비교문화연구 제53집 (2018.12) pp.163~192

다국어 범용 의존관계 주석체계(Universal Dependencies) 적용 연구
– 한국어와 일본어의 비교를 중심으로*

한 지 윤
(연세대학교 박사과정생, 1저자)
이 진
(연세대학교 박사과정생, 2저자)
이 찬 영
(연세대학교 박사과정생, 3저자)
김 한 샘
(연세대학교 부교수, 교신저자)

국문초록

이 논문은 형태통사적 특성이 유사한 한국어와 일본어의 다국어 범용 의존관계 주석체계(Universal Dependencies, 이하 UD) 적용 사례를 살펴보고 비교 분석을 통해서 한국어의 UD 적용 및 개선 방안을 고찰하는 것을 목적으로 한다. 한국어와 일본어는 교착어적 특성으로 인하여 어미와 조사가 매우 발달되어 있다. 그러므로 영어와 같은 굴절어를 중심으로 설계된 UD를 적용하는 데에 많은 어려움이 있다. 이에 본고에서는 UD를 구성하는 범용 품사 주석(Universal POS, 이하 UPOS)과 범용 의존관계 주석(Universal Dependency Relations, 이하 DEPREL)의 적용과 그에 따른 논의들을 검토하였다. UPOS의 경우 AUX(조동사 표지), ADJ(형용사 표지), VERB(동사 표지)처럼 서술어와 관련된 주석 표지의 처리와 조사, 어미와 같은 기능어의 처리 방안을 살펴보았으며 접속사 및

* 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2009-361-A00027)

이와 관련된 단위를 어떻게 처리하고 있는지 검토하였다. DEPREL과 관련해서는, 구문 표지를 주석하는 기본 단위의 문제에서 출발하여 통사적 문제를 어떻게 반영하고 있는지 살펴보았다. 지배소 설정 방식과 병렬 구조의 주석 방식, case(격 관계 표지)와 aux(조동사 관계 표지) 주석 방식을 검토하였다. 다양한 관계 주석 표지 중에서 특히 case와 aux에 집중하여 논의한 것은 한국어와의 주석 표지 적용 양상을 비교했을 때 분포 상 가장 두드러지는 차이를 나타내기 때문이다. case는 한국어와 일본어 모두 조사와 관련이 있고, aux는 한국어에서는 보조용언, 일본어에서는 조동사와 관련이 있는 표지이다. 구체적인 주석 양상을 살펴본 결과 일본어의 aux는 서법 조동사뿐만 아니라 동사에 문법적 의미를 더하는 요소, 한국어의 어미에 해당하는 형태에도 aux를 할당하기 때문에 주석이 차지하는 비율이 크게 차이가 나는 것으로 밝혀졌다. iobj(간접목적어 관계 표지)와 관련해서는 일본어에서 간접목적어를 인정하는 데에 반해 한국어에서는 간접목적어를 인정하지 않는 경우가 더 많았다. 일본어의 UD 주석에서 형태 분석 기본 단위인 '단단위'를 기본 구문 주석 단위로 하되 '장단위'와 문절 정보를 이용하는 것처럼, 한국어에서도 형태 분석 단위를 의존관계 주석의 정보로 활용하는 방안에 대해서 고려할 필요가 있다.

주제어 : Universal Dependencies, UD, 의존관계 분석, 형태 분석, 한국어, 일본어

Ⅰ. 머리말

이 논문은 한국어와 일본어의 다국어 범용 의존관계 주석체계(Universal Dependencies, 이하 UD)의 적용 양상을 비교 분석하여 한국어에 UD를 효율적으로 적용하고 기구축된 UD 주석 말뭉치의 개선하기 위한 기초연구이다. UD는 다국어 트리뱅크(treebank) 말뭉치를 일관되게 주석하기 위한 프레임워크(framework)로 50개 이상의 언어에 적용되어 있다. 이 프레임워크는 유형적으로 다른 언어 사이의 유사성을 포착하는 데 초점을 두고 있으며, 범언어적 언어 처리를 위해 서로 다른 특성을 지닌 언어에도 공통적으로 적용할 수 있는 품사 주석 표지체계(Universal POS, 이하 UPOS)와 구문 분석 표지체계(Universal Dependency Relations, 이하 DEPREL)를 제안하고 있다. 이는 언어 자원을 하나의 통일된 형식으로 변환하여 통합 처리를 수

월하게 할 수 있도록 하기 위함이다. 이를 통해 자연어 처리 기술의 성능을 높이고, 언어 간 대조 연구를 돕는 것을 목적으로 한다.

UD 이전에도 언어 자원을 다루기 위한 언어 주석체계에 대한 연구는 활발하게 이어져 왔다. 주석체계는 언어 자원을 처리하는 기준이 되므로, 다양한 언어 자원이 동일한 주석체계로 주석되어 있을수록 그 효용가치가 높아진다. 따라서 언어마다 그 언어를 대표하는 표준적인 주석체계를 확립하려는 연구가 존재한다. 한국어의 경우 품사 주석 표지체계는 '21세기 세종계획'에서 설계한 형태 주석체계가 보편적으로 사용되고 있다. 구문 주석의 경우도 '21세기 세종계획'에서 제시한 표지체계를 일반적으로 활용하고 있다. 다만 '21세기 세종계획'의 구문 주석 말뭉치는 구구조(phrase structure)를 기반으로 구축되었기 때문에 의존구조의 표현에 적합한지에 대한 검토가 필요하다. 따라서 한국정보통신기술협회(TTA)에서는 기존 세종 구문주석 표지체계를 의존관계 표지체계로 전환하고 의존관계를 어떻게 설정해야 하는지에 대한 지침을 제시하였다. 이는 한국어의 특성을 잘 표현하기 위해 설계된 것이기 때문에 여러 언어에 두루 적용될 수 있는 범용 주석체계로서의 역할을 하기에는 한계가 있다. 기계 번역과 같이 다국어를 병렬적으로 처리해야 하거나, 서로 다른 유형의 언어를 대조하여 연구할 때 전산적인 방법을 도입하기 위해서는 범용으로 활용할 수 있는 국제적인 표준이 필요하다. 이에 따라 UD와 같은 다국어 범용 의존관계 주석체계가 고안된 것이다. 그런데 UD는 언어 자원이 풍부한(resource-rich) 언어인 영어와 같은 굴절어를 중심으로 구축되었기 때문에 교착어인 한국어에 적용하는 데에 어려움이 발생한다. 한국어는 교착어적 특성으로 인해서 조사와 어미가 매우 발달해 있는데 이것이 UD를 한국어에 적용하는 데 있어서 큰 걸림돌로 작용하고 있다. 이러한 문제는 한국어뿐만 아니라 일본어에도 유사하게 나타나는 문제이므로 이를 해결하기 위해서는 일본어와 한국어의 UD 적용 양상에 대해 비교 분석할 필요가 있다. 한국어의 유형론적 특징으로 다음 (1)과 같은 점이 언급되는데 이러한 특징이 일본어에도 그대로 적용되기 때문이다.

(1) ㄱ. 주어-목적어-동사 어순을 기본으로 하되, 어순이 자유롭다.
ㄴ. 조사와 어미가 발달된 지배소 후위 언어로 영어와 같이 어순에

의해서가 아니라 기능 형태소들이 문법 관계를 결정한다.
ㄷ. 내포절이 선행하는 언어이다.

일본어는 한국어의 경우에 비해 상대적으로 UD 적용에 관한 논의가 활발할 뿐만 아니라 지침과 예문이 UD 지침에 공개되어 있기 때문에, 이들을 참고하여 일본어의 UD 적용 양상을 한국어의 사정과 비교 분석하는 것은 한국어의 UD 적용 방안을 모색하는 데 도움이 될 것이다. 이에 본고에서는 우선 교착어라는 공통점을 지닌 한국어와 일본어에 UD를 적용하는 데 있어서 생기는 여러 문제점들을 형태 주석체계와 의존관계 주석체계로 나누어 검토한다.

II. 관련 연구

본 장에서는 한국어와 일본어의 UD 적용과 관련된 연구를 살펴보고자 한다. 우선 한국어 관련 UD 연구는 박혜진 외(2018a, b)에서 정리한 바와 같이 국외 프로젝트를 중심으로 진행되었다. 그 출발은 영어를 주석하기 위한 스탠포드 주석체계와 구글의 다국어 주석체계를 결합하고자 한 2013년의 The Google Universal Dependency Treebank(UDT) project(McDonald *et al.* 2013)이다. 이 논의는 한국어를 포함한 여섯 개 언어의 UD 적용 방안을 다루고 있다. 한편 UD 주석체계를 한국어에 적용하고자 하는 노력 역시 꽤 활발하게 이루어져 왔다. 오진영・차정원(2013), Choi & Palmer(2011), Choi *et al*.(2013) 등에서는 기존에 구축되어 있는 주석체계, 즉 '21세기 세종계획'에서 구축된 POS 주석이나 구문 주석체계를 UD 주석체계에 활용하려는 시도의 일환으로, 그 과정 속에서 검토하고 개선해야 할 점들에 대하여 구체적으로 논의하였다. Chun *et al*.(2018)에서는 지금까지 한국어에 적용된 대표적인 세 가지 의존관계 주석 말뭉치인 Google UD Treebank, the Penn Korean Treebank, KAIST Treebank에 대해 구체적으로 논의하였다. 한편 Park *et al*.(2016)에서는 기존 형태 주석체계와 UD의 한국어 UPOS의 대응 관계를 최초로 제시하였다.

