하여 감탄사 표지(IC)를 부여한다. 물결표는 분석하지 않는다.

[예시] 저~, 음~, 저기~ [저/IC, 음/IC, 저기/IC]
어~, 그~ [어/IC, 그/IC]

마 관계언

조사는 주로 체언과 결합하여 다른 말과의 문법적 관계를 나타내거나, 특별한 뜻을 더해 주는 품사를 말한다. 조사는 그 수효가 많으므로 본 지침에서는 일부 사례만을 제시하였으며, 조사의 전체 목록은 <우리말샘>을 따르는 것을 원칙으로 한다. 조사는 크게 격조사, 보조사, 접속조사로 나뉘는데, 그 구분 역시 <우리말샘>을 따른다.

주의사항

한국어는 조사가 여러 개 결합하는 경우가 많은데, 조사 결합형은 아래와 같은 방식으로 세분 여부를 결정한다.

① 조사 결합형이 <우리말샘>에 등재되어 있지 않으면 각 조사를 분리하여 분석한다.

[예시] 부산에서도 대형 사고가 있었다. [부산/NNP+에서/JKB+도/JX] (‘에서도’ 미등재)
그녀와의 약속이 갑자기 잡혔다. [그녀/NP+와/JKB+의/JKG] (‘와의’ 미등재)

② 조사 결합형이 <우리말샘>에 등재되어 있으면, 사전의 뜻풀이를 참고하여 결합형 자체에 ‘격 조사’나 ‘보조사’라고 풀이되어 있으면 더 분석하지 않고 하나의 조사로 둔다.

[예시] 에다가 [에다가/JKB]
(사전: 일정한 위치를 나타내는 격 조사. 격 조사 ‘에’에 보조사 ‘다가’가 결합한 말이다.)

③ 만약 조사 결합형이 <우리말샘>에 등재되어 있는데 ‘어떤 조사와 어떤 조사가 결합한 말’로만 풀이되어 있으면 두 개의 조사로 분리하여 분석한다.

[예시] 에는 [에/JKB+는/JX]
(사전: 부사격 조사 ‘에’에 보조사 ‘는’이 결합한 말. 강조와 대조의 뜻을 나타내는 조사이다.)

1) 격조사(JK)

이는 체언과 다른 성분 간의 일정한 문법 관계를 나타내는 조사이다.

가) 주격조사(JKS)

선행 체언으로 하여금 주어가 되게 하는 조사이다.

이/가	산이 보인다.	[산/NNG+이/JKS]
	우리 둘이 갈게.	[둘/NR+이/JKS]
께서	선생님께서 오신다.	[선생/NNG+님/XSN+께서/JKS]
(이)서	둘이서 그 일을 꾸몄다고?	[둘/NR+이서/JKS]
	혼자서 그 일을 꾸몄다고?	[혼자/NNG+서/JKS]
께오서	부대장님께오서	[부대장/NNG+님/XSN+께오서/JKS]
께옵서	황제께옵서 드나드신다.	[황제/NNG+께옵서/JKS]

주의사항

'이서'의 경우, <우리말샘>에서는 '이'를 접미사로, '서'를 주격조사로 보고 있으나 여기에서는 '이서' 전체를 주격조사로 본다.

주격조사 '이/가'에 대하여 <우리말샘>에서는 '앞말을 지정하여 강조하는 뜻을 나타내는 보조사' 용법을 설정하고 있다. 여기에서는 보조적 연결어미 '-지' 뒤에 나온 '가'만을 보조사로 구별하여 분석한다.

[예시] 예쁘지가 않다.	[예쁘/VA+지/EC+가/JX]

나) 보격조사(JKC)

선행 체언으로 하여금 서술어 '되다, 아니다'의 보어가 되게 하는 조사이다. '되다, 아니다' 앞, 주어가 아닌 요소에 결합한 '이/가'를 보격조사로 분석해야 함에 유의한다.

이/가	얼음이 물이 되었다.	[물/NNG+이/JKC]
	철수는 범인이 아니다.	[범인/NNG+이/JKC]

다) 목적격조사(JKO)

선행 체언으로 하여금 목적어가 되게 하는 조사이다.

ㄹ/을/를	너는 바람소리를 들었다.	[바람/NNG+소리/NNG+를/JKO]

주의사항

목적격조사 'ㄹ/을/를'에 대하여 <우리말샘>에서는 '강조하는 뜻을 나타내는 보조사' 용법을 설정하고 있다. 여기에서는 보조적 연결어미 '-지' 뒤에 나온 'ㄹ/를'만을 보조사로 구별하여 분석한다.

[예시] 밥을 먹질 않는다.	[먹/VV+지/EC+ㄹ/JX]

라) 관형격조사(JKG)

선행 체언으로 하여금 관형어가 되게 하는 조사이다.

의	나의 친구는 너 하나뿐이다	[나/NP+의/JKG]

주의사항

구어에서 '의'가 '에'로 발음되어 '에'로 전사한 경우, 그 '에'는 관형격조사(JKG)로 분석한다.

[예시] 우리에 문제가 바로 그거야.	[우리/NP+에/JKG]

마) 부사격조사(JKB)

선행 체언으로 하여금 부사어가 되게 하는 조사이다.

(으)로	망치로 못을 박아야지.	[망치/NNG+로/JKB]
(으)로서	장관으로서 책임을 다해야 한다.	[장관/NNG+으로서/JKB]
(으)로써	돌로써 지붕을 만든다고?	[돌/NNG+로써/JKB]

같이	바보같이 웃고 다닌다.	[바보/NNG+같이/JKB]
더러	나더러 이것도 하라고 한다.	[나/NP+더러/JKB]
랑	너랑 많이 닮았다.	[너/NP+랑/JKB]
(으)로부터	TV로부터 받는 영향력이 너무 크다.	[TV/SL+로부터/JKB]
마냥	기영이마냥 놀 수만은 없다.	[기영이/NNP+마냥/JKB]
마따나	네 말마따나 나도 그래야 한다.	[말/NNG+마따나/JKB]
만큼	눈물만큼 콧물도 흐른다니까.	[눈물/NNG+만큼/JKB]
보고	영자보고 놀자고 좀 해라.	[영자/NNP+보고/JKB]
보다	직관보다는 논리가 동원돼야 한다.	[직관/NNG+보다/JKB+는/JX]
에	나는 너에 대해 아무것도 모른다.	[너/NP+에/JKB]
에게	너에게 말하기 싫다.	[너/NP+에게/JKB]
에게서	나는 철수에게서 그 말을 들었다.	[철수/NNP+에게서/JKB]
에서	집에서 학교까지 너무 멀다.	[집/NNG+에서/JKB]
에서부터	연구소에서부터 가게까지는 너무 멀다.	[연구소/NNG+에서부터/JKB]
와/과	경미와 함께 다닌다면,	[경미/NNP+와/JKB]
처럼	사람처럼 행동하는 동물이 있다.	[사람/NNG+처럼/JKB]
하고	그 일하고 관련된 사람은 아무도 없다.	[일/NNG+하고/JKB]

바) 호격조사(JKV)

주로 사람을 가리키는 체언 뒤에 연결되어 그것으로 하여금 부름의 대상이 되게 하는 조사이다.

아	호동아! 이제 그만 일어나거라.	[호동/NNP+아/JKV+!/SF]
야	철수야! 밥 먹어라.	[철수/NNP+야/JKV+!/SF]
여	주여, 우리에게 힘을 주소서.	[주/NNG+여/JKV+,/SP]
(이)시여	신이시여! 우리를 저버리지 마소서.	[신/NNG+이시여/JKV+!/SF]

주의사항

호격조사와 어말어미는 구분해서 분석해야 한다.

[예시] 저기 오는 것이 철수야.	[철수/NNP+이/VCP+야/EF+./SF]

사) 인용격조사(JKQ)

인용문이나 인용구를, 동사에 대한 부사적 성분으로 도입하는 조사이다.

고	그는 "이제 가도 좋다"고 말했다.	[좋/VA+다/EF+"/SS+고/JKQ]
(이)라고	"문제가 심각하다"라고 보고했다.	[심각하/VA+다/EF+"/SS+라고/JKQ]

주의사항

① 인용격조사는 연결어미와 구별하기 어려운 경우가 있으므로 주의한다.

[예시] 철수는 자기가 학생이라고 말했다.	[학생/NNG+이라고/JKQ] (×) [학생/NNG+이/VCP+라고/EC] (○)
철수는 “다음 주에 놀러 가도 좋다”고 말하였다.	[좋/VA+다/EF+”/SS+고/JKQ] (○) [좋/VA+다/EF+”/SS+고/EC] (×)

② 인용격조사는 형태만으로 확인할 수 없고 발화 상황까지 고려해야 하는 복잡한 표지이다. 게다가 인용격조사로 인정되는 형태인 ‘라고’ 등은 원래 용언의 활용형에 불과하다. 하지만 인용격조사를 설정하지 않을 경우에는 인용부호가 들어간 어절의 처리가 어색해진다. 따라서 우리는 인용격조사를 설정하되, 그 쓰임은 인용부호(”, ’,), },] , 〉 , …)가 있는 경우로만 제한하기로 한다. 물론 인용부호가 빠진 경우에는 어미로 분석한다.

[예시] 철수는 영희가 좋다고 말했다.	[좋/VA+다고/EC]

③ 명사 뒤에 따옴표와 ‘라고/이라고’가 이어지는 경우에도 따옴표 뒤의 ‘라고/이라고’를 인용격조사로 분석한다.

[예시] “그것이 우리의 목표”라고 말했다.	[목표/NNG+”/SS+라고/JKQ]

④ 단, 종결어미 + ‘라고’(직접 인용)의 경우에는 인용부호가 드러나지 않아도 조사로 분석한다.

[예시] 집에 간다라고 했다.	[가/VV+ㄴ다/EF+라고/JKQ] (○)
집에 간다라고 했다.	[가/VV+ㄴ다라고/EC] (×)
[참고] 집에 간다라는 말	[가/VV+ㄴ다/EF+이/VCP+라는/ETM]

⑤ 다음의 경우는 '이/VCP'가 생략된 것이므로 '이/VCP'를 복원하여 분석한다.

[예시] "집에 간다"라는 말에 놀랐다.	[가/VV+ㄴ다/EF+"/SS+이/VCP+라는/ETM]
"바보"라며	["/SS+바보/NNG+"/SS+이/VCP+라며/EC]
"밥을 먹고"라면서	[먹/VV+고/EC+"/SS+이/VCP+라면서/EC]

2) 접속조사(JC)

두 단어를 같은 자격으로 이어 주는 구실을 하는 조사를 말한다. 아래는 그 예시이다.

와	그 아주머니는 딸기와 사과를 샀다.	[딸기/NNG+와/JC]
과	그 기계는 사람과 컴퓨터를 구별하지 못한다.	[사람/NNG+과/JC]
나	사과나 배는 모두 몸에 좋은 과일이다 .	[사과/NNG+나/JC]
랑	머루랑 다래랑 먹으며 청산에 살고 싶어라.	[머루/NNG+랑/JC]
하고	이번 준비물로 칼하고 연필을 샀다.	[칼/NNG+하고/JC]

주의사항

'함께 함'의 뜻을 나타내는 접속조사는 부사격조사와 형태상 동일하므로 주의할 필요가 있다. 체언과 체언 사이에서 두 체언을 이어주는 요소는 접속조사이고, 그 외의 경우에는 부사격조사이다.

[예시] 철수와 영희가 왔다.	[철수/NNP+와/JC]
철수와 같이 놀았다.	[철수/NNP+와/JKB]

3) 보조사(JX)

체언이나 부사 또는 용언의 연결 어미나 종결 어미의 뒤에 쓰여 특별한 뜻을 더해 주는 조사를 말한다. 아래는 그 예시이다.

그려	먹습니다그려.	[먹/VV+습니다/EF+그려/JX+./SF]
까지(꺼정/까장)	너까지 나에게 이럴 줄이야.	[너/NP+까지/JX]
깨나	너도 사람깨나 울렸겠구나.	[사람/NNG+깨나/JX]
(이)나	너나 가라!	[너/NP+나/JX]
(이)나마	빵이나마 먹어라.	[빵/NNG+이나마/JX]
ㄴ/은/는	이 종이는 어제 사 온 것이다.	[종이/NNG+는/JX]
ㄴ커녕/은커녕/는커녕	돈은커녕 먹을 쌀도 없다.	[돈/NNG+은커녕/JX]
다	그 물건을 거기다 놓아라.	[거기/NP+다/JX]
다가	책상을 어디다가 둘까요?	[어디/NP+다가/JX]
대로(대루)	너는 너대로 살아라.	[너/NP+대로/JX]
따라	오늘따라 택시도 안 잡힌다.	[오늘/NNG+따라/JX]
도/두	강아지도 주인은 알아본다.	[강아지/NNG+도/JX]
(이)란	코알라란 호주에 사는 초식동물이다.	[코알라/NNG+란/JX]
만	인간은 빵만으로 살 수 없다.	[빵/NNG+만/JX+으로/JKB]
밖에	그래 봐야 죽기밖에 더 하랴.	[죽/VV+기/ETN+밖에/JX]
부터/부텀	우선 노인부터 태워라.	[노인/NNG+부터/JX]
뿐	가진 건 고작 집 한 채뿐.	[채/NNB+뿐/JX]
(이)야	그가 인간성이야 그만이지.	[인간/NNG+성/XSN+이야/JX]
요	나는요 그림을요 예쁘게 그립니다.	[나/NP+는/JX+요/JX]
조차	이젠 집조차 빼앗기는구나.	[집/NNG+조차/JX]
치고	값싼 물건치고 쓸 만하다.	[물건/NNG+치고/JX]

(1) 보조사 분석 기준

앞에 '이'가 개재될 수 있는 조사는 지정사 '이다'에 어미가 결합한 형태와 구분하기 어려운 경우가 있다. 본 분석에서는 <우리말샘>을 따라 '이'형 조사와 지정사 '이다'의 활용형을 구분하는 것을 원칙으로 한다. '라든지'처럼 <우리말샘>에서 조사로만 다루어지는 것은 조사로 처리하면 되지만, '든지'처럼 <우리말샘>에서 조사와 어미 모두로 등재되어 있는 것은 문맥과 <우리말샘>의 예문을 참조하여 조사인지 '이다'의 활용형인지를 판단해야 한다. 조사와 '이다' 활용형의

기본적인 구별 기준은 다음과 같다.

(가) '이-' 뒤에 '-시-'나 '-었-' 등의 선어말어미가 결합할 수 있으면 그 뒤의 요소는 어미이다. '이-' 뒤에 선어말어미가 결합할 수 없으면 전체가 '이'를 포함하는 조사이다.

(나) '체언+이-'의 주어를 상정할 수 있으면 그 뒤의 요소는 어미이다. '체언+이-'의 주어를 상정할 수 없으면 전체가 '이'를 포함하는 조사이다.

[예시] 학생이라도 지원할 수 있습니다. [학생/NNG+이/VCP+라도/EC]
cf) 학생이시라도 지원하실 수 있습니다.
cf) [철수가 학생이라도] 지원할 수 있습니다.
[예시] 사람이 부족하니 선생님이라도 빨리 오세요. [선생/NNG+님/XSN+이라도/JX]
cf) *선생님이시라도 빨리 오세요.
cf) *[당신이 선생님이라도] 빨리 오세요.

(다) 다음의 형태는 지정사 '이다'의 활용형과는 관계가 없으므로 모두 보조사가 된다.

[예시] 까지, 깨나, 는(은/ㄴ), 대로, 도, 따라, 마다, 마저, 만, 밖에, 부터, 뿐, 조차, 치고, ㄴ커녕

주의사항

① 다음의 형태들은 분석 결과에 중의성이 생기므로, 이들을 분석할 때는 특히 주의해야 한다.

[예시] (이)란 코알라란 동물은 호주에 주로 서식한다. [코알라/NNG+이/VCP+란/ETM]
코알라란 매우 귀여운 동물이다. [코알라/NNG+란/JX]
(이)나 밥이나 빵을 먹도록 해라. [밥/NNG+이나/JC]
밥이나 먹자. [밥/NNG+이나/JX]
그가 비록 열심히 하나 능력은 부족하다. [하/VV+나/EC]
어제 내가 술을 마셨나? [마시/VV+었/EP+나/EF+?/SF]
(이)야 철수야 당연히 그 일을 할 수 있지. [철수/NNP+야/JX]
내가 좋아하는 것은 철수야. [철수/NNP+이/VCP+야/EF+./SF]
철수야! 부르는 소리 [철수/NNP+야/JKV+!/SF]
(이)요 밥을 먹다가요 [먹/VV+다가/EC+요/JX]
밥이요 빵이요 [밥/NNG+이/VCP+요/EC]

② 구어에서 받침 있는 말 뒤에서 '요' 대신 쓰이는 '이요'는 보조사로 분석한다.

[예시] A: 넌 머 먹을래? B: 전 밥이요. [밥/NNG+이요/JX+./SF]

③ '종결어미+요(보조사)'는 <우리말샘>에 '어미'로 등재되어 있는 '어요, 지요, 래요' 등을 제외하고 모두 원래의 범주인 종결어미와 보조사로 분리하여 분석한다.

[예시] 우리 집에 갈까요? [가/VV+ㄹ까/EF+요/JX+?/SF]
어디서 저녁 먹나요? [먹/VV+나/EF+요/JX+?/SF]
빨리 공부해야지요. [공부/NNG+하/XSV+아야지/EF+요/JX+./SF]
[예시] 철수는 이제 집에 간대요. [가/VV+ㄴ대/EF+요/JX]
→ <우리말샘>에 '-ㄴ대요'가 등재되어 있지만, 이 예에서처럼 인용의 의미를 가지며 '-ㄴ다고 해요'가 줄어든 말인 경우에는 '어미'가 아니라 '줄어든 말'로서 올라 있다. <우리말샘>에 '어미'로 등재된 경우가 아니므로 종결어미와 보조사로 분리하여 분석한다.
[예시] 선생님, 철수는 장난친대요. [장난치/VV+ㄴ대요/EF]
→ 이 예에서는 '-ㄴ대요'가 인용의 의미 없이 남에게 일러바치거나 남을 놀리는 의미로 사용되었다. 이러한 용법의 '-ㄴ대요'는 <우리말샘>에 '어미'로 등재되어 있으므로, '-ㄴ대요' 전체를 종결어미로 처리한다.

④ '비종결어미+요(보조사)'는 통합하지 않고 각각 분석해 준다.

[예시] 제가 몸이 좀 아파서요 지각을 했어요. [아프/VA+아서/EC+요/JX]
내가요, 왜요 [내/NP+가/JKS+요/JX], [왜/MAG+요/JX]

⑤ '말고'는 용언 '말다'의 활용형으로 처리한다.

[예시] 돈말고 지혜가 필요하다. [돈/NNG+말/VV+고/EC]

바 의존형태

1) 어미(E)

가) 선어말어미(EP)

용언이 활용할 때, 어간과 어말 어미 사이에 나타나는 것으로 높임법이나 시제, 양태를 나타내는 문법적인 요소이다. 선어말어미의 목록은 연구자에 따라 다를 수 있으나 이 분석에서는 아래의 것만을 선어말어미로 인정한다.

-겠-	그 일은 내일 처리하겠다.	[처리/NNG+하/XSV+겠/EP+다/EF+./SF]
-(으)시-	선생님께서 손수 만드신	[만들/VV+시/EP+ㄴ/ETM]
-옵-	어머님께 선물을 바치옵고	[바치/VV+옵/EP+고/EC]
-았/었-	우리가 먹었던 음식에 문제가 있다.	[먹/VV+었/EP+던/ETM]
-았었/었었-	거기는 우리가 전에 갔었던 곳이야.	[가/VV+았었/EP+던/ETM]

주의사항

① 어간 '하-' 뒤에 과거 시제 선어말어미가 결합하여 '했'의 형태로 나타나거나 '하였'의 형태로 나타날 수 있는데, 본 분석에서는 이 경우 '-였-' 형태를 인정하지 않고 모두 '-았-'으로 분석한다. 이 외에 '아/어'를 포함하고 있는 모든 어미 역시, '하-' 뒤에 나타난 경우에는 '아X'형으로 분석한다.

② 다음의 선어말어미는 그 어간이 생략되었을 경우에 어간을 복원해 준다.

-겠-	이것은 그대로 두어야겠다.	[두/VV+어야/EC+하/VX+겠/EP+다/EF+./SF]
-았/었-	철수가 그것을 가져오랬다.	[가져오/VV+라/EF+하/VV+았/EP+다/EF+./SF]
-시-	선생님께서 가자시오.	[가/VV+자/EF+하/VV+시/EP+오/EF+./SF]

③ 위의 선어말어미가 포함되지 않은 어미 형태는 그대로 어미로 분석한다.

-랄까, -대야, -래야

나) 종결어미(EF)

용언의 어간이나 선어말어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 한 문장을 끝맺는 역할을 한다. 본 지침에서는 <우리말샘>에 따라 종결어미를 구분한다. 다음은 종결어미의 일부 사례이다.

