# 구어 의존 구문 분석을 위한 비유창성 처리 연구

박석원○, 최현수, 한지윤, 오태환1, 안의정2, 김한샘†

연세대학교 언어정보학협동과정, 국어국문학과1, 언어정보연구원2†
{ pswon27, choehyonsu, clinamen35, ghksl06041, khss†}@yonsei.ac.kr, snoopyjinu@gmail.com2

# A Study of Disfluency Processing for Dependency Parsing of Spoken

Seokwon Park○, Hyonsu Choe, Jiyoon Han, Taehwan Oh1, Euijeong Ahn2, Hansaem Kim†
Institute of Language and Information Studies, Yonsei University

**요 약**

비유창성(disfluency)은 문어와 같이 정연한 구조로 말하지 못하는 현상 전반을 지칭한다. 이는 구어에서 보편적으로 발생하는 현상으로 구어 의존 구문 분석의 난이도를 상향시키는 요인이다. 본 연구에서는 비유창성 요소 유형을 담화 표지, 수정 표현, 반복 표현, 삽입 표현으로 분류하였다. 또한 유형별 비유창성 요소를 실제 말뭉치에서 어떻게 구문 주석할 것인지를 제안한다. 이와 같은 구어 데이터 처리 방식은 대화시스템 등 구어를 처리해야 하는 도메인에서의 자연언어이해 성능 향상에 기여할 것이다.

주제어: Dependency parsing, disfluency, 구어 말뭉치, 의존 구문 분석, 비유창성

## 1. 서론

본 연구는 세종 현대 구어 말뭉치에서 나타난 비유창성을 살펴보고 이를 의존 구문 분석에서 처리하는 방안을 제시한다. 문어 말뭉치와 다르게 구어 말뭉치에는 다양한 비유창성 (disfluency)이 나타난다. '비유창성'(disfluency)이란 문어와 같이 정연한 구조로 말하지 못하는 현상 전반을 지칭하는 용어이다.[1] 이러한 비유창성이 구어 구문 분석에 어떠한 영향을 주고 이를 어떻게 처리하는 것이 효율적인지 밝히는 것이 본 연구의 목적이다.

현재 자연어처리 분야에서 의존 구문 분석 연구는 문어 텍스트 위주로 진행되어 왔다. 문어 텍스트를 기반으로 학습한 구문 분석기는 QA시스템, 기계독해 등의 도메인에 적합하다. 하지만 AI스피커나 챗봇 등 대화 시스템을 기반으로 한 서비스에는 적용에 어려움이 따른다. 한국어의 구어는 자유로운 어순, 비유창성, 생략 등이 나타나기 때문이다. 이에 본 연구는 구어 구문 파서 학습을 위한 구어 말뭉치 구축에 앞서 비유창성 요소를 추출하는 방법과 비유창성을 포함한 구문 주석 방법을 제안한다. 특히 의존 구문 주석 방법은 한국전자통신연구원(ETRI)과 UD(Universal Dependencies)의 구문 분석 방법론에 적용할 수 있는 비유창성 처리 방안을 제안한다. 이를 통해 의존 구문 분석 연구뿐만 아니라 구문 분석 결과를 이용한 다양한 연구에 도움이 되고자 한다.

2장에서는 비유창성을 고려한 파서, 말뭉치, 주석체계 등의 관련 연구를 설명한다. 3장에서는 세종 현대 구어 말뭉치에서 나타난 비유창성 요소를 유형화하고 추출 방법을 제시한다. 4장에서는 비유창성 요소 유형별로 구문 주석 하는 방법을 제시하고 비유창성 요소의 구문 주석에서 어려운 점을 설명한다.