다음으로 일본어의 UD 적용과 관련된 연구를 살펴보면 가장 많은 논의

가 이루어지고 있는 부분은 주석 단위의 문제이다. 일본어는 한국어와 달리 띄어쓰기가 없으므로 가시적으로 분할된 표층 단위가 없다. 종래의 일본어 의존구조 분석에서는 통사론적 단위로 문절을 이용했지만 최근에는 단단위(Short unit word, 短單位)을 기본으로 하고 장단위(Long unit word, 長單位)와 문절 정보를 활용하고 있다.

金山博 외(2015)는 UD의 일본어 버전 설계의 첫 번째 보고서로 단단위를 기초로 하는 UD 적용을 논의하였다. 그 이후 논의인 大村舞 외(2017)에서는 주석 단위로 단단위와 장단위를 모두 적용해 볼 필요가 있다고 밝히고 있다. 현대일본어문어균형말뭉치(Balanced Corpus of Contemporary Written Japanese, 이하 BCCWJ)를 UD Japanese BCCWJ로 변환하면서 단단위에 대한 주석과 장단위에 대한 주석을 모두 수행하고 있다. 이렇듯 초기 일본어 UD 적용 논의는 단단위를 기준으로 하는 연구가 많았으며 장단위에 대해서는 별로 검토되지 않았다. 그러나 최근에는 단단위와 장단위에 대한 논의가 동시에 이루어지고 있으며 두 단위의 효용성을 비교 분석하는 논의가 진행 중이다. 한국어는 일본어와 달리 어절 단위의 구분이 있지만 한 어절 안의 내용어와 기능어가 같이 존재하므로 이러한 문제를 해결하기 위해서 일본어의 주석 단위와 관련된 논의를 자세히 살펴볼 필요가 있다. 이에 대한 논의는 3장에서 상세히 기술한다.

그 다음으로 UPOS 태그셋의 적용에 관한 논의가 있다. 일본어에서는 Unidic 형식을[1] UPOS 태그셋로 변환하는 논의가 진행되고 있다. 大村舞 외(2017), 大村舞 외(2018a)에서는 단단위를 기반으로 한 말뭉치의 경우에는 기본적으로 어휘주의(lexicon-based, 語彙主義) 기반 품사 매핑이 이루어지고 있지만 사변동사(サ変動詞)나[2] 형상사(形狀詞)와[3] 같은 몇 가지 단어에 대

1) Unidic의 형식은 '명사(名詞)-보통명사(普通名詞)-부사가능(副詞可能)'의 형태로 되어 있는데, 여기서 '부사 가능'은 명사 용법뿐 아니라 부사 용법으로도 쓰일 수 있는 어휘라는 의미이다.

2) '경동사(輕動詞, Light Verb)'라고도 부르며, 동작성 명사 뒤에 붙어서 타동사를 만드는 역할을 한다.

3) '형용동사' 혹은 '나(ナ)형용사'라고도 부르며, 동사와 형용사의 특성을 모두 가지고 있다. 현대 일본 학교문법에서는 독립적인 품사로 설정되어있으며, '이(イ)형용사'와 함께 형용사에 포함되기도 한다.

해서만 맥락을 고려하여 용법주의(usage-based, 用法主義)에 기반한 품사 체계를 적용한다고 밝혔다. 반면에 장단위를 기반으로 한 말뭉치는 맥락을 고려하여 용법주의 기반 품사 체계를 적용하였다.

마지막으로 관계 주석의 변환에 대한 논의를 살펴보면 大村舞 외(2017), 大村舞 외(2018b)에서는 일본어는 csubj, advcl, acl 등 절 여부 판별에 대한 주석 표지와 cc, conj 등과 같은 병렬 구조 관련 주석 방안이 아직 명확하게 정의되어 있지 않았다고 하면서 더 검토의 여지가 있음을 밝혔다. 최근에는 Hiroshi *et al*.(2018)에서 핵어가 문말에 오는 언어에서 나타나는 대등 구성 간의 의존관계라는 주제로 한국어와 일본어를 함께 다루어 논의하였다.

Ⅲ. 다국어 범용 의존관계 주석체계 적용 관련 논의

1. 주석체계 적용에 영향을 미치는 한국어와 일본어의 특성

한국어와 일본어의 UD 적용 논의에 앞서 한국어와 일본어의 형태통사적 유사점과 차이점을 비교 분석하고자 한다. 이는 한국어와 일본어의 형태통사적 유사점과 차이점에 대한 분석이 전제되어야 일본어의 UD 적용 양상을 바탕으로 한국어 UD 적용 및 개선 방안을 고찰할 수 있기 때문이다. 한국어와 일본어는 형태통사적으로 유사점이 많다. 교착어에 속하고 '주어-목적어-서술어'의 어순을 가진다는 공통점이 있다. 또한 한국어 및 일본어와 같이 '주어-목적어-서술어'의 어순을 가진 언어들의 가장 큰 특징은 후치사 언어에 속하며 조사와 어미가 발달된 언어라는 것이다. 이는 문법적 기능을 하는 기능어들이 내용어의 앞에 위치하는 영어 등의 언어와 비교할 때 가장 큰 차이점 중의 하나라고 할 수 있다.

이외에도 한국어와 일본어는 분류사(classifier)가 비교적 잘 발달되어 있는 언어이지만 영어의 경우는 그렇지 않다. 이때 분류사란 명사와 함께 쓰이는 단위명사로서 '강아지 두 마리, 책 한 권'과 같은 표현에서 '마리, 권'과 같이 명사의 의미적 자질을 분류해 주는 기능을 하는 표지들이다. 한편 한국어의 의존명사와 이에 대응되는 일본어의 형식명사는 관형어가 필수적

으로 선행해야만 하는 통사적 특성을 가지는데, 이러한 구성 역시 영어에는 존재하지 않는다. 물론 UD 주석체계에도 분류사를 위한 ‘clf’라는 주석이 존재한다. 그런데 이에 대해서는 주로 분류사의 사용이 한국어와 일본어에 비해 필수적이며 보다 실질적인 의미를 지니는 중국어의 예시들이 제시되어 있어서 한국어 및 일본어에 직접적으로 적용하는 데에는 어려움이 따른다. 이처럼 영어에는 존재하지 않는 형식들이 한국어와 일본어에 존재하는 경우, UD 체계를 적용하는 데 있어서 걸림돌로 작용하게 된다.

다음으로 한국어 학교 문법과 일본어 학교 문법에서 제시하고 있는 품사 체계를 비교 분석해 보면 아래 <표 1>과 같다.

〈표 1〉 한국어 학교 문법과 일본어 학교 문법의 품사 체계 비교

한국어 학교 문법	일본어 학교 문법
명사	명사
수사	
대명사	
관형사	연체사
형용사	형용사
	형용동사
부사	부사
	접속사
조사	조사
(어미)	
(어미)	조동사
감탄사	감동사

두 언어가 일치하는 품사 유형도 있지만 서로 상이한 품사 유형도 보인다. 한국어의 경우에는 ‘명사’, ‘수사’, ‘대명사’를 그 기능에 따라서 각각 분류하여 제시하고 있지만 일본어에서는 이들을 모두 ‘명사’로 제시하고 있다. 한편 한국어의 ‘관형사’는 일본어의 ‘연체사’에 대응될 수 있는데, 이는 지시사, 수량사 및 일부 형용사를 아우르는 개념이다(리우완잉, 2017). 형용사의 경우에는 한국어는 ‘형용사’를 하나로 분류하고 있지만 일본어는 ‘형용사’ 외에 ‘형용동사’라는 품사를 따로 설정하고 있다. ‘형용동사’는 ‘형용사’와