-게	그만한 돈이 있으면 좋게.	[좋/VA+게/EF+./SF]
-ㄴ가	이것이 무엇인가?	[무엇/NP+이/VCP+ㄴ가/EF+?/SF]
-ㄴ걸	이제 시작인걸.	[시작/NNG+이/VCP+ㄴ걸/EF+./SF]
-ㄴ다	이건 말도 안 된다.	[되/VV+ㄴ다/EF+./SF]
-나	자네 그리로 가나?	[가/VV+나/EF+?/SF]
-냐	키가 얼마나 크냐?	[크/VA+냐/EF+?/SF]
-네	정말 큰일 났네!	[나/VV+았/EP+네/EF+!/SF]
-는걸	그는 벌써 갔는걸.	[가/VV+았/EP+는걸/EF+./SF]
-는구나	앞이 잘 안 보이는구나.	[보이/VV+는구나/EF+./SF]
-는구려	잘도 먹는구려.	[먹/VV+는구려/EF+./SF]
-는구먼	공부를 잘하는구먼.	[잘/MAG+하/XSV+는구먼/EF+./SF]
-는다	아이가 글을 잘 읽는다.	[읽/VV+는다/EF+./SF]
-다	그게 사실이다.	[사실/NNG+이/VCP+다/EF+./SF]
-ㄹ게	그렇게 할게.	[하/VV+ㄹ게/EF+./SF]
-ㅂ니까	이제야 옵니까?	[오/VV+ㅂ니까/EF+?/SF]
-ㅂ니다	이렇게 합니다.	[하/VV+ㅂ니다/EF+./SF]
-습니까	그래도 되겠습니까?	[되/VV+겠/EP+습니까/EF+?/SF]
-습니다	정말 재미있습니다.	[재미있/VA+습니다/EF+./SF]
-ㅂ시다	다시 만납시다.	[만나/VV+ㅂ시다/EF+./SF]
-ㅂ시오	서둘러 주십시오.	[주/VX+시/EP+ㅂ시오/EF+./SF]
-으냐	물이 얼마나 깊으냐?	[깊/VA+으냐/EF+?/SF]
-은가	그것이 좋은가?	[좋/VA+은가/EF+?/SF]
-오/으오/소	물이 깨끗하오.	[깨끗하/VA+오/EF+./SF]
-ㅂ디다/습디다	참 좋은 곳입디다.	[곳/NNB+이/VCP+ㅂ디다/EF+./SF]
-거든	나는 이것이 좋거든!	[좋/VA+거든/EF+!/SF]
-ㄴ걸/은걸	힘이 꽤 센걸.	[세/VA+ㄴ걸/EF+./SF]
-ㄹ걸/을걸	모른다고 할걸.	[하/VV+ㄹ걸/EF+./SF]
-ㄹ까	이제 밥을 할까?	[하/VV+ㄹ까/EF+?/SF]

-다오	그가 가지고 있다오.	[있/VX+다오/EF+./SF]
-다네	일을 망쳤다네	[망치/VV+었/EP+다네/EF+./SF]
-다구	돈이 많다구?	[많/VA+다구/EF+?/SF]
-다니까	돈이 없다니까!	[없/VA+다니까/EF+!/SF]
-냐고/느냐고	그가 누구냐고?	[누구/NP+이/VCP+냐고/EF+?/SF]
-도다	꽃이 아름답도다.	[아름답/VA+도다/EF+./SF]
-다니	그가 책을 읽다니!	[읽/VV+다니/EF+!/SF]
-는가	같이 가겠는가?	[가/VV+겠/EP+는가/EF+?/SF]
-ㅂ디까/습디까	보기에 좋습디까?	[좋/VA+습디까/EF+?/SF]
-다면서	술은 싫다면서?	[싫/VA+다면서/EF+?/SF]
-다나	그도 가겠다나.	[가/VV+겠/EP+다나/EF+./SF]
-렴/으렴	맘대로 해 보렴.	[보/VX+렴/EF+./SF]
-려무나	책이나 읽으려무나.	[읽/VV+으려무나/EF+./SF]
-라니까	그 사람이 아니라니까.	[아니/VCN+라니까/EF+./SF]
-세	일이나 하세.	[하/VV+세/EF+./SF]
-자꾸나	약속을 좀 미루자꾸나.	[미루/VV+자꾸나/EF+./SF]
-자니까	그만 따지자니까.	[따지/VV+자니까/EF+./SF]
-아/어/야	밥 먹어!	[먹/VV+어/EF+!/SF]
-ㅁ세/음세	그날 꼭 옴세.	[오/VV+ㅁ세/EF+./SF]
-단다	애들이 다쳤단다.	[다치/VV+었/EP+단다/EF+./SF]
-더라고	아까 보니 철수가 집에 가더라고.	[가/VV+더라고/EF+./SF]

주의사항

① '종결어미+요(보조사)'는 <우리말샘>에 '어미'로 등재되어 있는 '어요, 지요, 래요' 등을 제외하고 모두 종결어미와 보조사로 분리하여 분석한다.

[예시] 말씀대로 했는걸요.	[하/VV+았/EP+는걸/EF+요/JX+./SF]

② '-세요'는 다음과 같이 선어말어미까지 분석한다.

[예시] 어서 출근하세요.	[출근/NNG+하/XSV+시/EP+어요/EF+./SF]

③ ‘-죠’는 축약형을 그대로 태깅한다. 단, 종결어미 ‘-어야지’와 ‘요’가 결합하여 ‘-어야죠’ 형식으로 나왔을 때는 ‘-지’와 ‘요’를 분리한다.

[예시] 어서 출근하죠. [출근/NNG+하/XSV+죠/EF+./SF]
어서 출근해야죠. [출근/NNG+하/XSV+아야지/EF+요/JX+./SF]

④ “앞의 사실을 청자가 이미 알고 있음”을 나타내는 ‘잖’은, ‘-더-’, ‘-는-’ 등과 같이 본 지침상 분석하지 않는 선어말어미처럼 취급하여 다음과 같이 어말어미와 결합하여 표지를 부여한다.

[예시] 저 오늘 일찍 일어났잖아요. [일어나/VV+았/EP+잖아요/EF+./SF]
제가 갔잖습니까. [가/VV+았/EP+잖습니까/EF+./SF]

단, ‘하’ 생략과 함께 ‘지 않’이 줄어들어서 나타난 ‘잖’ 형은 ‘하’와 함께 ‘지 않’을 복원하여 분석한다. 이때의 ‘잖’은 용언 어간이나 선어말어미 뒤에 붙어 “앞의 사실을 청자가 이미 알고 있음”을 나타내는, 선어말어미에 준하는 ‘잖’이 아님에 유의한다.

[예시] 녹록잖은 일이다. [녹록하/VA+지/EC+않/VX+은/ETM]

⑤ ‘-려고’는 <우리말샘>에서 의심과 반문의 용법으로만 종결어미 자격을 갖는 것으로 등재되어 있다. 하지만 아래와 같이 뒤에 생략된 말 없이 주어의 의도만을 밝히며 문말에서 쓰이는 ‘-려고’는 종결어미 용법으로 볼 수 있으므로 종결어미로 분석한다.

[예시] 나는 오늘 집에 일찍 가려고. [가/VV+려고/EF+./SF]
→ ‘-려고’ 뒤에 ‘생각하다’ 정도의 동사가 생략되어 있다. 이때는 주어의 의도가 무엇인지만을 밝히며 문말에서 쓰인 ‘-려고’로 볼 수 있으며, 종결어미로 분석한다.
[예시] 나 요즘 매일 운동해. 살 빼려고. [빼/VV+려고/EC+./SF]
→ ‘-려고’ 뒤에 ‘생각하다’가 생략된 것이 아니며 ‘운동하다’와 같이 주어의 의도를 실현하기 위한 행동을 나타내는 말이 생략되어 있다. 이때는 ‘-려고’가 연결어미로 쓰인 것이다.

다) 연결어미(EC)

용언의 어간이나 선어말어미 뒤에 연결되어 용언의 형식을 완성시키는 어미로서 문장을 종결

시키지 못하고 뒤에 오는 절을 연결시켜 주는 어미를 말한다. 본 지침에서는 <우리말샘>에 따라 연결어미를 구분하는 것을 원칙으로 한다. 다음은 연결어미의 일부 사례이다.

-거나	누가 오거나 알은 체 할 것 없다.	[오/VV+거나/EC]
-거든	거기 가거든 김 사장이 있는지 보아라.	[가/VV+거든/EC]
-건대	내가 보건대, 네 말이 옳다.	[보/VV+건대/EC]
-건마는	말렸건마는 아직도 축축하다.	[말리/VV+었/EP+건마는/EC]
-게	개를 굶게 하지 마라.	[굶/VV+게/EC]
-고	일을 하고 밥을 먹자.	[하/VV+고/EC]
-곤	숙제한 것도 빌려가곤 한다.	[빌리/VV+어/EC+가/VV+곤/EC]
-기에	늦게라도 왔기에 용서해 주었다.	[오/VV+았/EP+기에/EC]
-ㄴ다기에	잠시 쉰다기에 승낙했다.	[쉬/VV+ㄴ다기에/EC]
-ㄴ다손/다손	밉다손 치더라도 구박하지 말자.	[밉/VA+다손/EC]
-ㄴ들/는들	간다 한들 아주 갈까?	[하/VV+ㄴ들/EC]
-ㄴ즉	배가 고픈즉 속이 쓰리다.	[고프/VA+ㄴ즉/EC]
-ㄴ지라/는지라	눈이 온지라 길이 미끄럽다.	[오/VV+ㄴ지라/EC]
-나	눈이 오나 비가 오나 같다.	[오/VV+나/EC]
-나마	맛이 좋지 못하나마 많이 드십시오.	[못하/VX+나마/EC]
-는다기에	빵을 먹는다기에 주었다.	[먹/VV+는다기에/EC]
-니까	너를 보니까 좋다.	[보/VV+니까/EC]
-다가	자랑하다가 망신당했다.	[자랑/NNG+하/XSV+다가/EC]
-다기에	꽃이 예쁘다기에 보러 왔소.	[예쁘/VA+다기에/EC]
-대도	시간이 있대도 만나 주질 않는다.	[있/VA+대도/EC]
-더라도	가더라도 꼭 돌아와라.	[가/VV+더라도/EC]
-던들	진작 알았던들 방법을 취했지.	[알/VV+았/EP+던들/EC]
-든지	외모가 어떠하든지 무슨 상관인가?	[어떠하/VA+든지/EC]
-ㄹ뿐더러	비가 올뿐더러 바람도 분다.	[오/VV+ㄹ뿐더러/EC]
-ㄹ수록	높이 올라갈수록 춥다.	[올라가/VV+ㄹ수록/EC]
-ㄹ지	비가 얼마나 올지 천둥이 다 친다.	[오/VV+ㄹ지/EC]
-ㄹ지라도	이길지라도 명예롭지는 않다.	[이기/VV+ㄹ지라도/EC]
-ㄹ지언정	죽을지언정 그 일은 못하겠다.	[죽/VV+을지언정/EC]
-라고	철수는 자기가 바보라고 생각한다.	[바보/NNG+이/VCP+라고/EC]
-락	자락 깨락 잠을 설쳤다.	[자/VV+락/EC]

-랍시고	그는 반장이랍시고 거드름만 피운다.	[반장/NNG+이/VCP+랍시고/EC]
-려니와	비용도 문제려니와 일꾼도 문제다.	[문제/NNG+이/VCP+려니와/EC]
-련마는	보면 반가우련마는 볼 수가 없네.	[반갑/VA+으련마는/EC]
-면	지옥이 존재하면 만원일 것이다.	[존재/NNG+하/XSV+면/EC]
-면서	푸르면서 검은 물빛	[푸르/VA+면서/EC]
-므로	비가 오므로 가지 않겠다.	[오/VV+므로/EC]
-아/어	입을 막아 버렸다.	[막/VV+아/EC]
-아도/어도	암만 봐도 모르겠다.	[보/VV+아도/EC]
-아서/어서	덫을 놓아서 꿩을 잡았다.	[놓/VV+아서/EC]
-아야	이 일은 잘해야 한다.	[잘/MAG+하/XSV+아야/EC]
-으나	밥을 먹으나 마나이다.	[먹/VV+으나/EC]
-으나마	맛은 없으나마 많이 드세요.	[없/VA+으나마/EC]
-자마자	집에 오자마자 씻었다.	[오/VV+자마자/EC]
-지	밥을 먹지 못했다.	[먹/VV+지/EC]
-지마는	비가 오지마는 가야 한다.	[오/VV+지마는/EC]

주의사항

① 어미에 따라서는 분석의 중의성이 생길 수 있으므로 문맥 확인을 통해 형태분석을 결정한다.

[예시]	너는 내가 왔는데 기쁘지도 않니?	[오/VV+았/EP+는데/EC]
	철수가 있는데가 어디지?	[있/VA+는/ETM+데/NNB+가/JKS]
	다들 만족하는지 아무런 불평이 없다.	[만족/NNG+하/XSV+는지/EC]
	다들 만족하는지는 모르겠다.	[만족/NNG+하/XSV+는지/EF+는/JX]
	너를 만난지도 꽤 오래구나.	[만나/VV+ㄴ/ETM+지/NNB+도/JX]

'-을까', '-는가, -은가'는 언제나 종결어미임에 유의한다. '누군가', '어딘가' 등도 '누구/NP+이/VCP+ㄴ가/EF', '어디/NP+이/VCP+ㄴ가/EF' 등으로 분석한다.

'-을지', '-는지, -은지'는 연결어미 용법과 종결어미 용법을 모두 갖는데, 뒤에 조사가 오거나 '모르다'의 목적어 자리에서 쓰이는 경우 종결어미 용법에 해당한다는 점에 유의한다.

② 통사적 구성에 나타나는 '-음직'은 '음직/EC'로 분석한다. 그러나 '바람직하다, 먹음직하다'

등 사전에 등재되어 있는 단어의 내부에서 확인되는 '-음직'은 더 이상 분석할 수 없다는 것에 유의한다.

[예시] 철수라면 외국에 갔음직 하다.	[가/VV+았/EP+음직/EC 하/VA+다/EF+./SF]
어른답고 믿음직하게 행동해라.	[믿음직하/VA+게/EC]
그것 참 먹음직스럽다.	[먹음직스럽/VA+다/EF+./SF]
그것은 매우 바람직한 일이다.	[바람직하/VA+ㄴ/ETM]

라) 명사형전성어미(ETN)

한 문장의 성격을 임시로 바꾸어 다른 문장 속에서 명사적인 역할을 하게 하는 어미를 말한다.

-기	그 일은 정말 중요하기 때문이다.	[중요/NNG+하/XSA+기/ETN]
-ㅁ/-음	장사는 신용을 얻음이 제일이다.	[얻/VV+음/ETN+이/JKS]

주의사항

① ㅂ 불규칙 용언 어간에 명사형 전성 어미 '-음'이 결합한 경우 '-음'이 아닌 '-ㅁ'으로 분석한다.
ㅅ 불규칙 용언 어간에 결합하는 '-음'은 '음/ETN'으로 분석한다.

[예시] 아니꼬움을 견디지 못하고	[아니꼽/VA+ㅁ/ETN]
[예시] 김철수 지음	[짓/VV+음/ETN]

② '음, 기'가 붙은 말이 단순히 명사형이냐 아니면 굳어진 명사이냐 하는 것은 물론 문맥에 따라 결정되어야 하지만 먼저 그것이 사전에 등재되어 있느냐의 여부를 살펴보아야 한다.

[예시] 책을 읽기가 어렵다.	[읽/VV+기/ETN+가/JKS]
읽기 교육이 문제가 된다.	[읽기/NNG]

③ 여러 개의 어미가 결합한 준말의 끝에 명사형 전성 어미가 나오는 다음과 같은 경우, 어미를 모두 묶어서 명사형 전성 어미로 표지를 부여한다.

[예시] 꼭 <u>그렇다기</u>보다는	[그렇/VA+다기/ETN+보다/JKB+는/JX]
그것이 <u>문제라기</u>에는	[문제/NNG+이/VCP+라기/ETN+에/JKB+는/JX]

마) 관형형전성어미(ETM)

용언의 성격을 임시로 바꾸어 다른 문장 속에서 관형사적인 역할을 하게 하는 어미이다.

-ㄴ/은	어제 먹은 빵에 이상이 있었다.	[먹/VV+은/ETM]
-는	잃어버린 물건을 찾는 일은 어렵다.	[찾/VV+는/ETM]
-던	이제까지 미루던 일을 오늘 해치웠다.	[미루/VV+던/ETM]
-ㄹ/을	나에게는 아직 처리할 일이 있다.	[처리/NNG+하/XSV+ㄹ/ETM]
-런	우리가 함께한 날이 어제런 듯하다.	[어제/NNG+이/VCP+런/ETM]

주의사항

① ㅂ 불규칙 용언 어간에 관형사형 전성 어미가 결합한 경우 '-은, -을'이 아닌 '-ㄴ, -ㄹ'로 분석한다. 이는 '-ㄴ, -ㄹ'을 포함하고 있는 '-ㄴ가', '-ㄹ까' 등에도 적용된다. 이러한 방식은 모든 불규칙 용언과 모든 매개모음 어미에 적용되는 것이 아니라, ㅂ 불규칙 용언 어간에 명사형 어미(-ㅁ)와 관형사형 어미(-ㄴ, -ㄹ)가 결합할 때 적용됨에 유의한다. ㅅ 불규칙 용언 어간에 결합하는 '-은, -을'은 '은/ETM, 을/ETM'으로 분석한다.

[예시] 그녀의 고운 얼굴	[곱/VA+ㄴ/ETM]
꽃밭은 매우 아름다울 것이다.	[아름답/VA+ㄹ/ETM]
얼마나 고울까?	[곱/VA+ㄹ까/EF+?/SF]
[참고] 얼굴이 고우니	[곱/VA+으니/EC]
[참고] 집을 지을 거야.	[짓/VV+을/ETM]

② 종결어미에 이어서 전성어미가 올 경우 통합해서 전성어미로 처리한다.

[예시] 어느 쪽에 더 비중을 두느냐는 것이	[두/VV+느냐는/ETM]

2) 접사(X)

주의사항

접사(접두사, 접미사)는 아래 가)~라)에 목록화된 접사가 등장한 경우에만 분리하여 분석한다. 접사의 분리 원칙은 다음과 같다.

2음절 단어의 처리

① <우리말샘> 등재어인 경우(즉 원어절에 나타난 2음절 단어와 같은 의미의 단어가 <우리말샘> 표제어로 올라 있는 경우), 표제어에 하이픈이 있는 경우에만 접사를 분리한다. 단, 접사를 분리하고 남는 요소가 어근에 해당하는 경우에는 접사를 분리하지 않는다.

[예시] 오형 (사전: 오-형, 혈액형의 하나)	[오/NNG+형/XSN]

<우리말샘>에 명사, 부사로 등재된 의미나 쓰임이 아니라, 그 앞뒤로 다른 말(단위성 의존명사 등)과 함께 쓰이는 '순서'의 '제일'은 '제/XPN+일/NR'로 분리한다.

[예시] 제일 차(회/조/항…) 회의	[제/XPN+일/NR] (○) [제일/NNG] (×), [제일/MAG] (×)

분석 대상이 되는 말이 <우리말샘>에 등재되어 있다고 하더라도 그 쓰임과 의미를 면밀히 확인할 수 있도록 주의한다.

[예시] 15일자 신문	[15/SN+일/NNB+자/NNG] (○) [15/SN+일자/NNG] (×)

② <우리말샘> 미등재어인 경우, 그 단어가 2음절 한자어이거나 접사 분리 시 어근이 남는다면 접사를 분리하지 않는다. 그 외의 경우에는 접사를 분리한다.

[예시] 뇌성마비(사전: 뇌성^마비)	[뇌성/NNG+마비/NNG]
→ <우리말샘>에서 '뇌성'은 단독 표제어로 올라 있지 않아 하이픈 유무를 참고할 수 없다. 하지만 2음절 한자어에 해당하는 말이 <우리말샘>에서 대체로 하이픈 없이 처리되고 있음을 참고하여 '-성'을 분리하지 않고 명사로 처리한다.	

[예시] 나는 아침형 인간이 아니라 밤형 인간이다. [밤/NNG+형/XSN]
→ '밤형'은 미등재어이지만 2음절 한자어가 아니다. 또한 '밤'이 명사이므로 '밤'과 '-형'을 분리한다.

③ 숫자, 로마자 등 기타기호에 접사가 결합한 것은 일반명사 지침의 (라)항을 우선 적용하여 처리한다.

[예시] 3분의 일 [3/SN+분/XSN]
→ 명사 '삼분'이 <우리말샘>에 하이픈 없이 등재되어 있지만, 기타기호의 처리 방법을 우선 적용하여 [3/SN+분/XSN]으로 분리하여 분석한다.

3음절 이상 단어의 처리

① 3음절 이상 복합어의 경우, <우리말샘>의 표제어 하이픈(-) 위치를 참고하여 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있는 경우에만 해당 접사를 분리한다.

[예시] 과보호(사전: 과-보호) [과/XPN+보호/NNG]
→ 하이픈 바로 앞에 놓인 접사인 '과-'가 분석 대상 접사이다. 이 경우 '과'와 '보호'를 분리한다.
[예시] 피보험자(사전: 피보험-자) [피보험자/NNG]
→ 하이픈 바로 뒤에 놓인 접사인 '-자'는 본 지침의 분석 대상 접사가 아니다. 따라서 더 이상 분리하지 않고 전체를 [피보험자/NNG]로 분석한다.

② 만약 접사를 분리했을 때 남는 단위가 어근(XR)이라면 접사 분리를 하지 않는다.

[예시] 비롯하다(사전: 비롯-하다) [비롯하/VV+다/EF]
→ 하이픈 바로 뒤에 놓인 접사인 '-하-'가 분석 대상 접사이지만, 이것을 분리하고 남는 단위인 '비롯'이 어근에 해당한다. 이 경우 '비롯'과 '하'를 분리하지 않는다.

③ 접사를 분리하고 남은 부분이 사전 미등재어인 경우가 있다. 그 미등재어가 홀로 쓰이지 않아 어근 자격을 갖는 것으로 판단된다면, 위 ②와 마찬가지로 접사를 분리하지 않는다.

접사를 분리하고 남은 미등재어가 합성어라면, 합성어 분석 원칙에 따라 합성어 구성 요소를 분리하여 분석한다.

[예시] 역세권(사전: 역세-권) [역세권/NNG]

→ 하이픈 바로 뒤에 놓인 접사인 '-권'이 분석 대상 접사이지만, 이것을 분리하고 남는 단위인 '역세'가 사전 미등재어이며 홀로 쓰이지도 않아 어근에 해당한다. 이 경우 '역세'와 '권'을 분리하지 않는다.

[예시] 중고생(사전: 중고-생) [중/NNG+고/NNG+생/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-생'이 분석 대상 접사인데, 이것을 분리하고 남는 단위인 '중고'가 사전 미등재어이다. 그런데 사전에 중학교를 뜻하는 '중', 고등학교를 뜻하는 '고'가 명사로 올라 있어 이 말은 합성어로 파악된다. 이 경우 접사를 분리하고 남은 미등재 합성어를, 여타 미등재 합성어의 처리 방식과 마찬가지로 분리하여 분석한다.