## 2. 관련 연구

일반적인 파서 모델 연구가 아닌 구어 특성의 비유창성을 고려한 파서 연구들이 있다. Honnibal et al.(2014)는 비유창성 탐색기를 조합한 파서 모델을 연구했다. 트랜지션 기반의 모델로 발화 수정을 처리할 수 있다. 학습 데이터는 비유창성이 주석된 Switchboard 코퍼스의 구 구조 구문 분석을 의존 구문 분석 코퍼스로 변환하여 사용했다.[2] Paria J. L. et al.(2018)는 자동 비유창성 탐색을 위한 모델을 개발했다. 모델은 auto-correlational nerural newtwork(ACNN) 기반이다. 이 모델은 특히 구어에서 발화를 수정하려는 현상을 쉽게 포착할 수 있다는 것이 특징을 가진다.[3]

해외 말뭉치 중에서 비유창성을 고려한 구어 구문 분석 말뭉치는 LDC에 공개된 Switchboard가 있다.(Godfrey et al., 1992)[4] 비유창성 주석 가이드(Meeter et al., 1995)[5]와 Penn Treebank(Marcus et al.1993)[6] 가이드에 따라 구문 분석이 주석됐다. Rehbein,I. et al.(2014)는 독일어의 구어 말뭉치 KiDKo를 기획한 연구를 진행했다. 비유창성 현상들(disfluencies)에 해당하는 어절 또는 구를 가상의 root에 연결할지 주절에 연결할지에 대한 논의를 진행했다. 비유창성에는 반복, 주저, 자기수정, 휴지 등이 있다.[7]

국내에서 비유창성 현상 주석 체계는 남길임(2011)이 있다. LDC의 MDE 연구, MATE의 Mate Telematics 프로젝트, W. DuBois의 전사 계층을 비교 분석하여 한국어에 맞는 비유창성 현상 주석 체계를 설정했다. 비유창성 현상 주석 체계의 대분류는 삽입, 대치, 도치로 선정하였다. 세종 현대구어 말뭉치에서 공적 독백_강의와 국립국어원의 지역어 구술 담화 말뭉치를 대상으로 비유창성 현상 빈도를 추출했다.[1]

## 3. 구어 말뭉치의 비유창성 요소 유형

연구 대상 말뭉치는 세종 현대 구어 형태분석 말뭉치에서 대분류 '일상대화'에 해당하는 말뭉치이다. '일상대화' 관련 말뭉치 파일은 45개로 전체 말뭉치 중에서 21%에 해당한다. 총 문장은 51,626문장(213,241어절)이고 본 연구는 이 중에서 임의 추출한 2,000문장(7,279어절)을 대상으로 한다. 일상대화에 한정하는 이유는 첫째, 독백은 대화 참여자간의 상호성이 떨어지며 둘째, 강연 등의 말뭉치는 순구어가 아닌 준구어에 해당하기 때문이다.

연구 대상에서 비유창성을 드러내는 요소를 분석하여 유형화하고 추출 방법을 제시한다. 비유창성을 추출하여 명제적 정보를 가진 시퀀스만으로 의존 구문 분석을 하기 위함이다.(그림1) 비유창성 주석 체계가 이미 존재하지만 본 연구는 언어학적 세분류보다는 자동 처리를 염두에 두고 있기 때문에 담화표지, 수정 표현, 반복 표현, 삽입 표현의 4종류로 단순화했다.(표1) 문장의 기본 단위는 억양 단위 <s></s> 기준이다.

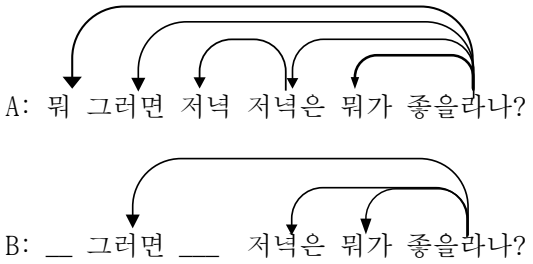



그림1 비유창성 요소 추출

| 유형 | 설명 | 빈도 |
|---|---|---|
| 담화 표지 (discourse marker) | 기존 품사의 의미, 기능을 가지지 않는 화용적 요소 | 105어절 (약 1.5%) |
| 수정 표현 (revision) | 이전 발화 수정 | 36어절 (약 0.4%) |
| 반복 표현 (repetition) | 동일 발화 반복 | 88어절 (약 1.2%) |
| 삽입 표현 (insertion) | 발화 중간 삽입 | - |