의미적으로 유사하지만 형식의 측면에서는 동사와 유사함을 가지기 때문에 이를 품사로 따로 설정하고 있다. ‘부사’의 경우에는 한국어와 일본어에서 그 범위가 다르게 나타나는데 한국어의 부사는 일본어에서 부사에 접속사가 포함된 개념으로 사용된다. 일본어의 접속사는 한국어의 부사에 대응되는데 한국어 학교 문법에서는 ‘그리고, 그러나’ 등을 ‘접속부사’로, ‘와/과’, ‘(이)랑’ 등을 ‘접속조사’로 분류하고 있다. 한국어의 접속부사에 대응되는 것을 일본어에서는 ‘접속사’라 하여 독립된 품사로 설정하고 있다. 마지막으로 조사와 조동사 역시 한국어와 일본어가 그 범위가 다르게 나타난다. 교착어적 특성을 지니는 한국어와 일본어에는 모두 조사와 어미라는 형식이 존재하며 이들이 내용어들 간의 문법적 관계 등을 나타내 주는 역할을 한다. 일본어의 경우에는 이와 관련하여 ‘조사’와 ‘조동사’ 범주가 품사 체계 내에 포함되는 반면에 한국어에는 ‘조사’만이 포함된다. 한국어에서는 전통적으로 조사만을 단어로 인정하고 어미는 용언 어간과 결합하여 별개의 단어를 이룬다고 보는 체계가 학교 문법에서 받아들여져 온 관계로 어미는 품사 체계에 포함되지 못한 것이다. 따라서 한국어의 어미 중 일부는 일본어의 조사에 대응되기도 하고 조동사에 대응되기도 한다. 상대적으로 그 경계가 뚜렷하고 어휘적 의미가 명확한 다른 품사에 비하여 조사와 어미는 경계가 모호할 뿐만 아니라 어휘적 의미보다는 문법적 의미가 강하기 때문에, 두 언어에 UD 주석 표지를 할당할 때 어려움을 야기하는 한 원인이 된다.

또한 UD 프로젝트에서 아래와 같이 제시한 설계 원칙도 교착어의 실정에는 맞지 않는 부분이 존재한다.

(2) ㄱ. 개별 언어의 언어학적 분석을 만족시켜야 한다.
ㄴ. 언어와 언어군 간의 비교에 기반한 언어 유형론 연구가 가능해야 한다.
ㄷ. 인간이 빠르고 일관성 있게 주석하기에 적합해야 한다
ㄹ. 높은 정확도로 자동 구문분석이 가능해야 한다.
ㅁ. 언어 학습자나 간단한 언어 처리를 하려고 하는 개발자 등 언어학자가 아닌 사람들도 이해하고 활용하기 쉬워야 한다.
ㅂ. 관계 추출, 독해, 기계 번역 등 기본적인 언어 이해 과제를 지원해야 한다.

한국어의 경우 UD 프로젝트에서 제시한 디자인 원칙 (2ㄱ)과 (2ㄴ)을 만족시키기는 쉽지 않다. Berdicevskis(2018)에서는 공개된 UD 언어 자원을 활용하여 언어 복잡성(Linguistic Complexity)을 측정하는 연구를 진행하였다. 총 36개 언어에 대해 TTR(Type-Token Ratio)을 비롯한 8종의 형태적 복잡성 측정 자질과 CR_POSP(POS 바이그램의 다양성) 등 7종의 통사적 복잡성 측정 자질을 적용하여 언어 복잡성의 양상을 분석하였으나 한국어, 일본어 등은 언어 분석 단위 대응이 쉽지 않아 포함되지 않았다. UD의 주석체계를 따르는 언어 자원은 10개의 열로 구성된 CoNll 형식의 데이터로 구축된다. 각 열은 분석 대상 텍스트와 그 원형, 형태 분석 정보, 구문 분석 정보 등의 정보를 보여준다. 이때 품사(Part of Speech, POS) 분석 결과는 UPOS 외에 XPOS(language specific part-of-speech layer) 필드를 만들어 개별 언어의 특성을 드러낼 수 있도록 하였다. 그러나 내용어와 기능어가 명확하게 별도의 어휘로 나뉘어져 있는 굴절어를 위해 만든 구문 분석 체계를 기능어가 내용어와 통합되어 단어, 혹은 어절 등의 단위를 이루고 있는 교착어에 바로 적용시키는 것은 쉽지 않다. 교착어의 UD 적용과 관련해 Park *et al*.(2017)에서는 우선 어절 단위로 분석을 하였고 여기에 한국어의 교착어적 특성을 살려 기호, 조사, 어미의 순으로 문법 형태소를 분리하여 분석하였다. 그 결과 문법 형태소를 더 분리할수록 자동 분석의 효율이 높아짐을 확인하였다.

2. 한국어와 일본어의 주석 단위

주석의 기본 단위를 설정하는 것은 분석의 가장 기초적인 단계이다. 형태 주석의 경우 의미 또는 기능을 지닌 언어의 최소 단위를 설정하는 것으로 언어별로 주석의 단위와 체계가 어느 정도 확립이 되어 있다. 그러나 구문 주석의 단위를 설정하는 문제는 아직 확립이 되어 있지 않으며, 언어적인 특성을 어떻게 반영할지에 대한 고민이 필요하다. UD 가이드라인은 의존관계를 통사적 단어(syntactic word) 간에 발생하는 것으로 정의한다. 그러나 통사적 단어를 정의하는 기준은 모호하게 제시하고 있어 언어마다 통사적 단어를 명징하게 정의하기 어렵다. 영어의 경우 일반적으로 품사를 주석하는 단어와 공백(white space)으로 분절되는 단위가 같아 형식적 경계와 UPOS, DEPREL의 주석 기본 단위가 일치하지만 한국어는 형태 분석의 기

본 단위가 공백으로 나누어지지 않는다. 한국어의 경우는 공백으로 분절된 단위를 어절이라 부르며, 어절은 내용어와 기능어가 조합된 형태소의 결합으로 구성되는 것이 일반적이다. 예를 들면 '학교를 가다'라는 구에서 '학교를'은 '학교'라는 명사와 '를'이라는 조사가 결합되어 구성된 어절이다. 형태 분석 단위와 공백으로 나뉘는 단위가 일치하지 않는 것이다. 영어의 경우는 동일한 의미를 'go to schoo'로 표현할 수 있는데 전치사인 'to'와 명사인 'school'이 공백으로 나뉘어져 있어 형태 분석의 기본 단위와 공백으로 나뉘는 단위가 서로 일치하는 것과 다르다.

구문 분석의 기본 단위를 구문 분석 전 단계인 형태 분석의 기본 단위와 일치시킬지, 공백으로 구분된 어절을 기본 단위로 설정할지에 대한 논의가 필요하다. 한국 국립국어원에서 구축한 21세기 세종 계획의 현대 문어 구문 분석 말뭉치(이하 세종 말뭉치)는 통사적 주석의 기본 단위를 어절로 채택하였다. 또한 임준호 외(2015)에서 밝힌 바와 같이 세종 말뭉치와 동일한 주석체계을 활용하여 의존관계를 표현한 엑소브레인의 언어 분석 말뭉치도 기본 주석 단위로 어절을 사용한다. 田中貴秋(2018)에서는 세종 말뭉치에 상응하는 일본 국립국어연구소의 BCCWJ에서 구축한 기존의 의존관계 주석 말뭉치 또한 한국어의 구문 주석 기본 단위인 어절에 대응되는 단위인 '문절(bunsetsu, 文節)'을 구문 분석의 기본 단위로 삼았다고 보고했다.[4] 그러나 추후 BCCWJ를 UD 주석체계로 변환한 UD Japanese BCCWJ를 제작하면서 기본 구문 주석 단위를 문절에서 장단위로 변환하였다. 이를 통해 해결되지 않는 것은 장단위와 문절을 이용한 규칙을 통해 사상(寫像)하였다. 단단위는 어휘에 기반한 품사 체계이고, 장단위는 구문적인 기능에 착안한 품사 체계라고 大村舞 외(2018a)에서 변별한 바 있다. 기본적으로 문절은 여러 개의 단단위와 장단위로 구성되는 경우가 있다. 한국어의 어절이 여러 개의 형태소 및 단어로 구성되는 것과 유사하다. 한국어의 경우 기존 구문 분석 말뭉치를 UD 체계에 맞춰 변환할 때 기존의 구문 분석 기본 단위인 어절을 그대로 사용할지, 일본어의 경우처럼 형태 분석 단위를 기본으로 하여 변환할지 논의가 필요하다.

4) 일본어 표기법의 특성상 문절은 공백으로 나누어지지는 않으나 해당 단위와 형태 분석 단위가 서로 다르다는 점에서 한국어의 경우와 유사한 문제에 봉착한다.

교착어인 한국어와 일본어는 어절 안에 포함된 조사와 어미가 해당 어절이 문장에서 어떠한 기능을 나타내는지 보여준다. 예를 들어 <표 2>에서 '학교 생활을'이라는 이어진 어절을 살펴보면 '학교', '생활', '을'이라는 세 개의 형태소로 구성된다. 이때 '학교'는 '생활'을 꾸며 주고, 조사인 '을'은 '학교 생활'이 문장에서 목적어의 역할을 한다는 사실을 알려 주는 목적경이다. 위의 예와 같이 형태소 간에도 중요한 문법적 정보를 전달하는 이러한 관계가 존재하지만, 어절을 기본 주석 단위로 할 경우 '학교'와 '생활을' 사이의 관계만 주석할 수 있게 된다. '을'을 별도로 주석할 경우 DEPREL 중 case 표지를 할당하여 '을'이 문장에서 격을 나타낸다는 정보를 줄 수 있으나 어절을 기본으로 주석할 경우 이 정보가 누락되는 것이다.