[참고] 일회용(사전: 일회-용) [일회/NNG+용/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-용'이 분석 대상 접사인데, 이것을 분리하고 남는 단위인 '일회'가 단독으로는 사전 등재어가 아니다. 하지만 '일회^결실성' 등의 구 표제어 속에서 한 단어로 나타나므로 더 분석하지 않고 한 단어로 취급한다.

④ 만약 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있어서 해당 접사를 분리해 냈는데, 접사를 뗀 나머지 부분에 또 분석 대상 접사가 포함되어 있을 수 있다. 그런 경우에는 그 나머지 단어를 <우리말샘>에서 검색하여 하이픈의 위치를 확인한 후, 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있다면 해당 접사를 다시금 분리해 낸다.

[예시] 비합리적(사전: 비합리-적) [비/XPN+합리/NNG+적/XSN]

→ 하이픈 바로 뒤에 놓인 접사인 '-적'이 분석 대상 접사이므로 '비합리'와 '적'을 분리한다. 그런데 '적'을 뗀 나머지 부분인 '비합리'(사전: 비-합리)에 분석 대상 접사인 '비'가 들어 있고, <우리말샘>에서 '비합리'를 검색했을 때 하이픈 바로 앞에 '비'가 놓여 있다. 이 경우 '비'와 '합리'를 다시금 분리한다. 결과적으로 '비합리적'을 [비/XPN+합리/NNG+적/XSN]으로 분석하게 된다.

⑤ 하이픈 바로 앞이나 바로 뒤에 분석 대상 접사가 있어서 해당 접사를 분리했을 때 남는 단위가 2음절 요소라면 위 '2음절 단어의 처리'에 따라 해당 2음절 요소를 처리한다.

⑥ 복합어가 <우리말샘> 미등재어여서 하이픈 정보를 참고할 수 없는 경우에는 <우리말샘> 등재 어휘를 참조하고 복합어의 의미 구조에 대해 직접 판단하여 처리한다.

[예시] 최대형 (미등재어) [최대/NNG+형/XSN]

→ 사전 등재어인 '최소형'(사전: 최소-형)을 참고하여 처리할 수 있다.

[예시] 대의원회 (미등재어) [대의원회/NNG]

→ '대의원의 모임'이라는 뜻이므로 의미 구조상 대의원-회로 나뉜다. 이때 하이픈 뒤의 '-회'는 본 지침의 분석 대상 접사가 아니므로 이 단어를 더 분리하지 않는다.

가) 체언접두사(XPN)

접두사는 명사와 수사에 결합하는 접사류를 묶어서 체언접두사만을 설정하기로 한다.

명사 접두사에는 한자어계 접두사와 고유어계 접두사가 있는데, 그 목록의 풍부함에 비해 대개가 생산성이 그리 높지 않다. 일단 여기서는 비교적 생산성이 높다고 인정되는 접두사에 대해 접두사 분석을 하기로 한다.

하나의 접두사가 여러 개의 다의를 갖는 경우가 있다. 아래에 제시한 접두사가 하나의 표제어 안에서 다의를 갖는 경우, 어떤 의미로 쓰인 것이든 관계없이 접두사를 분석해 낸다. ('친(親)-'이 "혈연관계로 맺어진"의 뜻으로 쓰였든 "부계 혈족 관계인"의 뜻으로 쓰였든 "그것에 찬성하는"의 뜻으로 쓰였든 모두 접두사 분석을 한다.)

가(假)	가건물	소(小)	소강당
고(高)	고물가	신(新)	신정당
과(過)	과보호	왕(王)	왕족발
구(舊)	구소련	재(再)	재충전
날	날음식	저(低)	저임금
노(老)	노부부	제(第)	제13차
대(大)	대선배	준(準)	준전시
맏	맏아들	초(超)	초만원
맨	맨몸	최(最)	최고급
무(無)	무의식	친(親)	친러시아
미(未)	미완성	탈(脫)	탈냉전
반(反)	반독재	폐(廢)	폐광산
범(汎)	범세계	풋	풋살구
부(不)	부도덕	피(被)	피고소인
불(不)	불합리	한	한가운데
비(非)	비논리	헛	헛고생
생(生)	생김치		

나) 명사파생접미사(XSN)

명사파생접미사는 명사나 다른 어근에 후행하여 그것이 명사의 기능을 수행할 수 있도록 만들어 주는 의존 형태이다. 그러나 명사파생접미사는 연구자에 따라 그 목록이 다르며, 실제로도 구분이 애매한 경우가 많다. 본 분석에서는 접미사의 생산성과 접미사를 제외한 형태의 독립성을 기준으로 다음과 같이 목록을 마련하였다.

하나의 접미사가 여러 개의 다의를 갖는 경우가 있다. 아래에 제시한 접미사가 하나의 표제어 안에서 다의를 갖는 경우, 어떤 의미로 쓰인 것이든 관계없이 접미사를 분석해 낸다.

가(哥)	김가	분지(分之)	삼분지 일
가(價)	매매가	뻘	조카뻘
가량	1시간가량, 다섯명가량	산(産)	중국산
간(間)	한 달간	상(上)	역사상
경(頃)	두 시경	생1(生)	갑자생
계(界)	교육계	생2(生)	견습생
계(系)	몽고계	성(性)	인간성
광(狂)	메모광	시(視)	영웅시
권(券)	만 원권	씩	만원씩
권(圈)	운동권	어치	만원어치
권(權)	참정권	여(餘)	삼십여
기(氣)	기름기	용(用)	전쟁용
께	10분께	율(率)	출산율
꼴	십 원꼴	장이	간판장이
꾼	노름꾼	쟁이	심술쟁이
끼리	전우끼리	적(的)	사상적
네	동이네	정(整)	일만 원정
님	선생님	제(制)	봉건제
당(當)	한 사람당	질	서방질
대(臺)	만 원대	짜리	백 원짜리
댁(宅)	청주댁	째1	이틀째
들	우리들	째2	옹기째

들이	1L 들이	쯤	내일쯤
론(論)	비평론	층(層)	선수층
류(類)	자연류	치(値)	기대치
률(率)	경쟁률	치레	인사치레
리(裡)	비밀리	투성이	먼지투성이
발(發)	서울발	풍(風)	복고풍
배기	열 살배기	하(下)	지배하
별(別)	가구별	형(型)	기본형
부(附)	12일부	형(形)	계란형
분(分)	3분의 일	화(化)	도구화

주의사항

① 명사파생접미사인 '-들'은 그 분포가 매우 다양하여 일부에서는 이를 보조사와 접미사로 나누어 분석하기도 한다. 그러나, 본 분석에서는 이들을 모두 명사파생접미사로 처리한다. '먹고들'의 '-들'도 선행성분이 어미이긴 하나, 일치하는 대상은 선행하는 명사로 해석할 수도 있기 때문이다.

[예시] 사람들이 우리 집에 왔다.	[사람/NNG+들/XSN]
그들은 밥을 먹고들 싶었다.	[먹/VV+고/EC+들/XSN]

② '-님'은 다음과 같이 세 가지의 분석 중의성을 가지므로 주의해서 분석한다.

'임'의 의미로 쓰인 경우: 보통명사

[예시] 님과 이별하다.	[님/NNG+과/JKB]

사람의 '이름'이나 '성' 뒤에서 쓰인 경우: 의존명사

[예시] 김철수님께서 오셨습니다.	[김철수/NNP+님/NNB+께서/JKS]

그 밖의 경우: 명사파생접미사

[예시] 과장님이 부르십니다.	[과장/NNG+님/XSN+이/JKS]

다) 동사파생접미사(XSV)

동사파생접미사는 어기 또는 어근에 붙어서 그것을 동사로 만들어 주는 기능을 갖는 접미사이다. 여기서는 그러한 접미사 중 현재 생산성을 가지고 쓰이는 것만을 인정하여 분석한다. 접사를 분석하고 난 나머지 언어 단위가 '어근(XR)'일 경우에는 더 이상 분리하여 분석하지 않고 통합한다.

당하	아군이 공격당하는 데에는 이유가 있다.	[공격/NNG+당하/XSV+는/ETM]
되	아침식사가 이미 준비되어 있었다.	[준비/NNG+되/XSV+어/EC]
시키	오늘 강아지를 운동시키려고 공원에 나갔다.	[운동/NNG+시키/XSV+려고/EC]
하	외국에서 공부하는 일이 쉬운 것은 아니다.	[공부/NNG+하/XSV+는/ETM]
받	몇몇은 집세 인상을 강요받았다.	[강요/NNG+받/XSV+았/EP+다/EF+./SF]

주의사항

'말씀드리다', '축하드리다' 등에서 볼 수 있는 '드리다'는 <우리말샘>에 동사 파생 접미사로 올라 있다. 하지만 본 지침에서 분리하여 분석하는 접미사 목록에는 들어 있지 않으므로, 앞에 오는 명사와 묶어서 '말씀드리/VV, 축하드리/VV' 등으로 분석해야 한다. '말씀, 축하, 인사, 감사, 사과' 등 동작을 나타내는 명사 뒤에서 '하다' 대신 공손의 의미를 더하며 쓰이는 '드리다'를 이처럼 앞말과 묶어서 처리함에 유의한다.

'출렁거리다', '출렁대다' 등에서 볼 수 있는 '거리다', '대다' 역시 <우리말샘>에 동사 파생 접미사로 올라 있으나 본 지침에서 분리하여 분석하는 접미사 목록에 들어 있지 않다. 따라서 앞에 오는 요소와 묶어서 '출렁거리/VV', '출렁대/VV' 등으로 분석한다.

라) 형용사파생접미사(XSA)

형용사파생접미사는 어기나 어근에 붙어서 그것을 형용사로 파생시키는 접미사이다. 여기서는 그러한 접미사 중 현재 생산성을 가지고 쓰이는 것만을 인정한다. 접사를 분석하고 난 나머지 언어 단위가 '어근(XR)'일 경우에는 더 이상 분리하여 분석하지 않고 통합한다.

답	사람이 사람답게 행동해야지.	[사람/NNG+답/XSA+게/EC]
되	거짓된 말은 들통나기 마련이다.	[거짓/NNG+되/XSA+ㄴ/ETM]
롭	어려운 일일수록 슬기롭게 대처하라.	[슬기/NNG+롭/XSA+게/EC]
스럽	그녀의 사랑스러운 표정을 보거라.	[사랑/NNG+스럽/XSA+ㄴ/ETM]
하	건강한 신체에 건강한 정신이 깃든다.	[건강/NNG+하/XSA+ㄴ/ETM]

주의사항

명사구에 결합하는 '만하'(예: 집채만 하다)는 '만'을 보조사로, '하'는 그 활용 양상을 참고하여 형용사로 분석한다. '만하'는 앞에 관형사형이 올 경우 '만/NNB+하/XSA'로 분석되는 경우도 있으므로 주의해야 한다.

[예시] 그 일을 처리하는 데 철수만한 인재가 없다.	[철수/NNP+만/JX+하/VA+ㄴ/ETM]
이 음식은 먹을 만하다.	[먹/VV+을/ETM] [만/NNB+하/XSA+다/EF+./SF]

3) 어근(XR)

국어에는 하나의 단어가 조사에 의해 분리되는 현상이 있다. 즉, 파생된 용언에서 보조사 등의 삽입에 의해 어근과 접사가 분리되는 현상이 있다. 어절 분석 표지에 어근(XR)이 포함되어 있으므로 분리된 어근에 어근 표지를 할당할 수 있다.

[예시] 따뜻도 하다	[따뜻/XR+도/JX] [하/VA+다/EF]

신문기사의 제목 등에서 어근으로 문장이 끝나는 경우에도 어근 표지를 할당할 수 있다.

[예시] 대회 3일 차 분위기 무난	[무난/XR]

주의사항

① 형용사의 어근 분리 시 어근 뒤에 오는 '하다'에는 형용사(VA) 표지를 부여한다. 실제 '하다'가 형용사라고 보기 어려운 측면이 있으나, 그 활용 양상을 참고하고 말뭉치에서의 활용

을 생각했을 때에는 형용사로 파악하는 것이 실익이 있다. 아래와 같은 경우도 '하다'를 형용사(VA)로 본다.

[예시] 영희는 키가 철수만 하다.	[철수/NNP+만/JX] [하/VA+다/EF+./SF]
비가 올 듯도 하다.	[듯/NNB+도/JX] [하/VA+다/EF+./SF]

② 용언을 파생하는 접미사(하다, 되다 등)를 분리하고 남은 단위가 어근(XR)일 경우에는 더 이상 분리하지 않고 통합형으로 분석한다.

[예시] 듬직하다	[듬직하/VA+다/EF]
취하다	[취하/VV+다/EF]

사 기타

1) 기호

가) 마침표

마침표(.)의 경우, 문장 끝에서 쓰인 것을 SF로 분석한다. 따옴표 안 인용문의 문장 끝에서 쓰인 마침표도 마찬가지이다.

[예시] 아이가 듬직하다.	[듬직하/VA+다/EF+./SF]
"아이가 착하다. 또 듬직하다."라고 말했다.	[듬직하/VA+다/EF+./SF+"/SS+라고/JKQ]

아래와 같이 장, 절 등의 항목을 구분하기 위해 숫자 사이에, 또는 숫자 끝에 마침표가 쓰이는 경우가 있다. 이때 숫자 중간에 있는 마침표는 SP로, 숫자 끝에 있는 마침표는 SF로 분석한다.

[예시] 1. 서론	[1/SN+./SF]
1.1. 연구 목적	[1/SN+./SP+1/SN+./SF]

소수점으로 쓰인 마침표, 날짜를 나타내는 숫자 사이에서 쓰인 마침표, 홈페이지 주소 속 마

침표 등 문장 종결의 의미가 없는 마침표는 SP로 분석한다.

[예시] 3.14	[3/SN+./SP+14/SN]
www.an.com	[www/SL+./SP+an/SL+./SP+com/SL]

말줄임표 대신 마침표가 쓰인 경우, 마침표의 개수에 관계없이 모두 묶어 SE로 처리한다.

[예시] 그리고....	[그리고/MAJ+..../SE]

단, 아래와 같이 말줄임표 대신 쉼표가 쓰인 경우에는 각각의 쉼표를 따로따로 SP로 처리한다. 말줄임표로서 마침표를 여러 개 찍는 것은 어문 규범에 부합하는 것인 데 반해, 쉼표를 여러 개 찍는 것은 어문 규범에서 인정하는 말줄임표가 아님을 고려한 것이다.

[예시] 안 기뻐요,,,	[기쁘/VA+어요/EF+,/SP+,/SP+,/SP]

나) 기타 기호(SW)

길이, 무게, 수효, 시간 따위의 수량을 수치로 나타내는 단위들 중 '미터, 그램, 리터' 등은 의존명사(NNB)로, 외국어로 된 'm, g, l' 등은 기호(SW)로 분석함에 유의한다. '제곱미터, 퍼센트포인트' 등 사전에 한 단어로 올라 있는 단위 명사를 기호로 나타낸 것도 아래와 같이 하나의 기호로 분석한다.

[예시] 5㎡	[5/SN+㎡/SW]
2%p	[2/SN+%p/SW]

한글이 원이나 괄호 속에 들어간 아래와 같은 기호는 SW로 처리한다.

[예시] ㈜/SW, ㈁/SW, ㉮/SW	
[참고] (주)	[(/SS+주/NNG+)/SS]
→ 이와 같이 괄호와 한글을 분리할 수 있는 경우에는 각각을 따로 분석한다.	

다) 한자(SH)

한자를 SH로 처리한다. 한자가 원이나 괄호 속에 들어간 기호도 SH로 처리한다.

[예시] ㊊/SH, ㊄/SH
[참고] (五) [(/SS+五/SH+)/SS]
→ 이와 같이 괄호와 한자를 분리할 수 있는 경우에는 각각을 따로 분석한다.

라) 외국어(SL)

한자(SH)를 제외한 외국 문자(로마자, 가나 등)를 SL로 처리한다. 로마자로 쓰인 숫자(로마숫자 I, II, III 등)도 로마자임에 주목하여 SL로 처리한다. 외국 문자가 원이나 괄호 속에 들어간 기호도 SL로 처리한다.

[예시] ⓐ/SL, Ⓐ/SL
[참고] (a) [(/SS+a/SL+)/SS]
→ 이와 같이 괄호와 외국 문자를 분리할 수 있는 경우에는 각각을 따로 분석한다.

아래와 같이 어떤 표현의 구체적인 내용을 숨기려는 의도로, 또는 구어 전사 시 말이 정확히 들리지 않아 로마자 X 표시를 사용하는 경우가 있다. 그런 경우 아래와 같이 처리한다.

[예시] 이런 버르장머리 없는 X [X/SL]
XXXX 이론 [XXXX/SL]
→ 이처럼 한 어절 전체가 X로 되어 있는 경우, 전체를 묶어 SL로 처리한다.
※ X 대신 O가 쓰일 때도 마찬가지이다.
※ 만약 로마자가 아닌 ×(곱셈표)나 △ 등이 쓰였으면 SW로 처리한다.
[예시] 어쩌라는 거야 씨X [씨X/NA]
XX스의 이론 [XX스/NF+의/JKG]
XXX의 이론 [XXX의/NA]
→ 이처럼 어절의 일부가 X로 되어 있는 경우, X를 포함하는 말의 품사를 고려하여 명사에 준하면 NF로, 용언에 준하면 NV로, 그 외에 해당하면 NA를 부여한다. 'XXX의'의 경우 XXX가 체언일 것으로 예상은 되지만 확실치 않으므로 전체 어절을 NA로 처리한다.

마) 숫자(SN)

아라비아숫자(0, 1, 2 등) 및 아라비아숫자가 원이나 괄호 속에 들어간 기호를 SN으로 처리한다.

[예시] ① 조리법	[①/SN]
⑴ 조리법	[⑴/SN]
[참고] (1) 조리법	[(/SS+1/SN+)/SS]
→ 이와 같이 괄호와 숫자를 분리할 수 있는 경우에는 각각을 따로 분석한다.	

바) 기호가 어절 중간에 개입한 경우

기호가 어절 중간에 개입한 경우, 기호를 뺀 말이 사전에 한 단어로 등재되어 있다면 기호가 있다 하더라도 전체를 통합하여 표지를 부여한다.

[예시] 농·수산물 (사전: 농수산-물)	[농·수산물/NNG]
초·중·고 (사전: 초중고)	[초·중·고/NNG]
의~리	[의~리/NNG]
사이~소 (사전: 어미 '-이소')	[사/VV+이~소/EF]

기호를 뺀 말이 사전에 한 단어로 등재되어 있지 않은 경우에도, 분리하여 분석할 경우 어근이 남는다면 전체를 통합하여 표지를 부여한다.

[예시] 당·정·청 (사전: 당정)	[당·정·청/NNG]

사전에 '당정'만이 등재되어 있어 이 어절을 '당·정/NNG'과 '·/SP', '청'으로 분리할 경우, 사전 미등재어이면서 홀로 쓰이지 않는 '청'이 남는다. 이 경우 '청'을 앞말에 통합하여 '당·정·청/NNG'로 표지를 부여한다.

단, 숫자나 외국어로만 표기된 경우에 기호가 포함되어 있으면 모두 각각 분석한다.

[예시] 6.25	[6/SN+./SP+25/SN]

아래와 같이 외국 문자나 숫자로 된 '주석'이 어절 중간에 개입하는 경우에는 각각의 요소를 분리하여 분석한다. 이 경우 표지를 줄 수 없는 불완전한 형태가 생길 수 있다.

[예시]마이크로소프트(microsoft)사	[마이크로소프트/NNP+(/SS+microsoft/SL+)/SS+사/NNG]

2) 준말

여러 개의 어미가 결합한 준말은 그 안에 분석 대상 선어말어미가 들어 있는 경우에 한해서만 복원한다.

[예시] 간다는	[가/VV+ㄴ다는/ETM] (○)
	[가/VV+ㄴ다/EF+하/VV+는/ETM] (×)
간댔어.	[가/VV+ㄴ다/EF+하/VV+았/EP+어/EF+./SF] (○)
	[가/VV+ㄴ댔어/EF+./SF] (×)

3) 분석불능범주

그 자체가 사전에 등재되어 있지도 않으면서, 축약의 정도가 심하거나 분석하기 어려운 방언형의 경우 분석불능범주로 처리한다. 분석이 어렵더라도 그 품사(범주)를 명확히 할 수 있는 경우에는, 추정 범주인 NF(명사 추정 범주), NV(동사 추정 범주)를 부여한다.

[예시] 담배가 쪼매턴게 하마 자라서 빼나?	[쪼매턴게/NA]

<우리말샘>에 접사로 등재되어 있으나 본 지침의 '분석 대상 접사'가 아닌 요소가 앞뒤의 기호 등 때문에 분리되어 홀로 남은 경우, 해당 요소를 분석불능범주로 처리한다. <우리말샘>에 등재되어 있지 않고 홀로 쓰이지도 않아 어근에 준하는 것으로 볼 수 있는 요소가 같은 이유로 홀로 남은 경우에도, 해당 요소를 분석불능범주로 처리한다.

[예시] 대(對)중국	[대/NA+(/SS+對/SH+)/SS+중국/NNP]
5인(人)승	[5/SN+인/NNG+(/SS+人/SH+)/SS+승/NA]

4) 합성어

<우리말샘>에 등재되어 있는 합성어를 한 단위로 둔다. 합성어가 북한어나 방언으로 등재되어 있어도 분석하려는 말과 의미가 동일하다면 표준어와 동일하게 처리한다.

주의사항

① <우리말샘>에 합성어로 올라 있는 단어는 한 단위로 분석한다.

[예시] 정치권력 (사전: 정치-권력)	[정치권력/NNG]

② 어절에 나타난 표기가 규범에 맞지 않아 사전에서 검색되지 않으나 규범에 맞게 표기된 단어는 사전 등재어일 때, 규범에 맞게 표기된 단어에 준하여 한 단어로 분석한다.