표1 비유창성 요소 유형 및 빈도

### 3.1 담화 표지(discourse marker)

담화 표지란 화용적 층위에서 여러 가지 기능(텍스트적 기능과 상호작용적 기능)을 수행하는 요소로, 구어에서 나타나며 문장의 명제적 의미에 영향을 미치지 않는 요소이다.[8] 세종 현대 국어 구어 말뭉치는 명제적 의미를 갖지 못하는 담화 표지가 주석되어 있어 이를 따로 추출하는 것이 가능하다. 추출해보니 105어절로 전체 7,279어절에서 약 1.5%에 해당한다.

"이, 그, 저, 아, 어" 등 동일한 형태로 기존 품사의 의미, 기능을 가지지 않는 것은 담화 표지로 보고 물결표(~)를 이용하여 표시하였다. 주로 머뭇거림의 이~, 그~, 저~, 어~, 아~ 등이 해당된다.[9]

예시 1

ㄱ. **그~** 살아 온 과정이,
ㄴ. **저~** 뭐야 밥 먹으러 갈 거니?
ㄷ. **아~** 이거 녹음을 해서 오빠 뭘 하는 건데?
ㄹ. **어~** 생각에는 이제,
ㅁ. 근데 그게 **뭐~** 말처럼 되냐?
ㅂ. 신규는 아직 **뭐~ 그~** 어떤 성과는 없어.

담화 표지는 위의 예시 'ㅁ, ㅂ'과 같이 담화 안에서의 위치가 자유롭다는 특성을 가지고 있다. 여기서 주의할 점은 표층 분석으로 담화표지 물결표(~) 주석으로는 실제 분석에서 중의성을 가진다. 이와 관련한 중의성 문제는 4장에서 다루도록 한다.

### 3.2 수정 표현(revision)

문어와 달리 구어에서는 불완전한 발화나 오류가 나타나고 이를 수정하려는 현상이 나타난다. 세종 현대 구어 말뭉치에는 불완전한 발화를 <trunc>태그로 주석해 놓아 이런 수정 어구를 자동적으로 추출할 수 있다. 임의 추출한 2,000문장(7,279어절)에서 <trunc>태그가 36개 추출되었고 약 0.4%에 해당한다.

예시 2

ㄱ. 빨리 타면 일곱시 **<trunc>사십</trunc>** 사십분 차를 타고 늦게 타면
ㄴ. **<trunc>어려</trunc>** 어려운 말 하네.
ㄷ. 근데 여름에 할려면 **<trunc>덥겠</trunc>** 덥겠다.
ㄹ. 사람이 좀 **<trunc>내리</trunc>** 좀 내리드라.

불완전한 발화는 <trunc>를 통해 추출 가능하고 완전한 발화는 이어서 나타난다는 것을 알 수 있다. 발화의 명제적 의미는 <trunc></trunc>로 주석된 내용을 제외하고 파악해야 한다.

### 3.3 반복 표현(repetition)

똑같은 시퀀스로 어절이 반복되는 표현이다. 수정 표현은 불완전한 발화 시퀀스와 완전한 시퀀스가 다르게 나타나는 것이고 반복 표현은 동일 시퀀스가 나타나는 것이다. 반복 현상은 두 어절이 인접해서 발생하는 경우(예시 3 ㄴ의 '너', ㄹ의 '우리', ㅁ의 '거기에 대해서')와 인접하지 않은 경우가 있다. 인접해서 반복하는 경우와 인접하지 않은 경우들 모두 후행 어절을 삭제하는 것이 합리적이다.

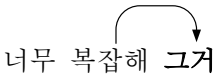
예시 3

ㄱ. **너** 감기 옮는다 **너,**

ㄴ. **너 너(인접 반복) 오씨엔** 보냐 **오씨엔?**
ㄷ. **준비를** 안해 갖고 왔어, **준비를.**
ㄹ. 아니 **우리** 샘플을 가져와 **우리 우리(인접 반복).**
ㅁ. **거기에 대해서** 자기 생각을 **거기에 대해서**
ㅂ. 그러면 니네들 답사 갈지 안 갈지 모르는 거야?