〈표 2〉 한국어와 일본어의 주석 단위 비교 예시

한국어 원문 : 학교 생활을 즐겁게 할 수 있을지도 모르는 방법

형태소 분석(세종)	학교 NNG	생활 NNG	을 JKO	즐겁 VA	게 EC	하 VV	ㄹ ETM
어절	학교	생활을		즐겁게		할	

형태소 분석(세종)	수 NNB	있 VX	을지 EC	도 JX	모르 VV	는 ETM	방법 NNG
어절	수	있을지도		모르는		방법	

일본어 원문 : 学校生活を楽しくするかもしれない方法

단단위 (Unidic)	学校 gakkou 名詞	生活 seikatu 名詞	を wo 助詞	楽しく tanosiku 形容詞
장단위 (Unidic)	学校生活 名詞		を 助詞	楽しく 形容詞
문절	学校生活を			楽しく

단단위 (Unidic)	する suru 動詞	か ka 助詞	も mo 助詞	しれ sire 動詞	ない nai 助動詞	方法 houhou 名詞
장단위 (Unidic)	する 動詞	かもしれない 助詞				方法 名詞
문절	するかもしれない					方法

Ⅳ. 다국어 범용 의존관계 주석체계의 적용 사례

1. 형태 주석체계의 적용

일본어는 UPOS 태그를 적용하기 위해서 UniDic 주석체계와 UPOS 주석체계의 대응표를 구축하여 UD의 품사를 정의하였다. 본고에서 세종 주석체계와 UPOS 주석체계의 대응표를 구축하여 한국어 UD의 품사를 정의하였고 이를 일본어의 UPOS 대응표와 비교 분석하고자 한다. 한국어와 일본어는 형태통사적 유사성이 많기 때문에 UPOS로 변환하는 과정에서 생기는 문제도 비슷하다. 본 장에서는 한국어와 일본어에서 교착어적 특성으로 인해 UPOS 주석에 문제가 되는 태그를 중심으로 검토해 보고자 한다.

〈표 3〉 한국어와 일본어 UPOS 변환표(일부)

UPOS	일본어(Unidic)	한국어(세종)
동사 (VERB)	동사-일반 명사-보통명사-사변동사 가능	VV+E ([NNG, NNP, MAG, XR])+XSV+E
형용사 (ADJ)	형용사-일반 형태사-일반 명사-보통명사-형태사 가능 연체사[5]	MM(성상 관형사) VA+E VCN+E ([NNG, NNP, MAG, XR])+XSA+E ([N, MAG, SN])+VCP+E
한정사 (DET)	연체사의 일부 ‘この’,‘その’,‘あんな’	MM(수・성상 관형사를 제외한 관형사)
부치사 (ADP)	격조사 계조사[6]	(JK, JX)
조동사 (AUX)	조동사 동사-비자립 가능 형용사-비자립 가능	VX+E

5) 오직 명사만을 수식하는 말로 ‘ある日’(어느 날)의 ‘ある’가 이에 해당한다.
6) 결합된 서술어와 호응관계가 나타나는 조사이다. ‘少ししか食べない。’(조금밖에 먹지 않는다.)에서 ‘しか’는 부정과 함께 호응되어 쓰이므로 계조사이다.

UPOS	일본어(Unidic)	한국어(세종)
불변화사 (PART)	종조사[7] 접미사{형태사+적(的)}	(EP, EC, EF, ET, XP, XS)
등위접속사 (CONJ)	접속사 조사-접속조사(등위접속사)	MAJ{및, 또는} JC
종속접속사 (SCONJ)	접속사 조사-접속조사(CONJ 제외) 준체조사(準体助詞)[8]	MAJ{‘및, 또는’을 제외한 모든 접속부사}

1) 조동사(AUX)

AUX는 한국어와 일본어에 유사하게 적용 가능할 것으로 보이나 주석 표지에 포함되는 구체적인 형태 범주는 다르다. 일본어는 Unidic 주석체계에서 ‘조동사’, ‘동사-비자립 가능’, ‘형용사-비자립 가능’이 AUX와 사상되었다. 일본어의 ‘동사-비자립 가능’과 ‘형용사-비자립 가능’은 한국어의 ‘보조용언’과 같은 역할을 하고 있기 때문에 한국어에 적용되는 형태 범주와 일치한다.

그러나 일본어의 경우 동사나 형용사 뒤에 붙어서 특별한 기능을 하는 것들에 ‘조동사’라는 독립적인 품사를 부여하기 때문에 이를 모두 AUX로 주석하고 있다. 그래서 (3ㄱ)의 경우 일본어에서 ‘た’를 과거를 나타내는 조동사로 보기 때문에 AUX로 주석한다. 한국어에서 일본어의 조동사 ‘た’에 해당하는 것은 ‘-았(었)다.’ 인데 한국어에서 어미는 별도의 어절을 구성하지 않고 품사로 인정되지도 않기 때문에 세종 주석체계에서 VX로 주석하는 ‘보조용언’만 UD의 AUX로 주석한다.

(3) ㄱ.

食べ tabe VERB	**た** **ta** **AUX**	。 PUNCT
먹었다 VERB		. PUNCT

7) 문말에 붙어서 화자의 감정과 심리 상태를 나타내는 조사이다. ‘行きますか?’(갑니까?)에서 ‘か’가 종조사이다. ‘か’는 문말에 붙어서 의문을 나타낸다.

8) 체언에 준하는 조사이다. ‘本を読むのが好き。’(책을 읽는 것을 좋아해.)에서 ‘の’가 준체조사이다.

ㄴ.

勉強	**する**	。
benkyou	**suru**	PUNCT
VERB	**AUX**	
공부하다		.
VERB		PUNCT

ㄷ.

食べて	**いる**	。
tabete	**iru**	PUNCT
VERB	**AUX**	
먹고	**있다**	.
VERB	**AUX**	PUNCT

ㄹ.

食べ	**ない**	。
tabe	**nai**	PUNCT
VERB	**AUX**	
먹지	**않다**	.
VERB	**AUX**	PUNCT

2) 형용사(ADJ)와 한정사(DET)

영어에서는 형용사가 서술어로 쓰일 때는 be동사가 필요하지만 한국어나 일본어는 형용사가 단독으로 서술어로 쓰일 수 있으며 수식어로도 쓰인다. 일본어에서 ADJ는 '형용사-일반'과 '형태사-일반'이 포함되고 그 외에 '명사-보통명사-형태사 가능', '연체사'가 포함된다. 일본어의 경우에는 형용사를 형용사와 형태사로 나눌 수 있으므로 '형용사-일반'과 '형태사-일반'으로 나누어 제시하였다. 그리고 '명사-보통명사-형태사 가능'은 명사이지만 형태사로 기능할 수 있는 것들로 (4ㄹ)과 같이 ADJ로 주석한다. 일본어에서 형태사 '健康だ(건강하다)'는 명사를 수식할 때 'だ'가 'な'로 바뀌어 '健康な(건강한) + 명사'의 형태가 된다. 이런 경우에 '健康(건강)'은 명사의 형태이지만 ADJ로 주석하고 'な'를 AUX로 주석한다.

마지막으로 일본어의 연체사는 한국어에서는 관형사에 가까운데 한국어는 관형사 중에서 성상 관형사만이 ADJ에 포함되고 지시 관형사는 DET, 수관형사는 NUM에 포함된다. 일본어에서도 'この(이)', 'その(그)', 'あんな(저런)', 'どんな(어떤)' 등 '한정'의 의미를 나타내는 연체사는 DET로 분류한다.

(4) ㄱ.

あかい	りんご
akai	ringo
ADJ	NOUN
빨간	사과
ADJ	NOUN

ㄴ.

大きな	かばん
ookina	kaban
ADJ	NOUN
큰	가방
ADJ	NOUN

ㄷ.

同じ	会社
ookina	kaisya
ADJ	NOUN
동	회사
ADJ	NOUN

ㄹ.

健康	な	人
kenkou	na	hito
ADJ	AUX	NOUN
건강한		사람
ADJ		NOUN

3) 동사(VERB)

일본어의 경우 ‘명사+する(사변동사)’ 형태를 동사로 처리하고 있다. 예를 들면 ‘食事する(식사하다)’의 경우에는 ‘食事(식사)’에 ‘する’(하다)가 붙어서 동사 역할을 하기 때문에 ‘食事(식사)’가 명사 형태임에도 불구하고 VERB로 주석하고 ‘する’(하다)를 AUX로 주석한다. 반면에 한국어의 경우에는 ‘식사하다’ 전체를 VERB로 주석한다.