[예시] 먼저번 (사전: 먼젓번)	[먼저번/NNG]
조랭이떡 (사전: 조롱이-떡)	[조랭이떡/NNG]

③ 단어 자체가 사전의 표제어로 등록되어 있지는 않으나 사전에 구로 등재되어 있는 말(A^B)의 일부에 해당하는 단어일 때에도 한 단위로 분석한다.

[예시] 사대강을 (사전: 사대강^수계법)	[사대강/NNG+을/JKO]

④ 사전에 구로 등재되어 있는 말(A^B)은 세분하여 분석하는 것을 원칙으로 한다.

[예시] 학생운동 (사전표기: 학생^운동)	[학생/NNG+운동/NNG]

구를 이루는 둘 이상의 요소를 분리했을 때, 어느 한 요소에 분석 대상 접사가 포함되어 있는 경우가 있을 수 있다. 이 경우 물론 분석 대상 접사를 분리해야 한다.

[예시] 인적사항 (사전: 인적 사항)	[인/NNG+적/XSN+사항/NNG]

⑤ 아래와 같은 혼성어는 분리하여 분석하기 어려우므로 한 단어로 보고, 의미에 따라 일반명사 또는 고유명사로 분석한다.

[예시] 아베노믹스 (아베+이코노믹스) [아베노믹스/NNG]
→ 이는 아베의 경제정책을 일컫는 말로, 의미상 본 지침의 고유명사 부류에 들지 않는다. 이에 따라 전체 단어를 일반명사로 처리한다.
[예시] 홍드로 (홍수아+페드로) [홍드로/NNP]
→ 이는 특정 인물을 가리키는 말로 사용되고 있으므로 의미상 고유명사 부류에 든다. 이에 따라 전체 단어를 고유명사로 처리한다.

⑥ 합성어로 등록되어 있지 않은 표제어는 분리해서 분석하되, 사전 표제어로 등록되어 있는 최대한 많은 음절수의 단어를 생성하도록 나눈다. 즉 다음 예와 같은 경우 3음절 어휘가 생성되는 첫 번째 분석을 취한다.

[예시] 영상학과 [영상학/NNG+과/NNG] (3음절+1음절)
영상학과 [영상/NNG+학과/NNG] (2음절+2음절)

⑦ 3음절 어휘와 같이 어느 쪽으로 나뉘어도 음절수가 같고, 양쪽 분석이 모두 사전 표제어라면 뒤쪽을 먼저 분석한다.

[예시] 차창밖 [차/NNG+창밖/NNG]
이등품 [이/NR+등품/NNG]

⑧ 합성어로 등록되어 있지 않은 표제어를 더 작은 요소로 분리했을 때 어근이 남거나 품사를 부여하기 어려운 요소가 남는다면, 해당 요소를 분리하지 않고 앞말 또는 뒷말과 결합하여 형태 표지를 부여한다.

[예시] 당정청 [당정청/NNG]

사전에는 '당정청'이 올라 있지 않고 '당정'만이 올라 있다. 청와대를 뜻하는 '청'은 미등재어이고 홀로 쓰이는 일이 드물어 어근으로 판단할 수 있다. 이런 경우 '청'을 앞말인 '당정'과 결합하여 처리한다.

[예시] 오인승 [오인승/NNG]

'오인승'은 수사 '오'와 명사 '인', 그리고 사전 미등재어인 '승'으로 구성되어 있다. '승'은 미등재어이지만 홀로 쓰이지 않으므로 어근으로 판단할 수 있고, 이런 경우 '승'을 앞말인 '인'과 결합하여 처리한다. 그런데 그렇게 해서 도출된 '인승' 역시 미등재어이고, '인승'의 '인'이 일반명사임을 고려하면 '인승'도 일반명사가 되어야 할 것이나 이 말이 홀로 쓰이지

않기 때문에 일반명사로 품사를 부여하기가 어렵다. 따라서 '인승'도 분리하지 않고 앞말과 결합하여 형태 표지를 부여한다.

5) 파생어

<우리말샘>에 등재되어 있는 파생어를 한 단위로 둔다.

주의사항

① 사전에 파생어가 등재되어 있어도 그 안에 분석 대상 접사가 포함되어 있으면 분석한다. 접사의 분석 범위는 접사 지침의 주의사항에 따른다.

[예시] 수습생 (사전: 수습-생)	[수습/NNG+생/XSN]

② 사전에 구로 등재되어 있는 말 안에 분석 대상 접사가 포함되어 있는 경우 역시 접사를 분석한다. 단, 접사를 분리해 냈을 때 어근이 남는 경우에는 접사를 분리하지 않는다.

[예시] 도선수습생 (사전표기: 도선^수습생)	[도선/NNG+수습/NNG+생/XSN]

③ 분석 대상 접사를 분리한 후 남은 단위가 사전 미등재어인 경우가 있다. 해당 미등재어가 홀로 쓰이지도 않고 조사와도 결합하지 않는다고 판단된다면 그것을 어근으로 보아 접사를 분리하지 않는다. 단, 해당 미등재어가 어근이 중첩된 형식이라면 접사를 분리한다.

[예시] 가급적이면	[가급적/NNG+이/VCP+면/EC]
대대적 개편	[대대적/MMA]

'-적'이 분석 대상 접사인데 '가급'과 '대대'가 미등재어이다. '가급' 및 '대대'는 홀로 쓰이지도 않고 뒤에 조사가 결합할 수도 없는 단위이므로 미등재어이더라도 어근으로 보아 '가급적', '대대적'에서 '-적'을 분리하지 않는다. '가급적'은 사전에 명사와 부사로, '대대적'은 관형사와 명사로 등재되어 있으므로, 뒤에 조사가 후행하는 경우에는 명사로, 조사 없이 체언이 후행하는 경우에는 관형사로, 그렇지 않은 경우에는 부사로 맥락에 맞게 분석한다.

[예시] 나른나른한	[나른나른/MAG+하/XSA+ㄴ/ETM]

'-하-'가 분석 대상 접사인데 '나른나른'이 미등재어이다. '나른나른'은 사전에 등재된 어근

'나른'의 중첩형이다. 이 경우 '-하-'를 분리하고, '나른나른'에 일반 부사 표지를 부여한다.

④ 분석 대상 접사 목록에 없는 접사(비분석 접사)가 결합한 단어는, 그것이 사전 미등재어여도 한 단위로 둔다.

[예시] 임명자	[임명자/NNG]

'임명자'는 미등재어이고 '-자'는 접미사인데 분석 대상 접사는 아니다. 이 경우 '-자'를 '임명'에 결합하여 처리하지 않으면 달리 처리할 수 있는 방법이 없다. 따라서 '임명자'를 한 단위로 두고 일반명사 태그를 부여한다.

아 구어

구어 자료의 형태 분석 방법은 기본적으로 문어 자료의 형태 분석 방법과 동일하다. 다만 구어에서 나타나는 준말과 형태 변이 현상을 되도록 분석에 반영하고, 구어 전사 시 이용된 특별한 마크업과 표지를 처리하기 위해 아래의 지침을 별도로 마련하였다.

1) 구어에서 나타나는 준말과 형태 변이 현상의 처리

가) 하나의 요소 내부에서 형태 변이가 일어난 경우

아래와 같이 하나의 형태 표지가 붙는 단위에 구어의 음성적 특성이 반영되어 형태 변이가 일어났을 때는, **원어절의 형태를 바꾸지 않되 표준형에 비추어 형태 표지를 부여**한다.

[예시] **건**(<그건) 어렵지 않아요 [**거/NP**+ㄴ/JX]
것두(<그것도) 좋은데 [**것/NP**+두/JX]
늦을까 봐 **날라**서 왔어. [**날르/VV**+아서/EC]
이**걸로** [이거/NP+**ㄹ로/JKB**]
좋**으까**? [좋/VA+**으까/EF**+?/SF]
여기 앉**어** [앉/VV+**어/EF**]
그렇게 하**더래도** [하/VV+**더래도/EC**]
학교 **간대더라** [가/VV+**ㄴ대더라/EF**]
할런지 모르겠다 [하/VV+**ㄹ런지/EF**]
갈 것 같**애** [같/VA+**애/EF**]
→ '애'가 이처럼 표기상으로 분리되어 드러난 경우에만 '애'로 분석한다. '가기를 바래', '나만 나무래'에서처럼 '애'가 표기상으로 분리되어 드러나지 않은 경우에는 모음조화에 따라 '아'로 분석한다.
[예시] 가기를 바**래** [바라/VV+**아/EF**]

나) 본래 둘 이상으로 분석되어야 하는 요소인데 축약되어 형태 분리가 어려워진 경우

(1) 용언 어간과 어미의 결합형인 경우

사전에 한 단어로 올라 있는 말이 아니어서 용언 어간과 어미로 분석해야 하는 말이 다음과 같이 축약되어 전사된 경우가 있다. 이때는 아래와 같이 구어의 변이 형태를 그대로 인정하면서 어간과 어미를 분리하여 분석한다. 형태상 분리가 어려움에도 어간과 어미를 분리하도록 한 것은, 절을 꾸리는 데 있어서 용언의 역할이 중요한 만큼 용언 어간의 모습을 드러낼 필요가 있기 때문이다.

(가) 형용사 '이렇-', '그렇-', '저렇-', '어떻-'류의 변이 형태

[예시] 일케 [잃/VA+게/EC]
이케 [잊/VA+게/EC]
이르케 [이릏/VA+게/EC]
요러케 [요렇/VA+게/EC]

요렇케 [요렇/VA+게/EC]

→ 어미가 '케'로 잘못 전사되었지만 'ㅎ'과 '게'가 만나 '케'가 된 것으로 보아야 하므로 '게/EC'로 분석한다.

요케 [욯/VA+게/EC]

요로케 [요롷/VA+게/EC]

그르케 [그륳/VA+게/EC]

그런케 [그럲/VA+게/EC]

그러치 [그러치/IC], [그렇/VA+지/EF]

→ 사전에 '그렇지'가 "틀림없이 그렇다는 뜻으로 하는 말"로서 감탄사로 올라 있다. 이에 따라, 일어난 사태에 대한 만족을 표시하며 혼잣말로 쓰이는 '그러치, 그릏치, 그치'는 더 분석하지 않고 감탄사로 보아야 한다. 이때 '그렇지'로 전사되어야 할 것이 '그러치'로 전사되었다. 하지만 위에서 본 '요렇케'의 경우와 달리 내부 형태 분석을 하는 상황이 아니므로 원문의 표기를 그대로 따른다.

→ 그런데 구어 말뭉치에서 나타나는 '그렇지'는 대부분, 일어난 사태에 대한 만족을 표시하는 혼잣말로서 쓰이는 것이 아니라 상대방에게 동의를 표시하는 말로 쓰인다. 이 경우에는 어간 '그렇-'과 어미 '-지'로 분리하여 분석해야 한다. 즉 '그렇/VA+지/EF'로 분석해야 함에 유의한다.

그치 [그치/IC], [긇/VA+지/EF]

→ 이 역시, 만족을 표시하는 감탄사로 쓰인 경우에는 '그치/IC'로, 상대방에게 동의를 표시하는 말로 쓰인 경우에는 '긇/VA+지/EF'로 분석한다.

그치 않습니까? [긇/VA+지/EC]

그릏치 않습니까? [그릏/VA+지/EC]

→ 이때는 '그치', '그릏치'가 감탄사로 쓰인 것이 아니다. 따라서 위와 같이 용언 어간과 어미로 분석해야 한다.

→ '그릏치'는 '그릏지'로 전사되어야 할 것이 잘못 전사된 것이다. 잘못 전사된 부분에서 형태 분석이 이루어지므로, 위에서 본 '요렇케'의 경우와 마찬가지로 '치/EC'가 아니라 '지/EC'로 분석한다.

그찮아 [긇/VA+잖아/EF]

어뜧케 [어뜧/VA+게/EC]

→ '어뜧게'로 전사되어야 할 것이 잘못 전사된 것이다. 잘못 전사된 부분에서 형태 분석이 이루어지므로, 위에서 본 '요렇케, 그릏치'의 경우와 마찬가지로 '케/EC'가 아니라 '게/EC'로 분석한다.

어뜨케 [어뜩하/VV+아/EF]

→ '어떡해'의 변이형이다. 사전에 '어떡하다'가 등재되어 있어 '어떡해'를 '어떡하/VV+아/EF'로 분석하는 것을 참고하여 '어뜩하/VV+아/EF'로 분석한다.

→ 물론, '어뜨케 됐어?'에서처럼 부사어로 쓰인 것은 '어뜧/VA+게/EC'가 될 것이다.

(나) 그 외 용언의 변이 형태

[예시] 따르케 [따릏/VA+게/EC]

다르케 [다릏/VA+게/EC]

요만하케 [요만핳/VA+게/EC]

→ '다르다', '요만하다'의 변이 형태가 나타났다. 역시 변이 형태를 그대로 인정하여 용언 어간과 어미를 분리한다. '따르+케', '요만하+케'로 분석될 수도 있을 것이나 가능한 한 용언 어간 쪽에서 변이 형태를 인정하기로 한다.

(다) 두 어절 이상에 해당하는 용언 어간+어미의 경우

[예시] 어케요 [엏/VA+게/EC+하/VV+아요/EF]

→ 두 어절에 해당하는 '어떻게 해요'가 줄어들었고, 그 속에 용언 어간과 어미가 있다. 용언 어간 '하'가 생략되었는데, 이때 '하'를 복원하지 않으면 어절 속에 동사의 어간이 없는 셈이 되므로 '하'를 복원해야 한다. '하' 앞의 '어케'는 '엏/VA+게/EC'로 분석한다.

이케서 [잏/VA+게/EC+하/VV+아서/EC]

→ 역시 두 어절에 해당하는 '이렇게 해서'가 줄어들었다. 용언 어간 '하'를 복원하지 않으면 어절 속에 동사의 어간이 없는 셈이 되므로 복원해야 한다. '하' 앞의 '이케'는 '잏/VA+게/EC'로 분석한다.

왜케 [왜케/MAG]

왤케 [왤케/MAG]

→ 두 어절에 해당하는 '왜 이렇게'가 줄어들었다. '왜케', '왤케'가 절에서 서술어로 쓰이는 일은 없으므로, 이 경우에는 예외적으로 더 분석하지 않고 '왜케, 왤케'를 일반부사로 처리한다.

(라) '이케~'와 같이 원문에 물결표 표시가 있는 것은 물결표를 제외하고 IC로 분석해야 함에 유의한다.

(마) 'X하-' 형태에서 '하'가 아예 생략되거나 'ㅎ'만 남은 아래와 같은 경우에는 '하'의 형태를 완전하게 복원한다. '하'를 복원하지 않으면 어절 속에 용언의 어간이 없는 셈이 되므로 복

원하지 않을 수 없다.

[예시] 논의토록 [논의/NNG+하/XSV+도록/EC]
생각지 못한 [생각/NNG+하/XSV+지/EC]

(2) 용언 어간과 어미의 결합형이 아닌 경우

용언 어간과 어미의 결합형이 아니라면, 아래와 같이 형태 분리가 어려운 구어의 축약형을 더 분석하지 않고 하나의 단어로 인정하는 방안을 취하기로 한다. 형태 분리가 어려운 경우란, 형태 분리를 했을 때 적어도 한 요소가 사전 미등재어이고 그 요소가 다른 환경에서는 나타나지 않는 경우를 말한다. 형태 표지는 해당 단어의 문장 성분을 고려하여 부여한다(예: 부사어→부사).

[예시] 내비둬 [내비두/VV+어/EF]
→ '두-'는 분석 가능하지만 '내비'가 사전 미등재어이다. 그리고 '내비'는 '두-' 앞 외의 다른 환경에서는 거의 나타나지 않는다. 이에 따라 '내비두-' 전체를 동사로 처리한다.
냅둬 [냅두/VV+어/EF]
여따(<여기에다가) 놔. [여따/MAG]
→ '여'는 등재되어 있지만 '따'가 미등재어이다. 이 '따'는 '여따, 거따, 저따' 외에서는 보기 어렵다. 이에 따라 '여따' 전체를 한 단어로 처리한다. 문장 속에서 부사어로 쓰이므로 일반부사로 처리한다.
언놈이(<어느 놈이) 그래? [언놈/NP+이/JKS]
→ '놈'은 분석 가능하지만 '언'이 사전 미등재어이다. 그리고 '언'이 '놈' 앞 외의 다른 환경에서는 나타나지 않는다.
얻다 대고 [얻다/MAG]
→ '-다'는 분석 가능하지만 '얻'이 사전 미등재어이다. 그리고 '얻'이 '-다' 앞 외의 다른 환경에서는 나타나지 않는다.
클났다. [클나/VV+았/EP+다/EF+./SF]
→ '클'이 사전 미등재어이고 '나다' 외의 다른 환경에서 나타나지 않는다.
어서(<어디서) 그래? [어서/MAG]
→ '어'가 사전 미등재어이고 '서' 외의 다른 환경에서 나타나지 않는다.
짱난다 [짱나/VV+ㄴ다/EF]
→ '짱'이 사전 미등재어이고 '나다' 외의 환경에서 나타나지 않는다.

주의사항

① '이리로, 그리로, 저리로, 요리로, 고리로, 조리로'뿐 아니라 '일로, 글로, 절로, 욜로, 골로, 졸로'가 <우리말샘>에 부사로 등재되어 있으므로 MAG로 분석해야 함에 유의한다.

② 아래와 같이 같은 모음이 겹치면서 축약된 경우에는 본래 형태를 복원한다.

[예시] 어딨어요.	[어디/NP+있/VA+어요/EF+./SF]

다) 비표준적인 준말 활용형

아래와 같이 비표준적인 준말 활용형이 나타난 경우, 용언 어간은 표준형으로 복원하되 어미에서 매개모음 '으'를 빼고 분석한다.

[예시] 여기다 논(<놓은) 거야.	[놓/VV+ㄴ/ETM]
여기다 노셨던(<놓으셨던) 거야.	[놓/VV+시/EP+었/EP+던/ETM]
찌시더니(<찧으시더니)	[찧/VV+시/EP+더니/EC]
아이를 나면은(<낳으면은)	[낳/VV+면/EC+은/JX]

라) 삼중모음

모음 'ㅟ'와 'ㅕ'가 축약되어 삼중모음 발음이 나타난 경우, '사귀'어요'와 같이 'ㅝ'로 표시되어 있다. 형태 분석 시에 이 'ㅝ'는 반영하지 않는다.

[예시] 바뀌'었었어요.	[바뀌/VV+었었/EP+어요/EF+./SF]

마) 관형격조사 '에'

관형격조사 '의'의 발음을 '에'로 전사한 경우가 있다. 이 경우 '에/JKG'로 형태 표지를 붙인다. '에'가 부사격조사인지 관형격조사인지 판단이 어려운 경우에는 부사격조사(JKB)로 형태 표지를 붙인다.

[예시] 나**에** 생각	[나/NP+**에/JKG**]

바) 지정사 '이다'

구어에서 이중모음이 단모음으로 발음되는 현상이 자주 일어나 그 결과 '예'가 '에'로 발음되고, 아래와 같이 지정사 '이다'가 생략된 것으로 보이는 현상이 있다. 이때는 문법적으로 지정사가 있으나 단지 이중모음이 단모음으로 발음된 것으로 보아 지정사를 복원한다.

[예시] 이렇게 얘기할 거에요. [거/NNB+**이/VCP**+에요/EF+./SF]

사) 구어에서 나타나는 사전 미등재 요소

사전에 등재되지 않은 문법 요소 '-ㄹ랑', '-ㄹ동'은 사전에 등재된 연결어미 '-ㄹ락'을 참고하여 연결어미로 분석한다.

[예시] 이해가 갈랑말랑 하길래 [가/VV+**ㄹ랑/EC**+말/VV+**ㄹ랑/EC**]

의성의태어를 구성하는 요소가 여러 번 반복되어 나오는 경우의 처리는 아래와 같다.

[예시] 지글지글지글지글 [지글지글/MAG+지글지글/MAG]
→ 사전에는 '지글지글'이 한 단어로 올라 있고, '지글'은 어근에 해당한다. 위의 경우 단어 '지글지글'이 두 번 연달아 나온 것으로 분석할 수 있으므로 두 단어로 나누어 처리한다.
지글지글지글 [지글지글지글/MAG]
→ '지글지글'을 한 단어로 처리하면 어근에 해당하는 '지글'이 남는다. 이런 경우에는 어근을 앞말에 붙여서 '지글지글지글' 전체를 하나의 단어로 처리한다.
지글지글지글지글지글 [지글지글/MAG+지글지글지글/MAG]
→ 위의 경우에는 뒤쪽에 더 많은 음절수가 남도록 지글지글/MAG+지글지글지글/MAG로 분석한다.

2) 구어 전사 시 이용된 마크업과 표지의 처리

가) 물결표(~)

머뭇거림을 나타내는 담화표지에 ~(물결표)가 붙어 있다. 이 경우 ~는 분석에서 제외하고 ~ 앞에 있는 말에 IC를 부여한다.

[예시] 아~	[아/IC]
그~	[그/IC]
뭐~	[뭐/IC]

단, 아래와 같이 머뭇거림을 나타내는 담화표지가 아닌 것에 ~(물결표)가 붙어 있는 경우가 있다. 그런 경우는 물결표를 넣지 않아야 할 곳에 넣은 전사 오류에 해당하므로, 물결표를 분석에서 제외하고, 남은 요소에 형태 표지를 부여한다.

[예시] 국호를~을~	[국호/NNG+를/JKO+을/JKO]

나) 마크업 기호

<trunc>, </trunc> 등의 마크업 기호는 한 어절로 두고 형태 표지를 부여하지 않는다. 단, 아래의 주의사항에 유의한다.

주의사항

① <note> </note> 마크업의 경우에는 마크업 기호와 그 안의 내용을 모두 한 줄에 보여주고, 어떠한 표지도 부여하지 않는다.

[예시] <note>배경 화면 잠깐 나옴</note> → 한 어절로 두고 분석하지 않음.