반복 현상은 세종 구어 말뭉치 지침에 다른 주석이 없기 때문에 억양 단위 <s></s> 내에서 동일 시퀀스가 2회 이상 발생한 어절로 자동 추출할 수 있다. 다만, 예시 'ㅂ' 처럼 반복 현상이 아니라 아닌 경우는 검수를 통해 제외해야 한다. 자동 추출한 115문장에서 해당되지 않은 경우를 검수하여 제외하니 반복 현상은 88어절이 추출됐다.

예시 4




인접하지 않는 반복 현상은 도치 현상(예시4)과 다르다. 도치는 일반적으로 서술어 앞에 위치하는 논항이 뒤에서 나타나는 현상이다. 즉, 도치 대상 어절은 문장에서 1회만 나타나고 반복 대상 어절은 2회 이상 나타나기에 구분하는 조건 추출이 가능하다.
반복 현상은 해당 어절들이 일부 명제적 정보를 가지기에 전처리를 위한 추출이 불필요한 것으로 보일 수 있다. 하지만 정보 추출(Information Extraction)에서 구문 분석 결과를 이용한 트리플 (Subject-Predicate-Object)을 추출한다면 트리플 내의 Subject 또는 Object가 중복되어 잉여적인 정보가 된다.

### 3.4 삽입 표현(insertion)

삽입 표현은 일부 명제적 의미를 가진 내용 삽입에 한정한다. 삽입은 구 또는 절이 될 수 있다. 남길임(2011)은 삽입의 하위 부류로 담화표지를 선정했는데 본 연구에서는 담화표지를 대분류로 선정하였다. 담화 표지가 고빈도로 나타나고 삽입 요소의 자동 처리가 어렵기 때문에 분류를 달리했다.

예시 5 나는 어제 그녀를 (**같은 동네 사는데**) 만났다.

내용 삽입은 원래 발화와 연속선상에서 분석될 수 없는 것이다. 문장 내에 포함된 어떤 성분에 대해 부연 설명하거나 다른 것을 덧붙이거나 하는 부분으로 담화 전체의 명제적 의미에 일정한 기여를 한다.[1]
세종 현대 구어 말뭉치에 관련 태그가 없고 실제 의미 분석을 고려해야 하기 때문에 자동 추출이 어렵다. 하지만 구문 분석 결과를 이용하여 의미 분석을 한다면 삽입 요소에 대한 구문 주석 방법이 필요하다.

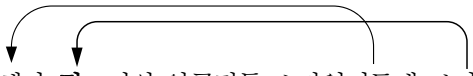
## 4. 비유창성 요소의 구문 주석 지침 제안

비유창성 요소 전처리를 거치지 않고 구어 구문 분석을 진행할 경우 비유창성 요소 유형에 따라 발생하는 문제를 분석하고 그에 따른 해결책을 제안한다. 비유창성 요소 유형별로 ETRI와 UD(Universal Dependencies) 기준의 의존 구문 분석 태그 주석과 의존 관계 지배소 설정 방법을 제시한다.
기존의 구문 주석 체계에서 구어에만 해당하는 주석 표지는 많지 않다. ETRI 주석 체계에서는 품사 태그 중 IP(감탄사구)를 활용할 수 있으며 의존 관계 태그 중에는 활용할 수 있는 태그가 없다.[10] UD에서는 품사 태그는 INTJ를 활용할 수 있고 의존 관계 태그 중에서 discourse(담화표지)와 reparundum(발화 수정) 태그를 활용할 수 있다.[11]