(5)

食事	**する**	。
syokuzi	**suru**	PUNCT
VERB	**AUX**	
식사하다		.
VERB		PUNCT

4) 부치사(ADP)와 불변화사(PART)

한국어와 일본어는 교착어이기 때문에 조사나 어미를 부착하여 여러 가지 문법 관계를 나타낸다. 그 중에 조사는 격을 표시하여 다른 단어와의 관계를 나타내거나 의미를 더하는 역할도 한다. 일본어는 한국어와 달리 띄어쓰기가 없어 단어 단위를 무엇으로 할 것인가에 대한 문제가 발생한다. 그러나 단어 분할이 없기 때문에 오히려 자유롭게 필요에 따라 단어를 분할할 수 있다는 장점이 있다. 그러므로 한 어절에 내용어와 기능어가 포함된 한국어와 달리 일본어는 조사나 어미, 접미사 등에도 별도의 태그를 부착하는 것이 가능하다. (6ㄴ)에서 'きれいですね(예쁘네요)'는 일본어에서는 '형용사+조동사+종조사'의 형태로 분석되지만 한국어에서는 어절을 기본 단위로 하기 때문에 '형용사'만으로 주석된다.

일본어는 다양한 종류의 조사를 가지고 있는데 ADP에는 격조사와 계조사가 포함되고 종조사는 PART에 포함된다. PART에서는 종조사 외에도 접미사 '的(적)'이 포함되어 있다. 한국어의 경우 어절 단위로 UPOS를 할당할 때 조사나 어미가 구두점이나 기호로 인해 분리되어 단독으로 어절을 구성하는 경우에만 ADP와 PART로 주석한다.

(6) ㄱ.

私	**は**	家	**に**	行く
watasi	**wa**	ie	**ni**	iku
NOUN	**ADP**	NOUN	**ADP**	VERB
나는		집에		간다
NOUN		NOUN		VERB

ㄴ.

きれい	です	**ね**	。
kirei	desu	**ne**	PUNCT
ADJ	AUX	**PART**	
예쁘네요			.
ADJ			PUNCT

5) 등위접속사(CONJ)와 종속접속사(SCONJ)

CONJ와 SCONJ도 영어와 같은 굴절어와 한국어나 일본어 같은 교착어가 많은 차이를 보이는 부분이다. 명확한 지침과 예문이 필요하지만 일본어의 경우 UD에 CONJ만 지침과 예문이 공개되지 않았다.

일본어에는 접속사가 독립적인 품사로 설정되어 있다. 그러므로 일본어

에서는 ‘접속사’[9]와 ‘접속조사’가 CONJ에 포함되며 한국어의 경우에는 접속부사와 접속조사가 CONJ에 포함된다.

SCONJ는 일본어의 경우 ‘접속사’와 ‘접속조사’ 그리고 ‘준체조사’가 포함된다. CONJ를 제외한 접속조사를 SCONJ에 포함시키고 있다. 한국어의 경우에는 ‘및’과 ‘또는’을 제외한 접속부사를 SCONJ에 대응시켰다. 이는 한국어의 접속부사가 형태만으로 종속 접속인지 대등 접속인지 구분하기 어렵기 때문이다.

(7) ㄱ.

食べ	て	寝る	。
tabe	**te**	neru	PUNCT
VERB	**SCONJ**	VERB	
먹고		자다	.
VERB		VERB	PUNCT

ㄴ.

食べる	の	が	好き	。
taberu	**no**	ga	suki	PUNCT
VERB	**SCONJ**	ADP	ADJ	
먹는 것이			좋아	.
VERB			ADJ	PUNCT

(7ㄱ)은 て(고, 어(여), 서)’는 한국어에서는 연결어미의 기능을 수행한다고 볼 수 있는데 이를 일본어에서는 ‘SCONJ’로 주석이 가능하다. (6ㄴ)은 준체조사의 예이다. 준체조사는 한국어로 ‘~는 것’으로 대역되는데 용언을 체언화하여 체언에 준하는 기능을 갖게 한다. 일본어에서는 준체조사를 ‘SCONJ’로 주석한다.

2. 의존관계 주석체계의 적용

본 절에서는 일본의 UD 적용 방안에 대해서 논의되고 있는 사안과 실제 UD 주석체계에 따라 주석된 공개 말뭉치에서 한국어와 일본어가 어떠한 차

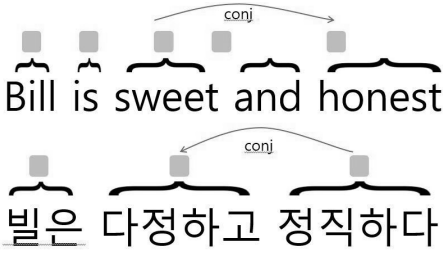
9) 일본어에서는 접속조사는 항상 앞말과 밀착되어 나타나지만 접속사는 독립하여 문절을 이룬다는 점에서 구별이 된다. 또한 접속사는 단어와 단어, 절과 절, 문과 문을 접속하는 기능을 한다면 모두 접속사로 분류하고 있다.

이를 보이는지 살펴본다. 먼저 한국어와 일본어가 일본어에 UD를 적용하는 방안에 대하여 大村舞 외(2018)에서 밝힌 세 가지 논의점은 다른 UD 적용 방안 연구에서도 공통적으로 제기하는 문제점이다. 첫 번째는 단어와 구, 절의 구분이다. UD는 단어, 구, 절을 나누도록 설계되어 있지만 일본어는 한국어와 같이 주어가 문장의 필수 요소가 아니다. 이에 따라 각 단위를 구분하는 기준이 모호해지므로 이를 결정하는 기준을 설정할 필요성이 있다. <표 4>는 田中貴秋(2018)에서 제시한 것으로 UD의 근간이 되는 Stanford Dendendencies(이하 SD) 일본어에 적용시킨 뒤 다시 한 번 UDv2에 맞춰 변환 작업을 거친 것으로 세분화된 구분에서 간략화를 지향하는 방향으로 변화하였다. 이러한 방향에서 가장 두드러지는 것은 기존에 구(*mod)로 주석되었던 사례가 절(acl)로 변환된 것이다. 한국어에 DEPREL를 적용할 때에도 이러한 방향성을 가져가야 할 필요가 있다. 언어적인 특성 상 구와 절의 구분이 명확히 되지 않을 경우, 주석을 어떻게 할당할 것인가에 대한 지침을 마련해야 한다. 주석체계 자체는 언어를 표현하는 수단이지만 효율적인 처리를 위하여 고안되었기 때문에 때로 언어학적으로 엄밀한 기준을 세우는 것보다 컴퓨터가 일관되게 처리할 수 있고, 주석 작업을 실제 담당하는 작업자들이 직관적으로 빠르게 주석 표지를 할당할 수 있는 기준을 세우는 것이 중요할 수 있기 때문이다.

〈표 4〉 일본어의 DEPREL 적용안 중 일부

분류		SD	UD_DEPREL
격관계	필수격	nsubj, dobj, iobj	nsubj, obj, iobj
	수의격	tmod, imod, arg	obl
절	관계절	rcmod_nsubj, rcmod_dobj, rcmod_iobj	acl
	보충절	ncmod	acl
	보족절	ccomp	ccomp
	부사절	advcl	advcl
내용어 간의 수식		amod, advmod, nmod, num	acl, advmod
기능어 관련		aux, pobj, post	aux, case
병렬, 동격		conj, appos	conj, appos

두 번째는 Hiroshi *et al.*(2018)에서 제시한 것과 같이 지배소 후위 언어의 특성에 기반하여 핵어(root)의 위치를 오른쪽에 두는 데서 발생하는 문제점이다. 특히 병렬, 대등 구조에서 이러한 문제가 드러나는데, SD의 기준이 되는 영어에서는 핵을 왼쪽에 두기 때문에 병렬관계에서 핵의 위치가 한국어나 일본어 같은 언어와 달라진다. 한국어와 일본어는 지배소인 핵어가 보통 오른쪽에 오기 때문에 대등, 병렬 요소를 주석하는 경우에도 지배소 후위 원칙에 따라서 오른쪽에 오는 요소에 root를 할당하기 때문이다. 이러한 혼란을 방지하기 위해서 이에 대한 기준 설정 역시 요구된다. [그림 1]은 UD 누리집에서 제시된 대등 접속 의존관계 표지인 conj의 영어 사례를 한국어로 번역하여, 지배소 후위 원칙에 따라 주석한 것을 보여준다. 의존관계 주석은 지배소와 의존소의 관계를 방향성을 지닌 그래프로 표현하는데, 이처럼 대등한 접속 관계를 표현할 때는 지배의존 관계를 나타내기 어렵다. 이에 따라 해당 언어의 보편적인 특질을 반영하여 핵어가 보통 선두에 오는 경우는 앞에 오는 요소를 핵어로, 핵어가 후위에 오는 언어의 경우는 뒤에 오는 요소를 핵어로 주석하는 것이 일반적이다. 그러나 UD의 경우는 다국어 범용 주석을 목적으로 하고 있어 이러한 기준 설정에 어려움이 따른다.