② 사람 이름, 주소 등 개인 정보 보호를 위한 마크업은 다음과 같은 방식으로 형태 표지를 부여한다.

[예시] <anon type="**name**" n="**1**"/>가	[**name1/NNP**+가/JKS]
<anon type="**name**"/>가	[**name/NNP**+가/JKS]
<anon type="**address**" n="**2**"/>은	[**address2/NNP**+은/JX]

구어 전사 지침상, 일반 대화에서 대화자들 및 관련인의 개인 정보가 드러난 경우에는 개인 정보 보호를 위하여 해당 정보를 위와 같이 마크업으로 가리도록 하였다. 그런데 전사 실수로 이름 등의 개인 정보가 마크업 없이 노출된 경우가 있다. 이때에는 해당 개인 정보에 NAP 표지를 부여하고, 향후 보완 방법을 모색할 수 있도록 한다.

NAP 표지는 일반 대화 자료의 형태 분석에서만 적용하며, 공적 방송 자료의 형태 분석

에서는 적용하지 않는다. 또한 일반 대화 자료에서도 정치인, 연예인 등 유명인의 이름에는 적용하지 않는다.

[예시] 지현이가 그러는데 [**지현이/NAP**+가/JKS]

③ 마크업으로 인해 사람 이름에서 분리된 접미사 '-이'에는 NA를 부여한다.

[예시] <anon type="name" n="5"/>**이** 말고 하나가 더 있니?
→ '말고' 앞에는 주격조사가 올 수 없다. 이때 '말고' 앞에 나온 '이'는 '영숙이'에서 볼 수 있는 접사 '이'로 판단 가능하다. 본 지침의 비분석 접사가 분리되어 나온 경우이므로 이러한 '이'는 '이/NA'로 처리한다.

다) <trunc> </trunc> 사이의 요소

<trunc> </trunc> 사이에 표시되어 있는 끊어진 어절(단어가 불완전하게 발화된 경우)에는 NA(분석불능범주) 표지를 부여한다.

[예시] 미국과 <trunc>같</trunc> 같은
→ <trunc>
같 [같/NA]
</trunc>

단, <trunc> </trunc> 사이에 있는 요소를 제외할 경우 앞뒤의 말이 이어지지 않는 경우라면, <trunc> </trunc> 사이에 있는 요소이더라도 형태 표지를 부여한다.

[예시] 또 <trunc>문의하< / trunc> 하기도 했습니다.
→ <trunc>
문의하 [문의/NNG+하/XSV]
</trunc>
하기도 [하/XSV+기/ETN+도/JX]

라) <unclear> </unclear> 사이의 요소

전사 시 잘 들리지 않은 부분은 <unclear> </unclear>로 표시되어 있다. 가능한 한 각 요소에 맞는 형태 표지를 부여하고, 형태 표지를 부여할 수 없는 경우에는 NA(분석불능범주),

NV(용언추정범주), NF(명사추정범주)를 부여한다.

(1) 정확히 들리지 않았으나 x 표시 없이 전사된 경우

가능한 한 각 요소에 맞는 형태 표지를 부여한다.

[예시] <unclear>더 힘들어</unclear>
→ <unclear>
더 [더/MAG]
힘들어 [힘들/VA+어/EF]
</unclear>
[예시] 있<unclear>어요</unclear>
→ 있 [있/VA]
<unclear>
어요 [어요/EF]
</unclear>
[예시] 있어<unclear>요</unclear>
→ 있어 [있/VA+어/EF]
<unclear>
요 [요/JX]
</unclear>

(2) 일부 음절이 들리지 않은 경우

일부 음절이 들리지 않은 경우에는 해당 음절이 x로 표시되어 있다. 이때는 x가 포함된 단어 부분에 NA(분석불능범주), NF(명사추정범주), NV(용언추정범주)를 부여한다.

[예시] <unclear>xx스의</unclear> 이론을
→ <unclear>
xx스의 [xx스/NF+의/JKG]
</unclear>
이론을 [이론/NNG+을/JKO]

(3) <unclear> </unclear> 마크업으로 인해 표지를 주기 어려운 음절이 발생하는 경우

아래와 같이 <unclear> 마크업으로 인해 단어가 분리되어 표지를 주기 어려운 음절이 발생하는 경우에는, 각 음절에 NA(분석불능범주)를 부여한다.

[예시] 임시정 <unclear>부</unclear>
→ 임시정 [임시/NNG+정/NA]
<unclear>
부 [부/NA]
</unclear>

3) 전사 오류 및 해석 불능 어절의 처리

가) 탈자로 인해 형태 표지 부여가 어려운 경우

아래와 같이 전사 과정에서 탈자가 발생하였거나 혹은 발화 실수로 과도한 생략이 일어나 형태 표지 부여가 어려워지는 경우가 있다. 이런 경우에는 해당 요소에 NA(분석 불능 범주), NV(용언 추정 범주), NF(명사 추정 범주) 중 하나를 부여한다.

[예시] 하지 못는(<못하는) [못/NV+는/ETM]

나) 잉여적인 요소가 덧붙은 경우

아래와 같이 전사 과정에서 첨자가 발생하였거나 혹은 발화 실수로 잉여적인 형태가 덧붙은 경우가 있다. 이런 경우에는 해당 요소에 최대한 그 요소에 맞는 형태 표지를 부여하고, 만약 형태 표지 부여가 어렵다면 NA(분석 불능 범주)를 부여한다.

[예시] 국호를을 [국호/NNG+를/JKO+을/JKO]
됐습니다다. [되/VV+었/EP+습니다/EF+다/EF+./SF]

다) 띄어쓰기 오류로 인해 형태 표지 부여가 어려운 경우

'그럴걸'이 '그럴 걸'로, '뒤치락거리다'가 '뒤치락 거리다'로 띄어쓰기와 함께 전사된 경우가 있다. 이때 띄어쓰기 오류로 인해 '걸'의 처리, '거리-'의 처리가 어려워진다. '거리-'같이 용언의 성격을 띠는 요소에 대해서는 최대한 형태 표지를 부여한다. 그 외의 경우에도 최대한 형태 표

지를 부여하지만, 형태 표지 부여가 어려운 요소에는 NA(분석불능범주), NV(용언추정범주), NF(명사추정범주)를 부여한다.

[예시] 뒤치락 거리고 [뒤치락/XR, 거리/VV+고/EC]
→ '거리-'는 본 지침에서 분석하지 않는 동사 파생 접미사이다. 용언의 성격을 띠는 요소에는 최대한 형태 표지를 부여하여 동사로 처리한다.

[예시] 그럴 걸 [그러/VV+ㄹ/ETM, 걸/NA]
→ 어미 '-ㄹ걸'이 분리되어 나왔는데, 앞의 '-ㄹ'에는 관형형 어미 표지를 줄 수 있지만 뒤의 '걸'은 처리가 어려우므로 분석 불능 표지를 준다.

[예시] 집에 갔는 지 모르겠다 [가/VV+았/EP+는/ETM, 지/NA]
→ 어미 '-는지'가 분리되어 전사되었는데, 앞의 '-는'에는 관형형 어미 표지를 줄 수 있지만 뒤의 '지'는 처리가 어려우므로 분석 불능 표지를 준다.

라) 표기법 오류로 인해 형태 표지 부여가 어려운 경우

아래와 같이 더 분석되어야 할 대상이 있음에도 표기법 오류 때문에 형태 분리 및 형태 표지 부여가 어려워지는 경우가 있다. 이 경우에는 올바른 표기법으로 수정한 형식을 상정하고 표지를 부여한다.

[예시] 공부를 하며는 [하/VV+면/EC+은/JX]
→ '면은'으로 써야 할 것을 '며는'으로 잘못 전사하였으며, 그 때문에 형태 분리가 어렵게 되었다. 이런 경우에는 올바른 표기형인 '면은'으로 복원하여 '면/EC+은/JX'로 형태 표지를 부여한다.

[예시] 너 때매 [땜/NNB+에/JKB]
→ '땜에'로 써야 할 것을 '때매'로 잘못 전사하였으며, 그 때문에 형태 분리가 어렵게 되었다. 이런 경우에는 올바른 표기형인 '땜에'로 복원하여 '땜/NNB+에/JKB'로 형태 표지를 부여한다.

[예시] 편찬되서 [편찬/NNG+되/XSV+어서/EC]
→ '돼서'로 써야 할 것을 '되서'로 잘못 전사하였다. 하지만 '되'와 '돼'는 발음이 동일하므로 단순한 표기법 차이 때문에 '어서/EC' 대신 '서/EC'로 형태 표지를 부여하는 것은 합리적이지 않다. 따라서 이 경우에도 올바른 표기형인 '편찬돼서'를 상정하여 '편찬/NNG+되/XSV+어서/EC'로 형태 표지를 부여한다.

아래와 같이 표기법 오류가 나타났으나 그 때문에 형태 분리 및 형태 표지 부여가 어려워지는 상황이 아니라면, 원문의 형식을 그대로 두고 표지를 부여한다.

[예시] 크리마트를 했어요. [크리마트/NNG+를/JKO]

→ '크림아트'로 써야 할 것을 '크리마트'로 잘못 전사했다. 하지만 그 때문에 형태 표지 부여가 어려운 상황은 아니다('아트'가 미등재어이기 때문에 외국어 지침에 따라 '크림아트' 전체를 일반명사로 분석해야 하는 상황임). 이런 경우에는 원문의 표기를 그대로 두고 '크리마트/NNG'로 형태 표지를 부여한다.

[예시] 그렇다고 봅니다만은 [보/VV+ㅂ니다만은/EC]

→ '봅니다마는'으로 써야 할 것을 '봅니다만은'으로 잘못 전사했다. 하지만 그 때문에 형태 표지 부여가 어려운 상황은 아니다('-다마는'이 하나의 어미로 등재되어 있고, 그 앞에 선어말어미에 준하는 '습니'가 결합한 것으로 보아 '-습니다마는'을 하나의 어미로 분석해야 하는 상황임). 이런 경우에는 원문의 표기를 그대로 두고 'ㅂ니다만은/EC'로 형태 표지를 부여한다.

※ '숙제를 하긴 했다마는(했다만).'에서처럼 '-다마는'이 종결부에서 쓰일 때에도 뒤에 생략된 말이 있는 것으로 보고 연결어미 표지를 부여한다.

[예시] 얼마나 슬펐는 줄 아네. [알/VV+네/EF+./SF]

→ '내'로 써야 할 것을 '네'로 잘못 전사했다. 하지만 그 때문에 형태 표지 부여가 어려운 상황은 아니다('내'는 '냐 해'의 준말로서 이처럼 '하'가 축약된 구성의 경우 그 속에 분석 대상 선어말어미가 들어 있지 않은 한 더 분리하지 않는다는 지침을 적용하여 '내/EF'로 분석해야 하는 상황임). 이런 경우에는 원문의 표기를 그대로 두고 '네/EF'로 형태 표지를 부여한다.

마) 어절의 의미 파악이 어려운 경우

어절의 의미를 파악하기 어려운 경우, 의미는 불분명하더라도 아래와 같이 해당 어절을 이루는 요소의 문법적 지위를 확정할 수 있다면 그에 따라 최대한 형태 표지를 부여한다. 문법적 지위를 확정하기 어렵거나 형태 표지를 부여하기 어려운 경우에 한해 NA를 부여한다.

[예시] 조선 웅 어~ 왕조 실의 역대 왕들의 왕릉 [실/NA+의/JKG]

→ 이때 '실의'의 의미를 정확히 파악하기가 어려우나, 맥락상 '조선 왕실의'에서 '실의'가 분리된 것으로 보인다. 이에 따라 실/NA+의/JKG로 형태 표지를 부여한다.

[예시] 아직까지 충청에 민심이 복마진입니다. [복마진/NNG+이/VCP+ㅂ니다/EF+./SF]

→ 이때 '복마진'의 의미를 정확히 파악하기가 어려우나 지정사 앞에 나타나는 문법적 특성으로 보아 일반명사로 판단할 수 있다. 이에 따라 복마진/NNG로 표지를 부여한다.

→ 위의 두 방식을 적용하여도 문법적 지위를 확정하기 어려운 요소가 있다면 해당 요소에 NA를

부여한다.

바) 한 어절 내에 마크업이 포함된 경우

다음과 같이 한 어절 내에 마크업이 포함된 경우에는, 마크업을 제외하고 남은 부분만을 지침에 따라 분석한다.

[예시] <trunc>아쉬< / trunc>아쉬움이	[아쉬/NA+아쉬움/NNG+이/JKS]

자 메신저 대화

메신저 대화 자료의 형태 분석 방법은 기본적으로 문어, 구어 자료의 형태 분석 방법과 동일하다. 다만 메신저 대화는 전형적인 문어나 전사된 구어와 구별되는 언어 특성 및 표기 특성을 보여 주기도 하므로, 메신저 대화에서 나타나는 특별한 언어 현상을 처리하기 위해, 또 원시 말뭉치에 포함된 특별한 표지를 처리하기 위해 아래의 지침을 별도로 마련하였다.

1) 원시 말뭉치에 포함된 표지 및 기호의 처리

가) 줄 바꿈 표지의 처리

하나의 메시지 안에 줄 바꿈이 포함된 경우, 줄 바꿈이 '\n'으로 표시되어 있다. 이 표지에는 NA(분석불능범주) 태그를 부여한다.

[예시] 별로지만\n　　[별로/MAG+이/VCP+지만/EC+**\n/NA**]

나) 개인정보를 치환한 표지의 처리

메신저 대화에 포함된 개인의 이름, 계정, 각종 번호, 주소, 소속 등의 개인정보는 메신저 대화 원시 말뭉치에서 name(이름), account(계정), telnum, cardnum, num(각종 번호), address(주소), affiliation(소속), others(기타)로 치환되어 있다. 이 중 **name, account, address, affiliation은 NNP(고유명사)**로, **telnum, cardnum, num은 SN(숫자)**으로, **others는 NNG(일반명사)**로 처리한다.

[예시] name1님　　[name1/NNP+님/NNB]
→ '님'은 사람의 성이나 이름 다음에 쓰이는 경우 의존명사(NNB)로 처리되어야 한다. 이에 따라 'name' 뒤의 '님'을 NNB로 처리함에 유의한다.
김name3　　[김name3/NNP]
name2이는　　[name2이/NNP+는/JX]
→ '성+name', 'name+이(받침 있는 사람 이름에 붙는 접미사)'는 묶어서 NNP를 부여한다.
[예시] 제 카트라이더 아이디랑 비슷하네요 account [account/NNP]
[예시] 전화번호는 telnum입니다.　　[telnum/SN+이/VCP+ㅂ니다/EF+./SF]

[예시] address으로 이사와. [address/NNP+으로/JKB]
[예시] 저는 affiliation에서 근무해요. [affiliation/NNP+에서/JKB]
[예시] 아직은 여성 others는 제주에서 저 혼자^^ [others/NNG+는/JX]

메신저 대화에 포함된 개인정보는 위와 같이 name 등으로 치환되어야 하나, 원시 말뭉치에서 미처 치환되지 않은 개인정보가 발견되는 경우가 있다. 그런 경우에는 NAP 표지를 부여하여 향후 보완 방법을 모색할 수 있도록 한다. 개인의 이름이 초성만으로 표시된 경우에도 NAP 표지를 부여한다.

[예시] 태화니랑 나랑 [태화니/NAP+랑/JC]
ㅇㅎ이가 와서 [ㅇㅎ이/NAP+가/JKS]

다) 이미지로 된 이모티콘의 처리

'😂'와 같이 이미지로 된 이모티콘에는 SW(기타 기호) 태그를 부여한다.

라) 문자나 기호를 사용한 이모티콘의 처리

아래와 같이 문자나 기호를 사용하여 표정이나 동작을 묘사한 이모티콘은 전체를 묶어 SW(기타 기호) 태그를 부여한다.

[예시] ㅠㅠ/SW, ㅜㅜㅜㅜㅜ/SW
→ 'ㅠ'와 'ㅜ'가 세 번 이상 입력된 경우에도 모두 묶어서 SW로 처리한다.
ㅡㅡ/SW
^^/SW
^^//SW
^^*/SW
^ㅠ^/SW
?.?/SW
ㅇㅅㅇ/SW
[예시] 응; [응/IC+;/SW]
→ ';'은 본래 SP 태그를 부여하는 기호이지만, 위와 같이 땀을 흘리는 모습을 묘사하기 위해 사용된 경우에는 '문자나 기호를 사용하여 표정이나 동작을 묘사한 이모티콘'의 일종으로 보고 SW 태그를 부여한다. 땀 흘리는 모습을 묘사하기 위해 ';'이 두 번 이상 입력

된 경우, 모두 묶어서 SW로 처리한다.

마) 이모티콘 외 기호의 처리

표정이나 동작을 묘사하지 않는 아래와 같은 기호에는 각 기호에 맞는 태그를 부여한다.

[예시] 그래♡ [그래/IC+♡/SW]
최고지~~ [최고/NNG+이/VCP+지/EF+~/SO+~/SO]
→ 구어 말뭉치에서는 머뭇거림 표시로 사용된 '~'(물결표)를 분석에서 제외한 바 있다. 하지만 메신저 대화 말뭉치의 '~'는 분석에서 제외하지 않고 SO 태그를 부여함에 유의한다.
죽는거임/ [죽/VV+는/ETM+거/NNB+이/VCP+ㅁ/ETN+//SP]

2) 메신저 대화에서 자주 나타나는 언어 현상의 처리

가) 다양한 의성의태어와 감탄사

메신저 대화에는 웃음소리, 울음소리 등의 각종 소리, 그리고 모양을 묘사하는 의성의태어가 다양한 형태로 나타나며, 느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말 등 감탄사도 다양한 형태로 나타난다.

'호호', '하하', '흑흑', '토닥토닥', '덜덜' 등 **소리나 모양을 묘사하는 의성의태어**는 <우리말샘>에 부사로 등재되어 있으므로, **이에 준하는 미등재어도 부사로 분석한다.**

또한 '아하', '응', '그래', '아니', '젠장', '빌어먹을', '안녕' 등 **느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말**은 <우리말샘>에 감탄사로 등재되어 있으므로, **이에 준하는 미등재어도 감탄사로 분석한다.**

[예시] 하하하/MAG, 흑흑/MAG, 흐규흐규/MAG, 아흑/MAG, 꺼어어억/MAG, 핰/MAG
[예시] 토닥토닥/MAG, 쥬륵/MAG
[예시] 오홍/IC, 핫/IC
으아앙 맛있겠다 [으아앙/IC]
왁! 부러워요 [왁/IC+!/SF]

나) 자음이 첨가된 형태

메신저 대화 원시 말뭉치에는 '~해용', '~해욤'처럼 주로 어미에 'ㅇ', 'ㅁ'과 같은 자음을 첨가한 형식이 많이 나타난다. 그러한 자음은 앞 형태와 묶어 형태 표지를 부여한다.

[예시] 최고징	[최고/NNG+이/VCP+**징/EF**]
학술대회얌	[학술/NNG+대회/NNG+이/VCP+**얌/EF**]
아파연(아파요)	[아프/VA+**아연/EF**]
그러셈	[그러/VV+**셈/EF**]

다) 어미 없이 용언 어간이 단독으로 쓰인 경우

아래와 같이 용언 어간이 어미 없이 단독으로 사용되는 경우가 있다. 그러한 경우에는 어미 없이 용언의 형태 표지만을 부여한다.

[예시] 고맙.	[고맙/VA+./SF]

때로는 용언 어간의 일부만으로 문장이 종결되는 경우가 있다. 그러한 요소는 어근으로 처리한다.

[예시] 속상..	[속상/XR+../SE]
부끄	[부끄/XR]

라) 사전 미등재 어미

메신저 대화 원시 말뭉치에는 '그러셈', '뭐하삼', '아니긔' 등에서 볼 수 있는 사전 미등재 어미가 종종 나타난다. 이때의 '-셈', '-삼', '-긔' 등을 종결어미로 처리한다.

[예시] 아니긔	[아니/VCN+긔/EF]

마) 사전에 등재된 요소가 다른 용법으로 사용되는 경우

본 지침은 분석된 각각의 요소에 대해 <우리말샘>의 품사를 따라 품사를 부여하는 것을 원칙으로 한다. 하지만 아래와 같이 '관형사형 어미 뒤에 명사 파생 접미사가 나오는 경우', '호칭에 해당하는 감탄사가 대명사로 쓰이는 경우', '명사 파생 접두사 또는 어근이 술부를 꾸미며 부사로 쓰이는 경우'에는 사전의 품사를 그대로 부여하기 어렵다. 따라서 이 세 가지 경우 및

이 외에 본 지침에서 명확히 제시한 경우에 국한하여 <우리말샘>과 다른 품사를 부여한다.

[예시] 청소기만 되는 용으로 샀어요. [용/NNG+으로/JKB]
→ '용(用)'은 <우리말샘>에 접사로 올라 있는데, 위 예문에서는 접사로 처리하기 어려우므로 일반명사로 처리한다.
[예시] 여보가 잘 골라 봐. [여보/NP+가/JKS]
→ '여보'는 <우리말샘>에 감탄사로 올라 있는데, 위 예문에서는 감탄사 '여보'가 따옴의 효과를 보이며 사용된 것이라고 하기 어렵고 2인칭 대명사로서의 용법을 보이고 있다. 이에 대명사로 처리한다.
[예시] 급 가을에서 겨울이 되었어요. [급/MAG]
→ '급'은 <우리말샘>에 접사와 어근으로 올라 있는데, 위 예문에서는 술부를 수식하고 있어 접사나 어근으로 처리하기 어렵다. 예문의 용법에 맞게 일반부사로 처리한다.

바) 미등재어

준말, 혼성어, 약어 등 <우리말샘> 미등재어가 출현한 경우, 아래의 조건에 부합하면 해당 미등재어를 한 단어로 처리한다.