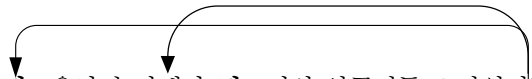
### 4.1 담화 표지의 구문 주석

담화 표지의 의존 관계 지배소 설정은 개별 담화표지가 독립적이라는 전제하에 인접하는 후행 어절을 지배소로 연결한다. 담화표지는 ETRI 기준으로 구문 태그 IP를 주석하고 기능 태그는 비워서 주석한다. UD 기준으로는 품사 태그 INTJ, 의존 관계 태그 discourse로 주석한다.
본 연구에서는 구문 주석의 기본 원칙인 투영성의 원칙(Projective Constraint)을 지키는 것을 전제로 하여 해결책을 제안한다. 만약 담화 표지의 위치가 자유로워서 의존 관계 설정 방법에 따라 교차 현상(non-projective)이 발생한다.
담화 표지가 명제적 정보가 없고 독립적 존재로 여겨질 수 있기에 어떤 지배소로 연결해도 무방해 보인다. 하지만, 담화표지의 지배소를 문장 전체의 head로 설정하면 아래와 같은 교차 현상이 일어난다. 이외에도 담화표지의 지배소를 근처 서술어에 연결하더라도 예시5의 ㄴ에서 볼 수 있듯이 교차 현상이 일어난다.

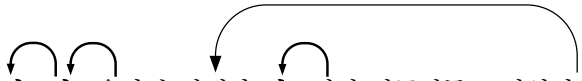
예시 6
ㄱ.문장 전체의 마지막 어절 연결




ㄴ. 인접 서술어 연결



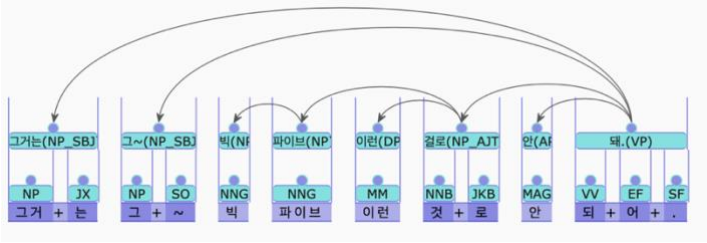

ㄷ. 후행 어절에 연결




이외에 담화 표지 주석에서 어려운 점은 담화 표지로 쓰이는 대상들의 중의성(disambiguation)을 가진다는 것이다. 담화 표지로 자주 쓰이는 '그'(대명사, 관형사), '뭐'(대명사, 감탄사), '좀'(부사, 명사) 등이 있다. 특히 구어에서는 문어에 비해 조사 생략이 빈번하여 중의성 해소가 보다 쉽지 않다.

예시6)에서는 '그~'는 담화 표지 분석이 아니라 NP_SBJ로 분석되었다. 특히 담화 표지의 중의성 해소가 되지 않으면 해당 어절 분석이 문장 전체 분석에 영향을 미친다. 이중 주어문으로 분석되어 구문 분석 기반의 의미역 분석에도 오류를 전파한다.

예시 6 ETRI 언어분석 API 분석 결과[12]




### 4.2 수정 표현의 구문 주석

수정 현상에서는 오류 어절의 지배소는 수정한 어절이다. 오류 어절의 품사 태그는 수정한 어절의 해당 품사 태그를 주석한다. ETRI 기준으로 구문 태그는 비워서 주석한다. UD 기준으로 의존 관계 태그는 reparundum(발화 수정)로 주석한다.

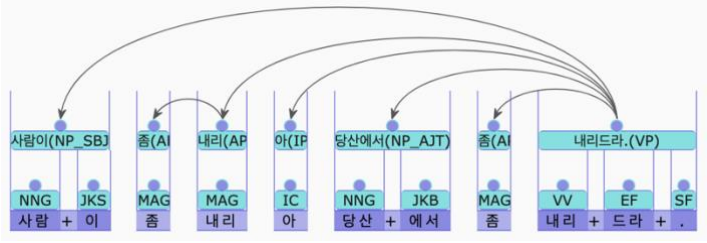
예시 7

ㄱ. 근데 여름에 할려면 **덥겠** 덥겠다.

ㄴ. 사람이 좀 **내리** 아 당산에서 좀 내리드라.

하지만 기존 파서로는 오류 어절과 수정 어절 구분이 쉽지 않을 뿐더러 오류 어절의 형태 분석이 중의성으로 다른 품사로 분석될 수 있다. 예시 8에서 '내리'는 '내리드라'의 불완전한 발화인데 부사 '내리'로 분석된 경우이다.