[그림 1] 영어와 한국어의 병렬 구조 주석의 예시

세 번째는 본고의 3장에서 별도로 논의한 것과 같이 DEPREL의 기본 주석 단위에 대한 문제이다. UDv1에서 공개된 말뭉치의 경우 기존에 문절을 기본으로 구문 관계를 주석한 말뭉치들을 UD로 변환하였기 때문에 문절을 기준으로 DEPREL을 주석하였다.[10] 大村舞(2017)에서 보인 대로 UDv2에서 새롭게 공개된 말뭉치는 이러한 흐름에서 벗어나 새로운 단어 단위를 설

정하고 있다. UD에서 요구하는 통사적 단어의 정의가 문절과 차이가 있다고 판단하여 단단위, 장단위, 문절의 조합을 이용해 UD의 단어 단위에 맞춰 변환하는 기준을 마련하려는 논의도 존재한다. BCCWJ의 경우 한국어의 형태소 분석 단위에 대응하는 단단위가 기본 DEPREL 주석 단위로 설정되어 있다. 이에 따라서 형태 분석 단위가 아닌 어절을 DEPREL 기본 주석 단위로 설정한 한국어 UD 말뭉치와 차이를 보인다. 이상과 같은 세 가지 논의점은 한국어에서도 충분히 숙고되어야 하는 문제점이다.

한국어와 일본어의 실제 주석 표지 적용 양상을 비교했을 때 두드러지는 차이를 보이는 주석 표지는 case, aux 및 간접 목적어와 관련된 주석 표지이다. case와 aux는 전체 말뭉치에서 해당 표지가 나타나는 상대 빈도를 기준으로 보았을 때, 가장 큰 차이를 보이는 표지이다. case는 한국어와 일본어에서 조사와 관련이 있고, aux는 한국어에서는 보조용언, 일본어에서는 조동사와 관련이 있는 표지이다. <표 5>는 UD 체계를 적용한 일본어 말뭉치인 UD-BCCWJ와 공개된 한국어 말뭉치 중 가장 큰 규모인 KAIST 말뭉치의 case와 aux의 빈도를 나타낸 것이다. 일본어에서는 case가 전체 표지의 19.85%로 압도적으로 고빈도를 차지하는 데 비해 한국어에서는 0.38%에 불과하다. aux 역시 일본어에서는 12.24%로 punct를 제외하면 두 번째로 고빈도인 데 비해 한국어에서는 5.41%에 불과하다. 이처럼 다양한 UD 의존관계 주석 표지 중에서 case 및 aux에서의 빈도 차이가 두드러지는 것은 앞서 언급한 것과 같이 한국어와 일본어가 지닌 교착어로서의 특성, 그리고 한국어와 일본어 간의 품사체계 및 문법형태소의 처리에 있어서 나타나는 차이에 기인하는 것으로 보인다. 한편 간접목적어와 관련되는 obl 및 iobj 역시 한국어와 일본어의 조사가 지닌 형태・기능적 특성과 깊은 관련이 있다. 즉 UD를 기준으로 했을 때 간접목적어에 상응하는 성분이 어떠한 표지에 의해 드러나는지, 그리고 이들 표지에 대하여 어떠한 주석을 부여해야 하는지에 대한 판단이 곧 obl 및 iobj의 처리 방안과 직결되는 것이다. 본 절에서는 각각의 적용 및 처리에 있어서 나타나는 차이점에 대해 살펴보도록 한다.

---

10) 일본어 구문 분석 말뭉치는 교토대학 텍스트 말뭉치와 일본어 의존관계 말뭉치, Kaede treebank 등이 있고 이러한 의존관계를 반영한 말뭉치 중 BCCWJ도 존재한다.

〈표 5〉 한국어와 일본어의 DEPREL 주석 표지 빈도 비교

한국어 태그	한국어 빈도	비율	일본어 태그	일본어 빈도	비율
aux	18,935	5.41%	aux	154,878	12.24%
case	1,343	0.38%	case	251,250	19.85%
전체	350,090	100%	전체	1,265,572	100%

1) 격표지(case)와 조동사(aux)

일본어는 UD의 DEPREL을 할당하기 위하여 격 표지(case-marking)와 술어-논항 구조(predicate-argument structure) 정보를 이용한다. 술어-논항 정보의 경우 기본적으로 의미 층위의 정보이기 때문에 주로 격 표지 정보를 이용하여 DEPREL를 부여한다고 밝히고 있다. 예를 들면 조사 'は(wa)'는 격 표지로 주어 관계(nsubj)를 나타내는 동시에 주제 표지가 된다. 이는 한국어 조사 중 '은/는'과 유사한 기능을 가지고 있다. 일본어에서는 단단위를 구문 분석의 기본 단위로 설정하기 때문에 조사가 독립적인 단위로 인정받아 관계 주석 표지를 할당받는다. 따라서 <표 5>에서 격 표지가 고빈도로 나타나는 것으로 풀이된다. 반면 한국어의 경우 조사는 자립 성분이 아니기 때문에 언제나 명사, 대명사, 수사와 같은 체언에 결합하여 실현된다. 기능적 측면에서만 본다면 조사는 선행하는 체언뿐만 아니라 선행 명사구 전체에 대하여 그 문법적 역할이 미치지만, 어절을 관계 주석의 기본 단위로 삼을 경우에는 조사가 독립적인 단위로 인정받지 못하고 선행 요소에 대하여 의존적인 형식으로 처리될 수밖에 없다. 조사가 독립적인 단위로 인정받아 관계 주석을 부여받는 것은 구두점 또는 기호 등으로 분리되어 단독 어절을 구성하는 예외적인 경우에만 가능하다. 이처럼 한국어와 일본어에서 조사가 지닌 형태적 지위나 역할은 흡사할 뿐만 아니라, 이들이 지닌 역할로 인하여 조사 자체는 고빈도로 출현하는 단위이다. 그러나 각 언어에 대한 주석체계 처리 방식에 따른 차이로 인하여 일본어에서는 case로 주석되는 경우가 고빈도로 나타나는 반면 한국어의 경우 매우 낮은 빈도를 보이게 된 것이다.

앞서 언급한 것처럼 aux의 경우 한국어에서는 보조용언에, 일본어에서는 조동사에 대응하는 주석이다. 이들은 모두 문장의 중심이 되는 용언의 의미를 보충해 주거나, 문장 전체에 대하여 일정한 의미를 더해 주는 역할을 한다는 공통점이 있다. 그런데 물론 앞서 살펴본 조사의 경우처럼 그 빈도 차

이가 현저하지는 않지만, aux의 경우에도 일본어에서는 적지 않은 빈도로 나타나는 반면 한국어에서는 보다 훨씬 적은 빈도로 출현한다는 점에서 그 원인을 살펴볼 필요가 있다. 결론부터 말하자면 이 역시 실제 언어 자료를 처리하는 데 있어서 발생하는 어절 단위 처리 문제와 깊은 관련이 있는 것으로 보인다. 우선 일본어에서 조동사는 UPOS로는 AUX에 할당되며 이들 대부분이 의존관계 주석체계에서는 aux에 대응된다. 이때 UPOS 차원에서 이미 별개의 단위로 처리되어 있기 때문에 의존관계 주석체계에서도 별도의 주석을 부여받는 것은 매우 자연스러운 현상이다. 또한 서술어의 의미를 보충해 주는 형식이 따로 존재하는 것은 교착어에서 흔히 나타나는 현상이므로 그 빈도 역시 높게 나타날 것으로 예상할 수 있다. 한국어의 보조용언 역시 그 기능은 일본어의 조동사와 유사하다. 즉 본용언에 후행하여 본용언이 지닌 어휘적 의미를 보충해 주거나 문법적 의미를 더해 주는 역할을 한다. 그런데 이때 실제 언어 자료상에서 한국어의 보조용언은 일본어의 조동사에 비해 그 독립성이 높지 않은 경우가 많다. 보조용언은 그 실질적 의미가 비교적 강하지 않기 때문에 붙여 쓰는 경우가 많고, 이를 반영하여 어문규정에서도 본용언과 보조용언을 띄어 쓰는 것을 원칙으로 하지만 붙여 쓰는 것도 허용하기 때문이다. 이로 인하여 만약에 '먹어 버렸다'와 같이 본용언과 보조용언을 띄어 쓸 경우 각각의 어절이 별개의 단위로 처리되어서 '버렸다'에 aux가 부여될 것이지만, '먹어버렸다'와 같이 하나의 어절로 쓴다면 해당 어절 전체가 서술어로서 처리될 수밖에 없기 때문에 aux가 부여될 기회가 없는 것이다. 이뿐만 아니라 일본어에서 aux는 용언류에 속하는 서법조동사(modal verb)뿐만 아니라 동사에 문법적 의미를 더하는 요소들까지 포함하는데, 이때 후자의 대부분의 한국어에서 보조용언이 아닌 어미에 해당한다. 즉 동일하거나 유사한 의미를 나타내는 문장이 있다고 하더라도 일본어에서는 조동사로 표현될 것이 한국어에서는 어미로 표현됨에 따라 aux에 대응되는 형식의 수 자체가 적게 나타나는 경향이 있는 것이다.