① 더 작은 요소로 분리하면 <우리말샘> 미등재어가 도출되는 경우

[예시] 혼영/NNG(혼자 영화), 혼치킨/NNG(혼자 치킨), 당빠/NNG, 딥빡/NNG, 흠좀무/NNG, 검핑/NNG(검정핑크),
펭하/IC(펭수 하이)

② 본말이 <우리말샘>에서 한 단어로 등재되어 있는 경우

[예시] 잼있다 [잼있/VA+다/EF]
→ <우리말샘>에 '재미있다, 재밌다'가 한 단어로 등재되어 있으므로, 이에 준하여 '잼있다'도 한 단어로 처리한다.

③ 두 요소의 의미 합으로 투명하게 설명되지 않는 합성어

[예시] 곱창떡볶이/NNG
→ 곱창과 떡볶이를 섞어 만든 음식으로, 단순히 '곱창과 떡볶이'가 아니므로 한 단어로 처리하기로 한다.
[예시] 엔빵/NNG

→ ‘엔분의 일’의 ‘엔’과 내기를 뜻하는 의존명사 ‘빵’이 결합한 것으로 볼 가능성이 있으나, 두 요소의 의미 합으로 투명하게 설명되지 않는 의미를 가지므로 한 단어로 처리한다.

[예시] 꿀과목/NNG, 칼약속/NNG

→ ‘꿀’과 ‘칼’의 의미가 본래의 의미로부터 변하여 합성어를 만들어 내는 경우이므로 한 단어로 처리한다.

④ ‘노’, ‘갓’, ‘충’, ‘개’, ‘핵’, ‘스/쓰’ 따위의 접사에 준하는 요소가 결합한 단어

[예시] 노야근/NNG, 갓겜/NNG, 파밍충/NNG

[예시] 궁금쓰/XR, 동생스/NNG, 대박쓰/NNG

→ 어근이나 명사 뒤에 ‘스/쓰’가 결합한 형태가 자주 보인다. 이때의 ‘스/쓰’는 접사에 준하는 요소이되 본 지침에서 분리하여 처리하지 않는 접사이므로, ‘스/쓰’를 앞말에 붙여 처리한다. 이때 어근 뒤에 ‘스/쓰’가 결합한 말은 그 품사도 어근으로 처리하고, 명사 뒤에 ‘스/쓰’가 결합한 말은 그 품사도 명사로 처리한다.

→ 단, ‘반갑쓰’처럼 용언 어간 뒤에 ‘스/쓰’가 결합한 경우에는 ‘반갑쓰’를 형용사로 처리하기 어렵다는 문제가 생긴다. ‘반갑쓰’의 문법적 성질에 비추어 볼 때 형용사로 볼 수 없기 때문이다. 따라서 이처럼 용언 어간 뒤에 ‘스/쓰’가 결합한 경우에는 ‘스/쓰’를 분석 불능 범주(NA)로 처리하여 ‘반갑/VA+쓰/NA’와 같이 분석한다.

[예시] 개꿀/NNG, 개빡/NNG, 개빡치/VV+다/EF, 개좋/VA+다/EF

[예시] 핵고통/NNG, 핵춥/VA+다/EF

→ ‘개’와 ‘핵’은 체언과 용언 모두에 결합하여 쓰인다. ‘개/핵’+체언, ‘개/핵’+용언 모두 한 단어로 처리한다.

[예시] 개 헬 [개/MAG]

[예시] 핵빵터짐 [핵/MAG+빵/MAG+터지/VV+ㅁ/ETN]

→ 단, ‘개’와 ‘핵’이 ‘개 헬’과 같이 뒷말과 공백으로 분리되어 있거나 ‘빵터지다(빵/MAG+터지/VV+다/EF)’처럼 한 단어가 아닌 말 앞에서 쓰인 경우에는, ‘개’, ‘핵’을 부사로 처리한다. ‘개’와 ‘핵’은 체언과 용언 모두와 결합하므로, 체언 앞에서 어절이 분리된 ‘개, 핵’은 관형사로, 용언 앞에서 어절이 분리된 ‘개, 핵’은 부사로 처리하는 방법도 있을 것이나, 부사가 “속성” 의미를 지닌 체언을 수식하는 경우가 있음을 고려하여 일괄 부사로 처리하기로 한다.

→ 결과적으로 ‘개’, ‘핵’은 한 단어 자격을 갖는 체언, 용언과 붙어서 쓰인 경우에는 접사처럼 취급되고, 그 외의 경우에는 부사로 취급되는 셈이다. 이는 ‘개’, ‘핵’이 접사에 준하는 요소가 되어 가고 있으나 접사로 지위를 완전히 고정하기는 어려우며 단어와 접사의 중간적인 성격을 지니고 있음을 반영한 것이다.

⑤ 문자 모양의 유사성에 기반하여 변형된 단어

[예시] 윾래/NNG
→ '유래'를 문자 모양이 유사한 '윾래'로 변형하여 쓴 단어이다. '유래'가 일반명사이므로 '윾래'도 일반명사로 처리한다.
[예시] 머한민국/NNP
→ '대한민국'을 문자 모양이 유사한 '머한민국'으로 변형하여 쓴 단어이다. '대한민국'이 고유명사이므로 '머한민국'도 고유명사로 처리한다.

사) 외국어

외국어로 된 단어의 분리 및 형태 표지 부여 방식은 기본적으로 문어 지침 16쪽 다)항의 외국어 처리 방식과 동일하다. 외래적 요소는 한국어에서 명사 자격을 갖는 것이 일반적이므로, 문어 지침에서 제시된 바와 같이 <우리말샘>에 등재되지 않은 외국어 단어의 품사는 의미에 따라 NNG 또는 NNP 둘 중 하나로 처리한다. 또한 외국어의 '한 문장'이 한글로 전사되어 나타난 경우에는 각 어절을 내부 분석 없이 NA로 처리한다.

다만, 메신저 대화에는 '하이', '바이'와 같이 관습적인 인사말로서 쓰이는 외국어가 자주 나타난다. 또한 '노', '예스', '예압'과 같이 대답하는 말로서 쓰이는 외국어가 종종 나타난다. 이들을 '외국어의 한 문장이 한글로 전사되어 나타난 경우'로 보아 NA로 처리할 수도 있지만, 본 지침에서는 이런 현상을 메신저 대화의 한 특성으로 인정하고 **인사말, 대답하는 말로서 메신저 대화에서 관습적으로 사용되는 외국어 단어에 감탄사(IC) 표지를 부여하기로 한다.**

[예시] 빠이요(bye) [빠이/IC+요/JX]
하이(hi) [하이/IC]
노(no) [노/IC]
예쓰(yes) [예쓰/IC]
→ 이처럼 외국어가 관습적인 인사말, 대답하는 말의 기능을 하며 독립어로 사용된 경우에는 감탄사 표지를 부여한다.
[예시] 굳!(good) [굳/NA+!/SF]
해피뉴이어 [해피뉴이어/NA]
→ <우리말샘> 미등재어로서 외국어의 한 문장이 한글로 전사되어 나타났으며 관습적으로 쓰이는 인사말, 대답하는 말도 아니므로 문어 지침에서 제시된 대로 NA로 처리한다.

아) 방언

문어 및 구어에서의 방언 처리 방식과 동일하게, 방언의 형태를 그대로 보존하는 방식으로 분석한다. 몇 가지 자주 나타나는 방언형의 분석 예시를 보이면 아래와 같다.

[예시] 키우나벼 [키우/VV+나/EF+**비/VX**+어/EF]
어딨는겨? [어디/NP+있/VA+는/ETM+**기/NNB**+이/VCP+여/EF]
올거가? [오/VV+ㄹ/ETM+거/NNB+이/VCP+**가/EF**]
잘생겼나베 [잘생기/VV+었/EP+나/EF+**보/VX+이/EF**]
놀래가(놀라 가지고) [놀래/VV+어/EC+**가/VX+아/EC**]
[예시] 신랑땜시 [신랑/NNG+땜시/NNB]
→ '땜시'가 '때문에'의 방언형으로 등재되어 있는데 품사 정보는 제시되어 있지 않다. 이때 '땜시'는 더 작은 요소로 분리하여 분석하기 어렵고 그 자체로 부사성 의존명사와 유사한 성격을 보이므로 의존명사로 처리한다.
[예시] 뭐라카노 [뭐/NP+이/VCP+라/EF+**카/VV**+노/EF]
→ '미안하다 안 캤나' 같은 예를 고려하면 '카/VV'를 설정할 수 있다.
[예시] 어대요 [**어대/NP**+이/VCP+오/EF]
→ 사용 맥락을 보면 하오체가 쓰인 방언형으로 볼 수 있다. 따라서 '요'를 해요체 보조사가 아닌, 지정사와 하오체 종결어미 '-오'의 결합으로 처리한다.

3) 메신저 대화에서 나타나는 특수한 표기법의 처리

가) 초성만으로 표기한 단어

메신저 대화에는 단어의 초성만을 표기하는 경우가 자주 나타난다. 단어를 초성만으로 표기한 경우, 본래 단어의 품사에 따라 태그를 부여한다.

(1) 소리(웃음소리, 울음소리 등), 모양을 묘사한 의성의태어

'호호', '하하', '흑흑', '토닥토닥', '덜덜' 등 소리나 모양을 묘사하는 의성의태어는 <우리말샘>에 부사로 등재되어 있으므로, 이러한 부류의 단어를 초성만으로 표기한 경우에도 MAG(일반부사) 태그를 부여한다.

다만 초성 단어를 영문 자판으로 입력한 경우(예: 'ㅋㅋ'를 'zz'로 입력)에는 해당 단어에 SL

(외국어) 태그를 부여한다.

[예시] ㅋㅋ/MAG, ㅎㅎ/MAG, ㅋㅋㅌㅌㅌㅌㅋ/MAG, ㄷㄷㄷ/MAG

[예시] 목표이긴한뎅ㅎㅎ [목표/NNG+이/VCP+기/ETN+ㄴ/JX+하/VX+ㄴ데/EC+ㅎㅎㅎ/MAG]

→ 이와 같이 초성 단어('ㅎㅎㅎ')의 첫 글자가 앞말의 받침으로 표기된 경우, 초성 단어와 앞말을 분리하여 태그를 부여한다.

[참고] zzz/SL

(2) 느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말

'아하', '응', '그래', '아니', '젠장', '빌어먹을', '안녕' 등 느낌을 나타내는 말, 대답하는 말, 욕하는 말, 인사말은 <우리말샘>에 감탄사로 등재되어 있으므로, 이러한 부류의 단어를 초성만으로 표기한 경우에도 IC(감탄사) 태그를 부여한다.

[예시] ㅇㅎ/IC (아하, 오호), ㅇㅇ/IC (응), ㅇㄴ/IC (아니), ㄴㄴ/IC (노노), ㅎㅇ/IC (하이), ㅂㄱ/IC (방가), ㅂㅂ/IC (바이바이)

[예시] ㅁㅊ/IC (미친)

→ 수식하는 체언 없이 단독으로 욕하는 말로 쓰인 경우이다. '미친'은 감탄사로 등재되어 있지 않으므로 동사 '미치다'의 활용형으로 분석되어야 하지만, 'ㅁㅊ'으로 표기된 것을 어간과 어미로 분리하기는 어렵다. 분석이 어려우므로 NA를 줄 수도 있을 것이나, 욕하는 말로 쓰여 감탄사에 준하는 기능을 한다는 점을 고려하여 IC 표지를 부여한다.

[참고] ㅈ같은 상황 [ㅈ같/VA+은/ETM]

→ 일종의 욕하는 말이라 할 수 있으나, 이때는 수식하는 체언과 함께 쓰였고 어간과 어미의 분리도 가능하므로 'ㅈ같/VA+은/ETM'으로 분석한다.

[예시] ㅁㅈ/IC (맞아)

→ 상대의 말에 맞장구치는 말로서 쓰인 경우이다. '맞아'는 감탄사로 등재되어 있지 않으므로 동사 '맞다'의 활용형으로 분석되어야 하지만, 'ㅁㅈ'으로 표기된 것을 어간과 어미로 분리하기는 어렵다. 분석이 어려우므로 NA를 줄 수도 있을 것이나, 대답하는 말로 쓰여 감탄사에 준하는 기능을 한다는 점을 고려하여 IC 표지를 부여한다.

(3) 기타

이 외에도 아래와 같이 초성 단어가 사용되는 경우가 있다. 모두 본래 단어의 품사에 따라 태그를 부여한다.

[예시] ㄹㅇ(레알) 재밌어 [ㄹㅇ/MAG]

→ <우리말샘>에 '레알'이 부사로 등재되어 있음을 참고하여 MAG(일반부사) 태그를 부여한다.

[예시] ㅇㅈ(인정) [ㅇㅈ/NNG]

[예시] 두부피만 있어도 ㄱㅊ(괜찮) [ㄱㅊ/VA]

→ '괜찮다'의 어간만으로 문장을 종결한 경우이므로 어미가 없더라도 VA(형용사) 태그를 부여한다.

[예시] 이거 ㄱㅊ함(괜춘함) [ㄱㅊ/NNG]

→ '괜찮다'는 '괜춘하다'로 변형되어 쓰이기도 하는데, 이때 '하다' 앞의 '괜춘'은 단독으로도 쓰여 명사의 자격을 갖는 것으로 볼 수 있으므로 NNG(일반명사) 태그를 부여한다.

용언의 활용형이 초성으로 표기된 아래와 같은 경우에는 어간과 어미를 분리하여 분석하기 어려우므로 활용형 전체에 NA 태그를 부여한다. 아래 예는 기능상 감탄사로 처리하기 어렵다는 점에서 위 (2)에서 든 'ㅁㅊ', 'ㅁㅈ'의 경우와 차이가 난다.

[예시] ㅁㅎ(뭐해) [ㅁㅎ/NA]

나) 음절을 첨가하여 장음을 표시한 경우

아래와 같이 음절을 첨가하여 장음을 표시한 경우에는, 첨가된 음절을 장음화된 형태와 묶은 후 형태 표지를 부여한다.

문어 지침 34쪽에서는 '그러어엄'처럼 한 어절이 비정상적으로 늘어난 경우 NA로 처리하도록 하였으나, 본 지침에서는 이러한 장음 표기법을 메신저 대화 표기법의 한 특성으로 인정하고, 장음화된 형태에 맞는 형태 표지를 부여하기로 한다.

[예시] 믓지겠어요오오 [믓지/VA+겠/EP+**어요오오/EF**]

→ 종결어미 '-어요'의 장음을 표시하기 위해 '오오'라는 음절을 첨가하였다. 장음 표시인 '오오'를 장음 관련 형태인 '어요'와 묶어서 '어요오오/EF'로 분석한다.

비켜어어어어어 [비키/VV+**어어어어어어/EF**]

→ '켜'에 포함된 '어'까지 포함하여 '어'를 6회 입력해야 함에 유의한다.

[예시] 네에~ [네에/IC+~/SO]

오오오오 [오오오오/IC]

아아아아아 [아아아아아/IC]

→ 이처럼 동일 모음이 반복되는 경우 장음을 표시하는 것으로 볼 수 있으므로 모두 묶어서

감탄사 표지를 부여한다.

[예시] 수제버거어 [수제/NNG+**버거어/NNG**]

[예시] 매에워어(원형: 매워) [**매엡/VA**+**어어/EF**]

다) 표음주의 표기법을 적용한 경우

(1) 연음 현상이 표기에 반영된 경우

메신저 대화 원시 말뭉치에는 앞말의 끝 자음이 다음 음절의 초성으로 발음되는 연음 현상이 반영된 비표준적 표기형이 많이 포함되어 있다. 가령 '맞아'를 '마자'로, '맛있는'을 '마싰는'으로 적는 것이 그 예가 된다. 그런데 '맛있다'와 '마싰다'는 동일한 언어 기호를 형태주의 표기법으로 적을 것인지 표음주의 표기법으로 적을 것인지에서 차이를 보인 것일 뿐 언어 기호 자체가 다른 것은 아니다. 이를 고려하여 연음 현상이 반영된 표기형을 '맞+아', '맛있+는'과 같이 분석하고 형태 표지를 부여하기로 한다.

[예시] 마자(맞아) [**맞/VV**+아/EF]

마쟈(맞아) [**맞/VV**+야/EF]

힘드렁(힘들엉) [**힘들/VA**+엉/EF]

쟁여놔써요(쟁여놨어요) [쟁이/VV+어/EC+놓/VV+**았/EP**+어요/EF]

마싰는(맛있는) [**맛있/VA**+는/ETM]

구어 지침 76쪽의 라)항에서는 전사자의 실수로 인한 표기법 오류가 나타났으나 그 때문에 형태 분리 및 형태 표지 부여가 어려워지는 경우가 아니라면, 전사된 형식을 그대로 두고 형태 표지를 부여하도록 한 바 있다. 예를 들어 구어 전사 시 [마신는]이라고 발음된 것을 '맛있는'으로 전사해야 하지만 실수로 '마싰는'으로 전사한 경우, 표기법 오류가 나타났지만 그 때문에 형태 표지 부여가 어려워지는 경우는 아니므로 '마싰/VA+는/ETM'으로 처리하도록 한 것이다. 구어 전사에서 이런 오류는 예외적으로 일부의 경우에 나타나는 것이므로 특별한 처리 지침을 마련하지 않고 원문의 표기 형태를 분석에도 그대로 반영하도록 한 것이다. 하지만 메신저 대화의 경우 의도적으로 표음주의 표기법을 취하여 언어 기호를 소리 나는 대로 적는 경우가 매우 빈번히 나타나므로, 그 특수성을 고려하여 표음주의 표기법이 반영된 '마싰는'을 형태주의 표기법이 반영된 '맛있는'과 동일하게 '맛있/VA+는/ETM'으로 분석하기로 하였다.

(2) 기본형이 아닌 이형태가 표기에 반영된 경우

메신저 대화 원시 말뭉치에는 기본형을 밝혀 적지 않고 발음 나는 대로의 이형태를 표기한 비표준적 표기형도 많이 포함되어 있다. 가령 '싫어'를 '시러'로 적는 것이 그 예가 된다. 이때에도 '싫어'와 '시러'는 동일한 언어 기호를 형태주의 표기법으로 적을 것인지 표음주의 표기법으로 적을 것인지에서 차이를 보인 것일 뿐, 언어 기호 자체가 다른 것은 아니다. 이를 고려하여 표음주의 표기법이 적용된 경우, 형태주의 표기법이 적용된 경우와 동일한 방식으로 형태 표지를 부여하기로 한다.

다만 분석의 일관성과 다)항의 과도한 적용을 방지하기 위해서는 다)항의 적용 범위를 제한할 필요가 있다. **표음주의 표기법이 적용된 표기에 대해 형태주의 표기법과 동일하게 형태 표지를 부여하는 것은, 한국어의 필수적인 음운 규칙이 적용된 형태로(경음화 제외) 표기되어 있는 경우에 국한하기로 한다.**

한국어의 필수적인 음운 규칙에는 아래와 같은 것이 있다. **아래와 같은 음운 규칙이 표기형에 반영되어 있을 경우에는, 음운 규칙 적용 전의 기본형을 밝혀 분석한다.**

① 종성에서의 평폐쇄음화

종성에 놓이는 /ㅍ/가 /ㅂ/로, /ㅌ, ㅅ, ㅆ, ㅈ, ㅊ, ㅎ/가 /ㄷ/로, /ㅋ, ㄲ/가 /ㄱ/로 대치되는 현상을 말한다.(예: 덮다→덥따, 있다→읻따, 묶다→묵따) 종성에서의 평폐쇄음화 현상이 표기에 반영된 경우, 평폐쇄음화 이전의 기본형을 밝혀 분석한다.

단, '있'을 '잇'으로 적는 등 'ㅆ'을 'ㅅ'으로 적은 것은 종성에서의 평폐쇄음화 현상과 무관하다. 따라서 '있/VA'으로 기본형을 밝혀 분석하지 않고 표기형 그대로 '잇/VA'으로 분석한다.

[예시] 이불덥고(덮고) [덮/VV+고/EC]
→ 용언 어간이 종성에서의 평폐쇄음화 현상이 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영된 표기이므로 본래의 형태를 밝혀 '덮/VV'로 분석한다.

[참고] 드실 수 잇는(있는) [잇/VA+는/ETM]
→ '있[읻]'을 '읻'으로 적지 않고 '잇'으로 적은 것은 평폐쇄음화가 적용된 것으로 볼 수 없다. 이 경우는 한국어의 필수적인 음운 규칙이 표기에 반영된 경우가 아니므로 표기된 'ㅅ' 형태를 그대로 반영하여 '잇/VA'로 분석한다.

낫엇잖아(원형: 났었잖아) [나/VV+앗엇/EP+잖아/EF]
→ '었'을 '엇'으로 적는 등 'ㅆ'을 'ㅅ'으로 적은 표기가 많이 나타난다. 역시 한국어의 필수적인 음운 규칙이 표기에 반영된 경우가 아니므로 표기된 'ㅅ' 형태를 그대로 반영하여 '앗엇/EP'로 분석한다.

터진거박에(거밖에) [거/NNB+박에/JX]
→ '밖' 뒤에 모음이 오는 경우이므로 평폐쇄음화의 적용 환경이 아니다. 이런 환경에서 받

침이 'ㄱ'으로 적힌 것은 한국어의 필수적인 음운 규칙과 무관한 것이므로 표기된 'ㄱ' 형태를 그대로 반영하여 '박에/JX'로 분석한다.

야채 복았는데(볶았는데) [복/VV+았/EP+는데/EC]

→ '볶' 뒤에 모음이 오는 경우이므로 평폐쇄음화의 적용 환경이 아니다. 이런 환경에서 받침이 'ㄱ'으로 적힌 것은 한국어의 필수적인 음운 규칙과 무관한 것이므로 표기된 'ㄱ' 형태를 그대로 반영하여 '복/VV'으로 분석한다.

마싯는(맛있는) [맛잇/VA+는/ETM]

→ '맛있[마싣]'을 '마싣'으로 적지 않고 '마싯'으로 적은 것은 평폐쇄음화가 적용된 것으로 볼 수 없다. 이 경우는 한국어의 필수적인 음운 규칙이 표기에 반영된 경우가 아니므로 표기된 'ㅅ' 형태를 그대로 반영하여 '맛잇/VA'로 분석한다.