이처럼 보조용언에만 aux를 할당하고 그 보조용언들도 붙여 쓰는 경우가 많은 한국어의 특성을 고려했을 때, 일본어에서의 aux 주석의 비율이 상대적으로 높은 것은 자연스러운 결과라고 할 수 있다. 물론 이러한 결과는 단순히 한국어와 일본어 간의 언어적 특성에 따른 차이로 처리할 수도 있겠지만, 결과적으로는 문장 내에서 일정한 기능을 하는 단위들이 일본어에서는

주석 표지로서 잘 드러나지만 한국어에서는 그렇지 못하다는 한계가 있다고 해석할 수도 있다. 실제 언어 현상을 비슷하지만 주석 단위의 빈도에 있어서 차이가 난다는 것은 곧 해당 언어에서의 의존관계를 여실히 드러내 주지는 못하는 결과로 이어질 수 있기 때문이다. 이러한 문제는 결국 한국어에서의 어절 또는 띄어쓰기 처리에 기인하는 것으로 보이는데, 이를 극복하기 위해서는 단순히 어절을 기본 단위로 하여 의존관계 주석을 부여할 것이 아니라 XPOS를 기반으로 한 형태 분석 단위를 기준으로 의존관계를 주석하는 방안을 고려해 볼 수 있을 것이다.

2) 간접목적어(iobj)와 부사어(advmod)

case와 aux 외에 한국어와 일본어의 처리 방식이 확연히 다른 관계 주석 표지로 간접목적어를 주석할 수 있는 표지인 iobj가 있다. 사실 간접목적어는 ‘She gave me a book.’과 같이 영어 등의 언어에서 두 개의 목적어 논항이 필수적으로 출현하고 이들이 구조상의 배열에 의해 구분되는 경우 이들을 구분하기 위한 성분이다. 즉 위의 문장에서 ‘me’가 간접목적어, ‘a book’이 직접목적어가 되는 것이다. 이러한 현상에 대하여 적절한 의존관계 주석을 부여하기 위해서 UD에서는 iobj라는 주석을 설정하고, 간접목적어에 부여하도록 지침을 제시하였다. 그런데 문제는 한국어와 일본어의 경우, 목적어 논항이 목적격 조사에 의해 구분되며, 특수한 경우를 제외하고는 한 문장에서 필수적인 목적어 논항이 두 개 이상 나타나는 경우는 거의 없다는 것이다. 대신 한국어와 일본어에서는 다양한 조사들이 결합하여 문장 내에서 나름대로의 관계와 역할을 나타내 주는 기능을 한다. 한국어의 예로 들면 영어의 간접목적어에 해당하는 성분은 ‘그녀가 나에게/한테 책 한 권을 주었다.’와 같이 ‘에게’, ‘한테’ 등의 부사격 조사가 결합한 명사구로 실현된다. 그런데 ‘에게’, ‘한테’는 소위 간접목적어 표지로도 쓰이지만, ‘일정하게 제한된 범위’(‘나에게 돈이 좀 있다.’)나 ‘행위의 도달점’(‘그녀는 나에게 살며시 다가왔다.’), ‘행위를 가하는 주체’(‘그는 나에게 미움을 샀다.’), ‘비교의 대상’(‘나는 동생에게 지기 싫다.’) 등 매우 다양한 의미를 나타낼 수 있다. 이러한 사정은 일본어의 경우에도 크게 다르지 않다. 이와 같은 언어적 특성을 고려했을 때, 한국어에서는 iobj만을 변별해 내기 어려울 뿐만 아니라 사격 논항에 부여되는 주석인 obl과의 혼동을 초래한다는 문제가 있다. 이로 인해

한국어와 일본어의 경우 동일한 성분에 대하여 사격 논항 주석인 obl과 간접 목적어 논항인 iobj 중 어느 것을 부여해야 하는가의 문제가 발생한다.

UD 프로젝트 누리집의 정보를 토대로 하여 한국어 UD 말뭉치별로 직접 목적어와 간접 목적어의 UPOS 결합 분포와 빈도를 살펴보면 전체적으로 간접목적어를 인정하지 않는 경향을 파악할 수 있다.

〈표 6〉 한국어의 목적어 표지 유형별 UPOS 결합 분포와 빈도

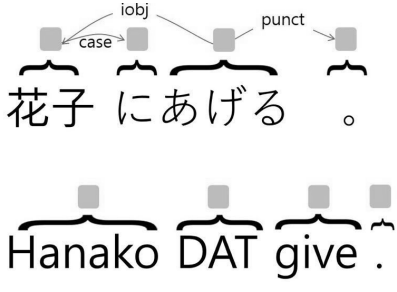
UD_Korean-GSD		UD_Korean-Kaist		UD_Korean-PUD	
obj		obj		obj	
VERB-NOUN	5062	VERB-NOUN	13411	VERB-NOUN	30
VERB-NOUN-ADP(는)	1	VERB-NOUN-ADP(과)	2	VERB-NOUN-Acc	333
VERB-NOUN-ADP(도)	1	VERB-NOUN-ADP(나)	1	VERB-PRON	1
VERB-NOUN-ADP(마저+를)	1	VERB-NOUN-ADP(들)	1	VERB-PRON-Acc	10
VERB-NOUN-ADP(만)	1	VERB-NOUN-ADP(를)	29	VERB-Fin-NOUN	19
VERB-NOUN-ADP(어+재연)	1	VERB-NOUN-ADP(만+을)	1	VERB-Fin-NOUN-Acc	108
VERB-NOUN-ADP(을)	1	VERB-NOUN-ADP(와)	2	VERB-Fin-PRON	1
VERB-NOUN-ADP(의)	1	VERB-NOUN-ADP(을)	56	VERB-Fin-PRON-Acc	2
VERB-NOUN-ADP(정도+를)	1	VERB-PRON	535	VERB-Ger-NOUN	2
VERB-PRON	65	VERB-PRON-ADP(를)	3	VERB-Ger-NOUN-Acc	20
		VERB-PRON-ADP(을)	1		
iobj		iobj		iobj	
VERB-NOUN	94	VERB-NOUN	1	VERB--NOUN	1
VERB-PRON	3	VERB-NOUN-ADP(들+에게)	1	VERB-NOUN-Advb	2
		VERB-NOUN-ADP(에게)	1	VERB-Fin-NOUN-Advb	2
				VERB-Fin-PRON-Advb	2

간혹 체언과 조사 ‘에게’가 결합한 경우에 대해 간접목적어로 관계 주석 표지를 부여한 예가 있는 것으로 보이나 대부분의 경우 아래 예와 같이 UPOS 자리에 ADV 표지를 할당하고 XPOS 필드에서 ‘에게’에 부사격 조사 표지 JKB를 할당하면서 관계 주석은 advmod 표지로 처리하고 있다. 이는 조사 ‘에게’가 의미적으로 영어의 간접목적어 역할을 하기도 하지만 선행 체언과 결합하여 일정하게 제한된 범위, 행위의 도달점, 행위를 가하는 주체, 비교의 대상 등 다양한 의미적 기능을 수행하기 때문이다.

〈표 7〉한국어의 advmod 표지 할당 예시

1	1969년	1969+년	NOUN	SN_NNB	3	nmod
2	경영에서	경영+에서	ADV	NNG+JKB	3	advmod
3	은퇴하며	은퇴+하+며	VERB	NNG+XSV+EC	7	advcl
4	전문경영인에게	전문+경영인+에게	ADV	NNG+NNG+JKB	7	advmod
5	유한양행의	유한양행+의	NOUN	NNP+JKG	6	det:poss
6	경영권을	경영+권+을	NOUN	NNG+XSN+JKO	7	obj
7	인계하였다	인계+하+았+다	VERB	NNG+XSV+EP+EF	0	root
8			PUNCT	SF	7	punct

이에 비해 일본어는 아래와 같이 UD에서 공개한 일본어 지침에서 간접목적어에 대응하는 격표지를 'に'로 제시하고 있다. 그러나 이에 대해서는 이론의 여지가 있다. 우선 일본어의 가장 보편적인 품사 주석체계인 IPA 품사체계에서 'に'를 부사화 조사로 분류하고 있으며 한국어의 '에게'와 유사하게 다양한 의미를 가지기 때문이다. 최근 발표된 Omura and Asahara(2018)에서도 일본 국립국어연구소의 BCCWJ-DepPara 말뭉치가 UD의 관계 주석 표지를 모두 반영하고 있지는 않다고 밝히고 있다. 한국어에 UD 주석체계를 적용하는 경우에도 UD의 주석체계를 전부 활용하는 것보다, 한국어의 특성을 반영하여 통용할 수 있는 주석 표지를 선택하고 그 표지를 어떤 기준으로 할당할 것인가에 대한 원칙을 정하는 것이 중요할 것이다. 이러한 문제는 특히 간접목적어와 부사어의 경계에 있는 성분을 주석하는 부분에서 먼저 고려되어야 할 것으로 보인다.



[그림 2] 일본어의 iobj 표지 할당 예시

Ⅴ. 맺음말

이 논문에서는 일본어 말뭉치의 UD 체계 적용 사례를 검토하여, 한국어 UD 말뭉치 구축 시 논의해 보아야 하는 내용을 살펴보았다. 우선 주석의 기본 단위 설정에 대한 문제를 확인하고 UD를 구성하는 UPOS와 DEPREL의 적용과 그에 따른 논의들을 검토하였다. UPOS의 경우 AUX, ADJ, VERB처럼 서술어와 관련된 범주의 표지를 할당하는 문제와 조사, 어미와 같은 기능어의 처리 방안을 살펴보았으며 접속사와 이와 관련된 단위를 어떻게 처리하고 있는지 검토하였다. DEPREL 또한, 구문 태그를 주석하는 기본 단위의 문제에서부터 통사적 문제를 어떻게 반영할 것인지에 대한 논의 사항을 살펴보았다. 한국어와 일본어가 지배소 후위 언어인데서 비롯하는 root 설정의 문제와 이어지는 병렬 구조의 주석 문제, 주석 기본 단위 설정의 문제에서 비롯되는 case와 aux, iobj 표지의 문제를 검토하였다. 상기의 논의들은 추후 한국어 UD 말뭉치 구축하거나 개선하기 위한 기준을 마련할 때 필수적으로 검토해야 하는 과제가 될 것이다.

UD의 DEPREL 주석 체계는 지속적인 연구와 워크숍을 통해서 진화하고 있으며 이에 발맞추어 UD 기반 의존 구문 주석 말뭉치 역시 종류와 버전이 다양해지고 있다. 역동적인 체계를 지닌 영어 중심 UD를 기반으로 영어와 언어 특성이 크게 차이가 나는 한국어와 일본어의 UD 주석 체계를 갱신하고 언어 자원을 구축하는 것은 지난한 작업이다. 한국어와 일본어의 UD 말뭉치가 여러 버전이지만 단위부터 주석 체계의 적용까지 일관성과 호환성이 떨어진다는 것이 그 방증이다. 그러나 기본 문법이 유사한 한국어와 일본어의 UD 주석 체계 관련 협력 연구 및 연구자 교류가 활발한 만큼 앞으로의 전망은 밝다고 볼 수 있다. 이 연구가 한국어와 일본어의 UD 주석 체계 논의 활성화에 기여하기를 기대한다.

참고문헌

문화체육관광부, 「21세기 세종계획 국어 기초 자료 구축 보고서」, 2006.

리우완잉, 「유형론적 관점에서 다시 본 한국어 관형사의 품사 처리 문제-중・일 양 언어와의 비교를 통하여」, 『이중언어학』 66, 2017.

박혜진・오태환・김한샘, 「Universal POS 태그셋의 한국어 적용」, 『제30회 한글 및 한국어 정보처리 학술대회(HCLT) 논문집』, 2018a.

박혜진・오태환・김한샘, 「Universal Dependency를 위한 한국어 형태 주석 체계연구」, 『언어와 정보』, 22-3, 2018b.

송경안 외, 「세계 주요 9개 언어의 유형론적 비교연구」, 한국연구재단 연구 보고서, 2007.

오진영・차정원, 「키어절을 이용한 새로운 한국어 구문분석」, 『정보과학논문지: 소프트웨어 및 응용』, 40-10, 2013.

임준호・배용진・김현기・김윤정・이규철, 「의존 구문분석을 위한 한국어 의존관계 가이드라인 및 엑소브레인 언어분석 말뭉치」, 『제27회 한글 및 한국어 정보처리 학술대회(HCLT) 논문집』, 2015.

최형용・유완영, 「한・중・일 품사 대조를 위한 품사 분류 기준 설정」, 『어문연구』 43-2, 2015.

Berdicevskins, A., Using Universal Dependencies in cross-linguistic complexity research, *EMNLP-UDW18*, 2018.

Choi, J. D., and Palmer, M., Statistical dependency parsing in Korean: From corpus generation to automatic parsing, *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, 2011.

Choi, K., Kawahara, D., and Kurohashi, S., Towards fully lexicalized dependency parsing for Korean, *The 13th International Conference on Parsing Technologies*, 2013.

Chun, J., Han, N., Hwang, J. D. and Choi, D., Building Universal Dependency Treebanks in Korean, *LREC* 2018, 2018.

Hiroshi, K., Han, N., Asahara, M., Hwang, J. D., Miyaho, Y., Choi, J. D.

and Matsumoto, Y., Coordinate Structures in Universal Dependencies for Head-final Languages, *EMNLP-UDW18*, 2018.

Omura, M. and Asahara, M., UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese, *EMNLP-UDW18*, 2018.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Castelló, N. B. and Lee, J., Universal dependency annotation for multilingual parsing, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 2, 2013.

Park, J., Segmentation Granularity in Dependency Representations for Korean, *Proceedings of the Fourth International Conference on Dependency Linguistics*, 2017.

Park, J., Hong, J. P., and Cha, J. W. Korean Language Resources for Everyone. *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation,* 2016.

金山博, 宮尾祐介, 田中貴秋, 森信介, 浅原正幸, 植松すみ れ. 日本語 Universal Dependencies の試案. 言語処理 学会第 21 回年次大会発表論文集, 2015.

大村舞, 浅原正幸, UD Japanese-BCCWJ の構築と分析, 言語資源活用ワークショップ2018発表論文集, 2018a.

大村舞, 浅原正幸. UD Japanese BCCWJ: 現代日本語書き言葉均衡コーパスのUniversal Dependencies. 言語処理 学会第 24 回年次大会発表論文集, 2018b.

大村舞, 浅原正幸. 現代日本語書き言葉均衡コーパスの universal dependencies. 言語資源活用ワークショップ 2017, 2017.

田中貴秋, UD Japanese-KTC: 京大コーパス句構造版からの Universal Dependencies化, 第1回 Universal Dependencies 公開研究会, 2018 (http://pj.ninjal.ac.jp/corpus_center/pdf/2018-06-16-tanaka.pdf).

UD annotation guidelines(http://universaldependencies.org/u/overview/tokenization.html.), 2018. 9.

ABSTRACT

A Case Study on Universal Dependency Tagsets

Han, Jiyoon
Lee, Jin
Lee, Chanyoung
Kim, Hansaem

The purpose of this paper was to examine universal dependency UD application cases of Korean and Japanese with similar morphological characteristics. In addition, UD application and improvement methods of Korean were examined through comparative analysis. Korean and Japanese are very well developed due to their agglutinative characteristics. Therefore, there are many difficulties to apply UD which is built around English refraction. We examined the application of UPOS and DEPREL as components of UD with discussions. In UPOS, we looked at category problem related to narrative such as AUX, ADJ, and VERB, We examined how to handle units. In relation to the DEPREL annotation system, we discussed how to reflect syntactic problem from the basic unit annotation of syntax tags. We investigated problems of case and aux arising from the problem of setting dominant position from Korean and Japanese as the dominant language. We also investigated problems of annotation of parallel structure and setting of annotation basic unit. Among various relation annotation tags, case and aux are discussed because they show the most noticeable difference in distribution when comparing annotation tag application patterns with Korean. The case is related to both Korean and Japanese surveys. Aux is a secondary verb in Korean and an auxiliary verb in Japanese. As a result of examining specific annotation patterns, it was found that Japanese aux not only assigned auxiliary clauses, but also auxiliary elements to add the grammatical meaning to the verb and form corresponding to the end of Korean. In UD annotation of Japanese, the basic unit of morphological

analysis is defined as a unit of basic syntactic annotation in Japanese UD annotation. Thus, when using information, it is necessary to consider how to use morphological analysis unit as information of dependency annotation in Korean.

Key Words : Universal Dependency, Dependency Relations, POS tagging, Korean language, Japanese language

▌논문접수일 : 2018. 11. 10
▌심사완료일 : 2018. 11. 30
▌게재확정일 : 2018. 12. 12