마싯써(맛있어) [맛잇ㅆ/VA+어/EF]

→ '맛있' 뒤에 모음이 오는 경우이므로 평폐쇄음화의 적용 환경이 아니다. 받침에 잉여적으로 'ㅅ'이 추가되었으며 이는 한국어의 필수적인 음운 규칙과 무관하므로 표기된 'ㅅ' 형태를 그대로 반영하여 '맛잇ㅆ/VA'로 분석한다.

② 비음화

비음이 아닌 자음이 비음의 앞 또는 뒤에서 비음으로 대치되는 현상을 말한다.(예: 먹는→멍는) 비음화 현상이 표기에 반영된 경우, 비음화 이전의 기본형을 밝혀 분석한다.

[예시] 수업 끈나써(끝났어) [끝나/VV+았/EP+어/EF]

→ 용언 어간이 비음화 현상이 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영된 표기이므로 본래의 형태를 밝혀 '끝나/VV'로 분석한다.

[예시] 궁물(국물) [국물/NNG]

→ 체언이 비음화 현상이 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영된 표기이므로 본래의 형태를 밝혀 '국물/NNG'로 분석한다.

③ 유음화

/ㄹ/와 /ㄴ/가 연쇄될 때, 또는 /ㄴ/와 /ㄹ/가 연쇄될 때 /ㄴ/가 /ㄹ/로 대치되는 현상을 말한다.(예: 칼날→칼랄) 유음화 현상이 표기에 반영된 경우, 유음화 이전의 기본형을 밝혀 분석한다.

[예시] 칼랄(칼날) [칼날/NNG]

→ 체언이 유음화 현상이 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영

된 표기이므로 본래의 형태를 밝혀 '칼날/NNG'로 분석한다.

④ 자음군 단순화

음절 종성에 두 개의 자음이 놓일 때 그 중 하나가 탈락하는 현상을 말한다.(예: 없고→업꼬) 자음군 단순화 현상이 표기에 반영된 경우, 자음군 단순화 이전의 기본형을 밝혀 분석한다.

[예시] 할수업는(할 수 없는) [하/VV+ㄹ/ETM+수/NNB+없/VA+는/ETM]
어이가업네(없네) [없/VA+네/EF]
힘듬(힘듦) [힘들/VA+ㅁ/ETN]
→ 용언 어간이 자음군 단순화가 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영된 표기이므로 본래의 형태를 밝혀 '없/VA', '힘들/VA'로 분석한다.

⑤ 용언 어간 말 /ㅎ/ 탈락

/ㅎ/로 끝나는 용언 어간 뒤에 모음으로 시작하는 어미가 올 경우 /ㅎ/가 탈락하는 현상을 말한다.(예: 좋아→조아) 용언 어간 말 /ㅎ/ 탈락 현상이 표기에 반영된 경우, /ㅎ/ 탈락 이전의 기본형을 밝혀 분석한다.

[예시] 시러(싫어) [싫/VA+어/EF]
너무마늠(너무 많음) [너무/MAG+많/VA+음/ETN]
조아(좋아) [좋/VA+아/EF]
→ 용언 어간이 어간 말 /ㅎ/ 탈락이 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영된 표기이므로 본래의 형태를 밝혀 '싫/VA', '많/VA', '좋/VA'로 분석한다.

⑥ 격음화

/ㅎ/와 /ㅂ, ㄷ, ㄱ, ㅈ/가 만나 /ㅍ, ㅌ, ㅋ, ㅊ/로 축약되는 현상을 말한다.(예: 놓고→노코) 격음화 현상이 표기에 반영된 경우, 격음화 이전의 기본형을 밝혀 분석한다.

[예시] 그러치(그렇지) [그렇/VA+지/EF]
→ 형용사 '그렇다'의 활용형이 격음화가 적용된 형태로 표기되었다. 한국어의 필수적인 음운 규칙이 반영된 표기이므로 본래의 형태를 밝혀 '그렇/VA'로 분석한다.
[참고] 왜케, 웰케 [왜케/MAG], [웰케/MAG]
→ 단, '왜 이렇게'가 줄어들면서 격음화가 표기에 반영된 '왜케', '웰케'의 경우, '웽/VA+게

/EC', '왫/VA+게/EC'로 분석할 수도 있겠으나 '왜 이렇-'이 한 단어가 아니므로 '왫/VA', '왫/VA'을 형용사로 처리하는 것이 부담스럽다. 또 '왫게', '왫게'와 같은 낯선 형태를 설정하여 일반부사로 처리하는 것도 부담스럽다. 따라서 언중에게 익숙한 표기형 그대로를 반영하여 '왜케/MAG', '왤케/MAG'로 처리하기로 한다.

위와 달리, **경음화, 구개음화 및 첨가 현상이 표기에 반영되어 있는 경우, 또 필수적이지 않은 음운 현상이 표기에 반영되어 있는 경우에는 표기된 형태를 그대로 보존하여 분석한다.**

경음화는 유형이 다양하고 그 중에는 규칙적인 경음화도 있지만 예측 불가능한 경우도 많으며, 메신저 대화에서 입력의 비경제성에도 불구하고 경음 표기를 한 데에는 특별한 의도가 있다고도 볼 수 있다. 따라서 경음화 현상이 반영된 표기는 원래 형태로 복원하지 않고 표기형을 그대로 반영하여 분석한다.

구개음화가 표기에 반영되는 경우는 '구지(굳이), 가치(같이)' 등 주로 부사 내부에서 나타난다. 이는 표음주의 표기법을 적용한 것이라기보다 '굳다', '같다'와의 의미적 관련성을 인식하지 못한 결과로 볼 가능성이 높으므로, 원래 형태로 복원하지 않고 표기형을 그대로 반영하여 분석한다. 다만, '끝이다'가 '끄치다'로 나타나는 등 형태를 분리해야 하는 부분에서 구개음화가 반영된 경우에는 기본형으로 복원하여 '끝/NNG+이/VCP+다/EF'와 같이 분석한다.

사잇소리 첨가나 /ㄴ/ 첨가, /j/ 첨가와 같은 각종 첨가 현상은 필수적이지 않은 음운 현상이므로 표기형을 그대로 반영하여 분석한다.

'전화→저놔'에서 볼 수 있는 공명음 사이 /ㅎ/ 탈락 현상, '문법→뭄뻡', '감기→강기'에서 볼 수 있는 양순음화, 연구개음화 현상 등도 필수적이지 않은 음운 현상이다. 이런 음운 현상이 표기에 반영되어 있는 경우에도, 기본형을 복원하지 않고 표기된 형태 그대로를 반영하여 분석한다.

[예시] 갈께(갈게) [가/VV+ㄹ께/EF]
안됐따니!(안 됐다니!) [안/MAG+되/VV+었/EP+따니/EF+!/SF]
치킨먹을꺼야(거야) [치킨/NNG+먹/VV+을/ETM+꺼/NNB+이/VCP+야/EF]
다섯씨에 만나(다섯시에) [다섯/NR+씨/NNB]
→ 경음화 현상이 반영된 표기이다. 표기된 형태 그대로를 반영하여 'ㄹ께/EF', '따니/EF', '꺼/NNB', '씨/NNB'로 분석한다.

[예시] 가치가자(같이 가자) [가치/MAG+가/VV+자/EF]
→ 구개음화 현상이 반영된 표기이다. 표기된 형태 그대로를 반영하여 '가치/MAG'로 분석한다.

[참고] 끄치 없다(끝이 없다) [끝/NNG+이/JKS]
→ 형태를 분리해야 하는 부분에 구개음화 현상이 반영되어 있다. 이런 경우에는 구개음화

이전의 기본형을 복원하여 '끝/NNG+이/JKS'로 분석한다.

[예시] 다섯시반이여서요(반이어서요) [반/NNG+이/VCP+여서/EC+요/JX]

→ '어서' 대신 '여서'가 쓰여 /j/ 첨가 현상을 보여 주고 있다. /j/ 첨가 현상은 한국어의 필수적인 음운 규칙이 아니고, 또한 입력의 비경제성에도 불구하고 첨가 현상을 표기에 반영한 것이므로, 표기된 형태 그대로를 반영하여 '여서/EC'로 분석한다.

[예시] 그래가꼬(그래 갖고) [그렇/VA+어/EC+가/VX+꼬/EC]

→ '갖고[갇꼬]'에서 종성의 /ㄷ/가 탈락하는 것은 한국어의 필수적인 음운 규칙이 아니다. 그러므로 표기된 형태 그대로를 반영하여 '가/VX'로 분석한다.

대다내(대단해) [대다나/VA+아/EF]

→ '대단해[대단해]'에서 /ㅎ/가 탈락하는 것은 한국어의 필수적인 음운 규칙이 아니다. 그러므로 표기된 형태 그대로를 반영하여 '대다나/VA'로 분석한다.

[예시] 재밌으(재밌어) [재밌/VA+으/EF]

→ '어' 대신 '으'가 쓰여 고모음화 현상을 보여 주고 있다. 고모음화 현상은 한국어의 필수적인 음운 규칙이 아니므로 표기된 형태 그대로를 반영하여 '으/EF'로 분석한다.

[예시] 하쉴분(하실 분) [하/VV+쉬/EP+ㄹ/ETM+분/NNB]

→ 모음 /ㅟ/가 /ㅜ/로 나타났다. 역시 한국어의 필수적인 음운 규칙과 무관하므로 표기된 형태 그대로를 반영하여 '쉬/EP'로 분석한다.

라) 이중모음이 단모음으로 표기되어 형태 분리가 어려운 경우

단모음화는 한국어의 필수적인 음운 규칙이 아니므로 기본형으로 복원하지 않고 단모음화가 적용된 표기형을 형태 분석에도 그대로 반영하는 것이 원칙이다. 하지만 다음과 같이 용언 활용형에서 단모음화가 적용되어 어간과 어미의 분리가 어려워지는 경우가 있다. 그러한 경우에는 단모음화 적용 전의 형식을 상정하여, 형태를 분리하여 분석하기로 한다.

[예시] 부러어(원형: 부러워) [부럽/VA+어/EF]

타밧어(원형: 타 봤어) [타/VV+아/EC+보/VX+앗/EP+어/EF]

낼바바(원형: 낼 봐 봐) [낼/NNG+보/VV+아/EC+보/VX+아/EF]

잘 달래조(원형: 달래 줘) [달래/VV+어/EC+주/VX+어/EF]

→ 어간과 어미가 결합한 음절의 이중모음에서 단모음화가 일어나 어간과 어미의 분리가 어렵게 되었다. 이런 경우에는 단모음화 이전의 형식을 상정하고 어간과 어미를 분리한다.

[참고] 몬운동이야(원형: 뭔 운동이야) [몬/MMD+운동/NNG+이/VCP+야/EF]

→ 이중모음의 단모음화가 적용되어 '뭔'이 '몬'으로 나타났다. 하지만 이 경우는 형태를 분

리해야 하는 경우가 아니므로 '몬' 형태를 그대로 두고 MMD 표지를 부여한다.

2처넌(원형: 2천 원) [2/SN+천/NR+언/NNB]

→ '원'이 '언'으로 나타났다. 이 경우 역시 형태를 분리해야 하는 경우가 아니므로 '언' 형태를 그대로 두고 NNB 표지를 부여한다.

마) 오타가 발생한 경우

(1) 하나의 형태 내부에서 오타가 발생한 경우

메신저 대화 원시 말뭉치에는 의도적이거나 비의도적인 오표기형이 많이 포함되어 있다. '끝'을 '끗'으로 적거나 '생각'을 '생가'로 적은 것이 그 사례가 된다. 이처럼 하나의 형태 내부에서 오표기가 발생했지만 본래 형태가 무엇인지 파악할 수 있는 경우에는, 오표기형을 그대로 두되 본래 형태에 맞는 태그를 부여하는 것을 원칙으로 한다.

[예시] 우리 셋찌가(원형: 셋째) [셋찌/NNG]

→ '셋째'가 '셋찌'로 잘못 표기되었다. 그러나 본래 형태가 '셋째'임을 파악할 수 있으므로, '셋찌'라는 오표기형은 그대로 두되 '셋째(셋째 자식의 의미)'의 태그인 NNG를 부여한다.

[예시] 끗(원형: 끝) [끗/NNG]

→ 의도적으로 '끝'을 '끗'으로 표기한 사례이다. 역시 본래 형태가 '끝'임을 파악할 수 있으므로, '끗'이라는 오표기형은 그대로 두되 '끝'의 태그인 NNG를 부여한다.

[예시] 전문가한티(원형: 한테) [전문가/NNG+한티/JKB]

→ '한테'가 '한티'로 잘못 표기되었다. 그러나 본래 형태가 '한테'임을 파악할 수 있으므로, '한티'라는 오표기형은 그대로 두되 '한테'의 태그인 JKB를 부여한다.

[예시] 생가중(원형: 생각 중) [생가/NNG+중/NNB]

→ '생각'의 받침이 탈락되었다. 그러나 본래 형태가 '생각'임을 파악할 수 있으므로, '생가'라는 오표기형은 그대로 두되 '생각'의 태그인 NNG를 부여한다.

[예시] 자잡아주면좋겠당(원형: 잘 잡아 주면) [자/MAG+잡/VV+아/EC+주/VX+면/EC+좋/VA+겠/EP+당/EF]

→ '잘'의 받침이 탈락되었다. 그러나 맥락에서 본래 형태가 '잘'임을 파악할 수 있으므로, '자'라는 오표기형은 그대로 두되 '잘'의 태그인 MAG를 부여한다.

[예시] 싫어하니다(원형: 싫어합니다) [싫어하/VV+니다/EF]

→ 어미 'ㅂ니다'의 'ㅂ'이 탈락되었다. 이 경우에도 오표기형은 그대로 두되 'ㅂ니다'의 태그인 EF를 부여한다.

[예시] 그러닐가(원형: 그러니까) [그러닐가/MAJ]

→ ‘ㄹ’이 첨가되었다. 그러나 맥락에서 본래 형태가 ‘그러니까’임을 파악할 수 있으므로, 오표기형은 그대로 두되 ‘그러니까’의 태그인 MAJ를 부여한다.

[예시] 식사는 하셨엉요?(원형: 하셨어요?) [하/VV+시/EP+었/EP+엉요/EF+?/SF]

→ ‘ㅇ’이 첨가되었다. 그러나 본래 형태가 ‘어요’임을 파악할 수 있으므로 오표기형은 그대로 두되 ‘어요’의 태그인 EF를 부여한다.

[예시] 그러시구뇽(원형: 그러시군요) [그렇/VA+시/EP+군/EF+용/JX]

→ ‘군요’가 ‘구뇽’으로 잘못 표기되었다. ‘군’와 ‘용’으로 분리하고, ‘용’에 ‘요’의 태그인 JX를 부여한다.

(2) 형태를 분리해야 하는 부분에서 오타가 발생한 경우

아래의 예도 오표기가 발생한 예인데, 오표기가 발생한 부분에 더 분석되어야 할 대상이 있는 경우이다. **이 경우 역시 오표기형을 그대로 인정하여 형태를 분리하고 표지를 부여한다.**

[예시] 반가위요(원형: 반가워요) [반갑/VA+이요/EF]

→ 어간과 어미를 분리해야 하는 부분에서 ‘반가워요’가 ‘반가위요’로 잘못 표기되었다. 이때에도 오표기형을 그대로 인정하여 형태를 ‘반갑’과 ‘이요’로 분리하고, ‘반갑’에는 VA를, ‘이요’에는 올바른 표기형인 ‘어요’에 부여되었을 형태 표지인 EF를 부여한다.

[예시] 집앞에 다너?(원형: 다녀?) [단/VV+어/EF+?/SF]

→ 어간과 어미를 분리해야 하는 부분에서 ‘다녀’가 ‘다너’로 잘못 표기되었다. 이때에도 오표기형을 그대로 인정하여 형태를 ‘단’과 ‘어’로 분리하고, 이 중 ‘단’에는 올바른 표기형인 ‘다니’에 부여되었을 형태 표지인 VV를 부여한다.

[예시] 안녕하셰요(원형: 안녕하세요) [안녕/NNG+하/XSA+시/EP+에요/EF]

→ 선어말어미와 종결어미를 분리해야 하는 부분에서 ‘하세요’가 ‘하셰요’로 잘못 표기되었다. 이때에도 오표기형을 그대로 인정하여 종결어미를 ‘어요’ 대신 ‘에요’로 처리하고, ‘에요’에는 ‘어요’에 부여되었을 형태 표지인 EF를 부여한다.

단, 아래와 같이 더 분석되어야 할 대상이 있는 자리에서 오표기가 발생했으나, 오표기의 발음과 올바른 표기의 발음이 동일한 경우가 있다. 이런 경우에는 **올바른 표기법으로 수정한 형식을 상정하고 표지를 부여한다.** 이는 구어 지침 76쪽의 라)항과 상통하는 것이다.

[예시] 걱정되(원형: 걱정돼) [걱정/NNG+되/XSV+어/EF]

→ 어간과 어미를 분리해야 하는 부분에서 ‘돼’가 ‘되’로 잘못 표기되어 ‘되+어’로의 형태 분리가 어렵게 되었다. 하지만 오표기인 ‘되’와 올바른 표기인 ‘돼’의 발음은 동일하다. 이런 경우에는 올바른 표기형인 ‘돼’를 상정하고 ‘되/XSV+어/EF’로 형태 표지를 부여한다.

[예시] 마지막으로 간개(원형: 간 게) [가/VV+ㄴ/ETM+거/NNB+이/JKS]
→ 의존명사와 격조사를 분리해야 하는 부분에서 '게'가 '개'로 잘못 표기되었다. 하지만 오표기인 '개'와 올바른 표기인 '게'의 발음은 동일하다. 이 경우 역시 올바른 표기형인 '게'를 상정하고 '거/NNB+이/JKS'로 형태 표지를 부여한다.

(3) 오타로 인해 불완전한 모아쓰기가 이루어진 경우

아래와 같이 오타로 인해 초성자 또는 모음자가 없이 불완전하게 모아쓰기가 이루어진 경우에는, 해당 요소를 포함하여 관련된 형태에 NA를 부여한다. 앞서 본 'ㅋㅋ'와 같이 의도적으로 초성자만을 사용한 경우가 아니라, 오타로 인해 초성자 또는 모음자만이 남은 경우에 NA 처리를 하는 것이다.

[예시] 핵ㅅ는데(원형: 했는데) [하/VV+앜ㅅ/NA+는데/EC]
→ '했'이 오타로 인해 '핵ㅅ'로 나타났다. '했'이 본래 '하/VV+았/EP'으로 분석됨을 고려하면, 이 경우에는 '하/VV+앜ㅅ'으로 분석할 수 있다. 이 중 불완전한 모아쓰기를 포함하는 단위인 '앜ㅅ'은 NA로 처리한다.
[예시] 눈온다고 하던데ㅛ(원형: 하던데요) [하/VV+던데/EC+ㅛ/NA]
요청 왔음ㄴ(원형: 왔으면) [오/VV+았/EP+음ㄴ/NA]
좋았던것 같ㅇ(원형: 같아) [같/VA+ㅇ/NA]
하지ㅏㄴ(원형: 하지만) [하지ㅏㄴ/NA]
무ㄴ의(원형: 문의) [무ㄴ의/NA]
먹을줌ㄹ아시네(원형: 먹을 줄 아시네) [먹/VV+을/ETM+줌ㄹ/NA+알/VV+시/EP+네/EF]
카톡하고 있ㄴㄴ(원형: 있는) [있/VX+ㄴㄴ/NA]
ㅡㄱ대노(원형: 극대노) [ㅡㄱ대노/NA]
절레절ㄹ레 [절레/XR+절ㄹ레/NA]
살펴아ㅓ됨 [살피/VV+어아ㅓ/NA+되/VV+ㅁ/ETN]
생각이 나사ㅓ [나/VV+아사ㅓ/NA]
빠3・지니까(원형: 빠지니까) [빠3・지/NA+니까/EC]
→ 모두 불완전하게 모아쓰기가 이루어진 부분을 포함하고 있는 예로서, 불완전한 모아쓰기와 관련된 형태에 NA 태그를 부여한다.
[예시] 아ㅏㅏㅏ니ㅣ [아ㅏㅏㅏ니ㅣ/NA]
타구시퍼ㅓㅓ [타/VV+구/EC+싶/VX+어/EF+ㅓㅓ/NA]
→ 긴 소리를 표시하기 위해 동일한 모음을 의도적으로 여러 번 적은 것으로 보이는 예이다. 그러나 불완전한 모아쓰기를 포함하고 있다는 점에서는 위에서 제시한 예들과 동일

하므로, 최대한 분석하되, 불완전한 모아쓰기와 관련된 형태에 NA 태그를 부여한다.

[예시] 공부했죠ㅡ [공부/NNG+하/XSV+았/EP+죠/EF+ㅡ/NA]

안됑ㅇ [안/MAG+되/VV+엉/EF+ㅇ/NA]

→ 역시 최대한 분석하되, 불완전한 모아쓰기에 해당하는 'ㅡ'와 'ㅇ'에 NA를 부여한다.

[예시] ㅇ.. 아닐거야 [ㅇ/NA+../SE]

→ 망설이는 어감을 위해 의도적으로 초성자만을 따로 쓴 것으로 보인다. 이 경우에도 불완전한 모아쓰기를 포함하고 있다는 점에서는 위에서 제시한 예들과 동일하므로, 불완전한 모아쓰기에 해당하는 'ㅇ'에 NA를 부여한다.

[예시] ㅁㅏ치 [ㅁㅏ치/NA]

→ 역시 의도적으로 초성자와 모음자를 분리하여 쓴 것으로 보인다. 하지만 이 경우에도 불완전한 모아쓰기를 포함하고 있으므로, NA를 부여한다.

바) 탈자로 인해 형태 표지 부여가 어려운 경우

아래와 같이 탈자로 인해 형태 표지 부여가 어려운 요소나 자음 하나로 구성된 형태가 발생하는 경우에는, 해당 부분에 NA(분석 불능 범주), NV(용언 추정 범주), NF(명사 추정 범주) 중 하나를 부여한다.

[예시] 하지 못는(원형: 못하는) [못/NV+는/ETM]

→ 보조용언 '못하-'에서 '하'가 탈락되었다. 부사 자격을 갖는 '못'에 VX(보조용언)를 부여하기에는 무리가 있으므로, NV(용언 추정 범주)로 처리한다.

[예시] 전 무섭더라궁(원형: 무섭더라구요) [무섭/VA+더라구/EF+ㅇ/NA]

→ 탈자가 발생하여 '구요'가 '궁'으로 잘못 표기되었고, 이에 따라 자음 하나로 구성된 'ㅇ' 형태가 남게 되었다. 이런 경우 'ㅇ'에 NA를 부여한다.

사) 띄어쓰기 오류로 인해 형태 표지 부여가 어려운 경우

'그럴걸'이 '그럴 걸'로, '엄청나다'가 '엄청 나다'로 띄어쓰기와 함께 전사된 경우가 있다. 이 때 띄어쓰기 오류로 인해 '걸'의 처리, '나-'의 처리가 어려워진다. '나-'같이 용언의 성격을 띠는 요소에 대해서는 최대한 형태 표지를 부여한다. 그 외의 경우에도 최대한 형태 표지를 부여하지만, 형태 표지 부여가 어려운 요소에는 NA(분석불능범주), NV(용언추정범주), NF(명사추정범주)를 부여한다.

[예시] 그럴 걸 [그러/VV+ㄹ/ETM, 걸/NA]

[예시] 엄청 나시네요 [엄청/MAG, 나/VA+시/EP+네/EF+요/JX]

→ '엄청나다'의 '나-'는 본 지침에서 분석하지 않는 형용사 파생 접미사이다. 용언의 성격을 띠는 요소에는 최대한 형태 표지를 부여하여 형용사로 처리한다.

[예시] 여기는 강화도 에요 [강화도/NNP, 이/VCP+에요/EF]

→ '강화도'와 '에요'가 띄어쓰기로 분리되었다. 이때에도 띄어쓰기를 하지 않은 경우와 마찬가지로 '강화도'와 '에요' 사이에 있는 지정사를 복원하여 분석한다. 이때 지정사는 '에요' 어절에서 복원한다.

[예시] 보고시 퍼 [보/VV+고/EC+시/NA, 퍼/NA]

→ '시'와 '퍼'가 분리됨으로써 형태 표지를 부여할 수 없게 되었다. 이때는 '시'와 '퍼' 각각을 NA로 처리한다.

아) 의미 파악이 어려운 요소

아래와 같이 의미 파악이 어려운 요소에는 NA를 부여한다.

[예시] 아닌것같기돗나그해서요 [아니/VCN+ㄴ/ETM+것/NNB+같/VA+기/ETN+도/JX+**ㅅ나그/NA**+하/VX+아서/EC+요/JX]

[예시] 마점자ㅓㅓ [**마점자ㅓㅓ/NA**]

→ 의미 있는 요소를 최대한 분석하되, 의미를 알 수 없는 'ㅅ나그', '마점자ㅓㅓ'와 같은 요소에는 NA를 부여한다.

의미 파악이 어려운 요소가 아니라면 NA를 부여하지 않고 최대한 적합한 형태 표지를 부여한다. 가령 아래 예는 대화 참여자가 오타와 교정자를 반복해서 입력한 경우인데, 앞의 '개'가 뒤에 오는 '게'의 오타임을 알 수 있고, '개'는 '거/NNB+이/JKS'로 분석할 수 있으므로 이를 고려하여 최대한 형태 표지를 부여한다.

[예시] 그런 개 개 게 여깄어? [거/NNB+이/JKS, 거/NNB+이/JKS, 거/NNB+이/JKS]

자) 하나의 형태 내부에 한글 외의 기호가 삽입된 경우

아래와 같이 하나의 형태 내부에 한글 외의 기호가 삽입되는 경우가 있다. 한글과 기타 기호는 분리하여 분석하는 것이 원칙이지만, 분리할 경우 형태 표지를 부여하기 어려운 요소 또는 어근이 남는다면 한글과 기타 기호를 묶은 단위에 형태 표지를 부여한다. 이는 문어 지침 5쪽의 (라)항, 61쪽의 바)항과 상통하는 것이다.

[예시] 당.연. [**당.연/NNG**+./SF]

→ 강조를 위해 음절 사이에 마침표를 삽입한 경우이다. 마침표를 따로 분리할 경우 '당'과 '연'이라는 어근이 남게 되므로, 마침표를 포함하여 '당.연'에 NNG를 부여한다.

[예시] 날씬~하구요 [**날씬~하/VA**+구/EC+요/JX]

→ '날씬'과 '~', '하'를 분리할 경우 어근에 해당하는 '날씬'이 남게 되므로 '날씬~하'를 묶어 VA를 부여한다.

[예시] 하2 (하이) [**하2/IC**]

→ '하이'라는 감탄사 자격을 갖는 인사말이 한글과 숫자를 이용해 표기되었다. 한글과 숫자를 분리할 경우 '하'라는 형태 표지를 부여하기 어려운 요소가 남게 되므로, '하2'를 묶어 IC를 부여한다.

[참고] RGRG(알지알지) [RGRG/SL]

→ 이처럼 형태 전체가 한글 외의 기호로 적힌 경우에는, 해당 기호에 할당된 형태 표지를 부여한다. 이 경우 SL(외국어) 표지를 부여한다.

[참고] 왜안3(왜 안 삼?) [왜/MAG+안/MAG+3/SN]

→ '사/VV+ㅁ/ETN'으로 분석되어야 할 요소가 숫자 3으로 적혀 더 이상 분석이 어려우므로 SN(숫자)으로 처리한다.

차) 끊어진 말

입력 실수로 오타가 발생한 부분만을 고쳐서 다시 입력하거나, 스페이스 또는 엔터를 잘못 누름으로써 한 단어 속의 일부 음절이 끊어져 나오는 경우가 있다. 끊어져 나온 요소에 형태 표지를 부여하기 어려운 경우, 해당 요소에 NA 표지를 부여한다.

[예시] 에밀리 블런드 [에밀리/NNP, 블런드/NNP]
트 [트/NA]

→ '에밀리 블런트'를 입력하려 했으나 '트'를 '드'로 잘못 입력하였고, 이에 '트'만을 다시 입력한 경우이다. 이때 끊어진 '트'에 품사 표지를 부여하기 어려우므로 NA 표지를 부여한다.

[예시] 구역식 구들 [구역/NNG+식/NA, 구/NA+들/XSN]

→ '식구들'이라고 입력해야 할 것을 스페이스를 잘못 눌러 '식 구들'로 입력하였다. 이때 끊어진 '식'과 '구'에 품사 표지를 부여하기 어려우므로 NA 표지를 부여한다.

제4장 결론

이 사업은 인공지능 발전을 위한 우리말 기초 자원으로 활용될 고품질의 한국어 어휘의미 분석 말뭉치 및 형태 분석 말뭉치를 구축하고, 분석 말뭉치 구축을 위한 표준적인 지침을 개발·정비하는 데 주요 목적이 있다.

사업의 범위는 크게 네 부분으로 나눌 수 있다. 첫째는 어휘의미 분석 말뭉치 구축 지침 수립으로, 2019년도에 마련된 체언류 어휘의미 분석 지침에 더해 용언류 어휘의미 분석 지침을 새로이 마련한다. 둘째는 어휘의미 분석 말뭉치 구축으로, 어휘의미 분석 말뭉치 구축 지침을 바탕으로 총 400만 어절 규모(문어 200만 어절, 구어 100만 어절, 메신저 대화 100만 어절)의 어휘의미 분석 말뭉치를 구축한다. 셋째는 형태 분석 말뭉치 구축 지침 수립으로, 2019년도에 마련된 문어·구어 형태 분석 말뭉치 구축 지침에 더해 메신저 대화 형태 분석 지침을 새로이 마련한다. 넷째는 형태 분석 말뭉치 구축으로, 형태 분석 말뭉치 구축 지침을 바탕으로 총 100만 어절 규모(메신저 대화 100만 어절)의 형태 분석 말뭉치를 구축한다.

○ 용언류 어휘의미 분석 말뭉치 구축 지침 수립

용언류의 어휘의미 분석을 위하여 2019년에 마련한 체언류의 어휘의미 분석 지침을 기반으로 삼되 비유 표현 등에 나타나는 용언의 비유적 의미를 허용하는 방향으로 분석 지침을 마련하였다. 그리고 용언류의 어휘의미를 분석하는 과정에서 활용할 수 있는 <우리말샘>의 뜻풀이와 예문, 문장 구조, 공기하는 체언의 의미 부류 등의 다양한 기준을 지침에 명시하였다.

이렇게 수립된 체언류와 용언류의 어휘의미 분석 말뭉치 구축 지침은 메신저 대화 말뭉치에 나타나는 체언과 용언을 분석하는 데에도 크게 문제가 되지 않았다. 다만 초성

단어 등과 같은 메신저 대화에 나타나는 다양한 표기형의 분석 방안을 세부 지침으로 추가하였다. 이를 통해 다른 영역의 형태 분석 말뭉치를 대상으로 어휘의미 분석 말뭉치를 만들더라도 올해 마련한 어휘의미 분석 말뭉치 구축 지침이 적용될 수 있음을 확인하였다.

○ 어휘의미 분석 말뭉치 구축

본 사업에서는 2019년에 구축된 국립국어원 어휘의미 분석 말뭉치(300만 어절)에 나타나는 용언류의 어휘의미를 분석하여 분석 대상 범주가 확장된 어휘의미 분석 말뭉치를 구축하였다. 또한 본 사업에서 올해 구축한 메신저 대화 형태 분석 말뭉치를 대상으로 메신저 대화 어휘의미 분석 말뭉치를 구축하였다.

어휘의미 분석 말뭉치 구축은 어휘의미 분석 지침 수립 → 분석 도구(워크벤치) 구현 → 작업 교육 → 의미번호 부착 → 말뭉치 검증 → 최종 결과물 산출의 순으로 이루어졌다.

이 중 말뭉치 검증은 분석이 완료된 어휘의미 분석 말뭉치와 형태 분석 말뭉치에서 무작위로 5,000개 어절을 추출하여 상위 작업자 그룹이 만든 정답 말뭉치와 비교하는 방식으로 진행하였다. 그 결과 문어 어휘의미 분석 말뭉치는 93.01%, 구어 어휘의미 분석 말뭉치는 95.41%, 메신저 형태 분석 말뭉치는 99.37%, 메신저 어휘의미 분석 말뭉치 95.96의 일치율을 보였다.

○ 메신저 대화 형태 분석 말뭉치 구축 지침 수립

2019년도에 구축된 메신저 대화 원시 말뭉치를 대상으로 형태 분석을 수행하기 위하여, 2019년도에 마련된 형태 분석 말뭉치 구축 지침을 기반으로 삼되 메신저 대화의 특수성을 고려한 메신저 대화 형태 분석 지침을 새로이 마련하였다.

메신저 대화는 문자를 통하여 이루어진다는 점에서는 문어의 속성을 지니지만 실시간으로 즉각적인 양방향 소통이 일어난다는 점에서는 구어의 속성을 지니는데, 이에 따라

메신저 대화에는 전형적인 문어나 전사된 구어와 다른 특수한 언어 현상들이 포함된다. 이를 고려하여 본 사업에서는 아래의 세 가지 내용을 골자로 하는 메신저 대화 형태 분석 지침을 마련하였다.

① 원시 말뭉치에 포함된 표지 및 기호의 처리
- 개인정보를 치환한 표지의 처리, 이모티콘의 처리 지침을 명시하였다.

② 메신저 대화에서 자주 나타나는 언어 현상의 처리
- 사전에 등재되지 않은 다양한 의성의태어와 감탄사, 어미, 각종 신어의 처리 지침을 명시하였다.

③ 메신저 대화에서 나타나는 특수한 표기법의 처리
- 초성만으로 표기한 단어, 음절을 첨가하여 장음을 표시한 경우, 표음주의 표기법을 적용한 경우, 오타가 발생한 경우, 하나의 형태 내부에 한글 외의 기호가 삽입된 경우 등의 처리 지침을 명시하였다.

○ 100만 어절 규모의 메신저 대화 형태 분석 말뭉치 구축

본 사업에서는 2019년도에 구축된 메신저 대화 원시 말뭉치를 대상으로 100만 어절 규모의 형태 분석 말뭉치를 구축하였다. 원시 말뭉치의 각 어절을 대상으로 형태를 분리하고 형태 분류 표지(세분류 47종)를 부착하는 작업을 하였는데, 형태 분리의 기준이 되는 단위는 기본적으로 <우리말샘>에 등재된 단어이되, 생산성이 비교적 높은 접사도 분리하는 것을 원칙으로 삼았다. 또한 메신저 대화의 특수성을 고려하여 마련된 메신저 대화 형태 분석 지침의 내용을 적용하여 형태 분석을 수행하였다.

형태 분석 말뭉치 구축은 형태 분석 지침 수립 → 분석 도구(워크벤치) 구현 → 작업 교육 → 자동 형태소 분석 → 분석 오류 수정 → 최종 결과물 산출의 순으로 이루어졌다.

이 중 분석 오류 수정은 3단계로 이루어졌다. 1단계는 작업자가 원시 말뭉치에 대한

자동 형태 분석 결과를 수정하는 단계이다. 2단계는 자동 형태 분석 결과와 작업자의 오류 수정 결과를 비교하며 검수자가 형태 분석 결과를 검수하는 단계이다. 3단계는 전체 작업 결과물에 대해 상위 작업자 그룹이 형태 결합 오류 목록, 어절 분석 중의성 목록 등을 검토하며 오류를 수정하는 단계이다.

본 사업에서는 말뭉치 구축의 편의를 도모하고 정확성을 높이기 위하여 높은 분석 정확률을 갖춘 형태소 분석기(서울대 형태소 분석기)를 사용하였다. 서울대 형태소 분석기는 세종 형태의미 분석 말뭉치(약 1200만 어절 규모)의 오류를 철저히 수정한 결과를 딥러닝의 훈련 자료로 삼아 개발한 것이다.

한편으로 형태 분석 말뭉치 구축에 최적화된 워크벤치를 사용하였다. 워크벤치에서는 서울대 형태소 분석기의 어절 분석 결과를 보여 주되 그것을 손쉽게 수정할 수 있게 하였고, 드롭다운 선택 방식 및 오류 검사를 통해 입력 오류를 원천적으로 차단함으로써 형태 분석 및 검수의 효율을 높였다.

Abstract

Research and Analysis of Korean Sense-tagged corpus

The main purpose of this project is to build a high-quality Korean sense-tagged corpus and morphological analysis corpus that will be used as basic resources for the development of artificial intelligence, and to develop and maintain standard guidelines for building an annotated corpus.

The scope of the business can be largely divided into four parts. The first is to establish guidelines for constructing a sense-tagged corpus, and in addition to the guidelines for sense-tagging for nous classes prepared in 2019, a new guideline for lexical meaning analysis of terminology is prepared. The second is to construct a sense tagged corpus, and build a lexical semantic analysis corpus with a total of 4 million words (written 2 million words, spoken words 1 million words, messenger conversation 1 million words) based on the lexical meaning analysis corpus construction guidelines. The third is to establish guidelines for establishing a corpus for morphological analysis. In addition to the guidelines for building a corpus for morphological analysis of written and spoken words prepared in 2019, a new guideline for analyzing the form of messenger conversations is prepared. The fourth is to build a morphological analysis corpus, and build a morphological analysis corpus with a total size of 1 million words (1 million words in messenger dialogue) based on the guidelines for morphological analysis corpus construction.

○ Establish guidelines for constructing corpus by analyzing vocabulary meanings of terminology

In order to analyze the lexical meaning of the idiomatic expressions, the guidelines for the analysis of the lexical meanings of the prophecies prepared in 2019 are based

on the guidelines for the analysis of the vocabulary meanings of the prophecies, but the analysis guidelines have been prepared in a direction that allows the figurative meaning of the proverbs appearing in metaphors. In addition, various standards such as meaning interpretation and example sentences, sentence structure, and meaning classes of prose speeches that can be used in the process of analyzing the lexical meanings of the prongs are specified in the guidelines.

The established guidelines for vocabulary and meaning analysis of vocabulary and verbal idioms were not a big problem in analyzing vocabulary and verbal idioms appearing in the messenger conversation corpus. However, analysis methods of various notation types appearing in messenger conversations such as initial words were added as detailed guidelines. Through this, it was confirmed that even if a vocabulary semantic analysis corpus was created for morphological analysis corpora in other areas, the vocabulary semantic analysis corpus construction guidelines prepared this year can be applied.

○ Vocabulary meaning analysis corpus construction

In this project, a vocabulary semantic analysis corpus was constructed by analyzing the vocabulary meanings of terminology appearing in the vocabulary semantic analysis corpus (3 million words) established in 2019. In addition, a corpus for analyzing the vocabulary meaning of messenger conversations was constructed for the analysis corpus of the messenger conversation form established this year.

The vocabulary semantic analysis corpus was constructed in the order of vocabulary semantic analysis guide establishment → analysis tool (workbench) implementation → work training → semantic number attachment → corpus verification → final result calculation.

Among them, the corpus verification was conducted by randomly extracting 5,000 words from the analyzed sense-tagged corpus and POS annotated corpus and

comparing them with the correct answer corpus created by the upper worker group. As a result, there was a concordance rate of 93.01% for the written vocabulary meaning analysis corpus, 95.41% for the colloquial vocabulary meaning analysis corpus, 99.37% for the messenger form analysis corpus, and 95.96 for the messenger vocabulary meaning analysis corpus.

○ Establish guidelines for building a corpus for analyzing the form of messenger conversation

In order to perform shape analysis on the primitive messenger dialogue corpus established in 2019, a new guideline for analysis of messenger dialogue types was prepared based on the guideline for establishing the shape analysis corpus established in 2019, but taking into account the specificity of the messenger dialogue.

Messenger conversation has the property of written language in that it takes place through text, but it has the property of spoken language in that instant two-way communication takes place in real time.Therefore, messenger conversations have special language phenomena other than typical written or transcribed spoken language. Included. In consideration of this, this project has prepared guidelines for analyzing the form of messenger conversations, which focus on the following three contents.

① Handling of signs and symbols included in the raw corpus

-The instructions for the treatment of the cover with replaced personal information and the treatment of emoticons are specified.

② Handling of language phenomena that often appear in messenger conversations

-Guidelines for handling various onomatopoeia, interjections, endings, and various new words that were not previously listed were specified.

③ Handling of special notations in messenger conversations

-Instructions for handling words such as words and syllables marked only with

initial vocalizations, marked long notes, applied phoneticism, typos, and inserted symbols other than Hangul within one form were specified.

○ Build a corpus for analyzing the form of messenger conversations with a scale of 1 million words

In this project, a morphological analysis corpus with a scale of 1 million words was established for the original messenger conversation corpus established in 2019. For each word of the primitive corpus, the morphology was separated and a morphological classification mark (47 subclasses) was attached. The standard unit for morphological separation is basically a word registered in the <Korean Word>, but the productivity is relatively high. Separation of high affixes was a rule. In addition, shape analysis was performed by applying the contents of the guidelines for analyzing the form of messenger conversations prepared in consideration of the specificity of the messenger conversation.

The morphology analysis corpus was constructed in the following order: establishment of morphology analysis guidelines → implementation of analysis tools (workbench) → work training → automatic morpheme analysis → correction of analysis errors → final result calculation.

Among them, the correction of analysis errors was made in three stages. Step 1 is a step in which the operator corrects the results of the automatic shape analysis of the raw corpus. In the second step, the result of automatic shape analysis is compared with the result of error correction by the operator, and the inspector inspects the result of the shape analysis. Step 3 is a step in which the upper level worker group reviews the form combination error list, the word analysis ambiguity list, etc. for the entire work result and corrects the error.

In this project, a morpheme analyzer (Seoul National University morpheme analyzer) with a high analysis accuracy was used to facilitate the construction of the corpus and

increase the accuracy. The Seoul National University morpheme analyzer was developed using the result of thorough correction of errors in the Sejong type semantic analysis corpus (about 12 million words scale) as training data for deep learning.

On the one hand, a workbench optimized for morphological analysis corpus construction was used. The workbench shows the result of the word analysis of the SNU morpheme analyzer, but makes it easy to correct it, and by fundamentally blocking input errors through a drop-down selection method and error check, the efficiency of shape analysis and inspection has been improved.

연구 기관: 성신여자대학교 산학협력단 · 고려대학교 산학협력단 · (주)언어과학 · 홍익대학교 산학협력단

연구진	
연구 책임자	김일환(성신여자대학교)
공동 연구원	박진호(서울대학교)
	유현조(서울대학교)
	윤태진(성신여자대학교)
	이규범(고려대학교)
	이도길(고려대학교)
	장원철((주)언어과학)
	정슬아(성신여자대학교)
	정연주(홍익대학교)
	이수화(엠티콤)
	이바다(엠티콤)
	김만수(엠티콤)
연구 보조원	강수연(고려대학교)
	강주은(서울대학교)
	고동현(서울대학교)
	김다미(서울대학교)
	김동근(홍익대학교)
	김연우(고려대학교)
	김은진(서울대학교)
	김희숙(서울대학교)
	박종우(성신여자대학교)
	백인영(서울대학교)
	신현규(서울대학교)

	웅연(서울대학교)
	이강혁(서울대학교)
	이세은(고려대학교)
	이순욱(서울대학교)
	이윤경(서울대학교)
	이준희(고려대학교)
	임민식(홍익대학교)
	전진호(서울대학교)
	정우헌(서울대학교)
	최준호(서울대학교)
	최지선(성신여자대학교)
	홍승혜(고려대학교)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2020년 12월 11일

발행일: 2020년 12월 11일
인 쇄: 성신POD

※ 이 책은 국립국어원의 용역비로 수행한 '어휘의미 말뭉치 연구 분석' 사업의 결과물을 발간한 것입니다.