예시 8 ETRI 언어분석 API 분석 결과[12]

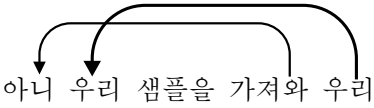



### 4.3 반복 표현의 구문 주석

반복 현상에서는 인접 반복 현상은 선행 어절의 지배소는 후행 어절이다. 후행 어절은 해당 품사 분석을 하고 실제 지배소를 결정한다. 인접 반복 현상에 한하여 품사 태그는 해당 어절의 품사로 주석한다. 의존 관계 태그 주석은 ETRI 기준으로 기능 태그를 비워서 주석하고 UD 기준으로 의존 관계 dep태그를 주석한다.

인접하지 않은 반복 현상은 반복하는 두 대상을 연결하지 않는다. 두 대상을 연결할 시에 예시 9와 같이 교차 현상이 일어나기 때문이다. 반복 대상을 연결하지 않는 대신에 표면상의 구문 분석을 진행한다.

예시 9




### 4.4 삽입 표현 구문 주석

삽입 구문 주석은 삽입된 구 또는 절의 마지막 어절에 따라서 품사 태그와 지배소를 결정한다. ETRI 분석 기준은 right arc를 허용하지 않기 때문에 삽입구의 마지막 어절의 지배소는 후행 어절이다. 하지만 UD 분석 기준은 right arc를 허용한다. 이에 삽입구가 선행 어절의 내용 부연일 경우에 삽입구의 마지막 어절의 지배소는 해당 선행 어절이다.

예시 10

ㄱ. ETRI 분석 방법

나는 어제 그녀를 **(같은 동네 사는데)** 만났다.

ㄴ. UD 분석 방법

나는 어제 그녀를 **(같은 동네 사는데)** 만났다.

## 5. 결론

본 논문에서는 비유창성 요소를 유형화하고 추출하는 방법과 구문 주석 지침을 제안했다. 전자는 비유창성 요소를 세종 구어 말뭉치 태그를 이용하거나 일부 조건을 제안하여 추출하는 방안이다. 추출을 통해 전처리를 한다면 명제적 정보만 남겨 구문 분석을 진행할 수 있다. 후자는 구문 분석 층위에서 비유창성을 구문 주석하는 방안이다. 구문 주석 지침은 ETRI와 UD의 의존 구문 분석 방법론에서 적용할 수 있도록 제안했다. 각 방법론의 품사 태그셋과 의존 관계 태그셋에서 비유창성 요소에 해당하는 태그로 설정하고 각 의존 관계 기본 원칙에 근거하여 의존 관계 지배소 설정 방법을 제안했다. 특히, 비유창성 요소로 인해 발생하는 교차 현상, 중의성 문제를 고려하여 의존 관계 지배소 설정 방법을 제안했다.한계는 삽입 표현 내에 나타날 수 있는 다양한 현상 또는 표현에 대한 연구가 부족했다. 예를 들어 삽입 어구 내의 head가 마지막 어절이 아닌 경우가 있을 수 있다. 또한 자동 전처리 모듈과 이를 포함한 구문 분석기 개발을 진행하지 못했다. 추후 담화 표지가 아닌 감탄사 처리에 대한 연구가 필요하다. 담화 표지 주석이 없을 경우에 담화 표지와 담화 표지가 아닌 감탄사가 동일한 형태이어서 구분하기 어렵기 때문이다. 더불어 삽입 표현의 다양한 유형과 처리하는 방법에 대해 연구를 진행할 예정이다.

## 참고문헌

[1] 남길임, “구어 비유창성 현상의 주석 체계 연구”, 한국텍스트언어학회, 텍스트언어학30(0), pp.45~72, 2011.
[2] Honnibal, Matthew et al. “Joint Incremental Disfluency Detection and Dependency Parsing”, *Transactions of the Association for Computational Linguistics*, 2014.
[3] Paria Jamshid Lou et al. “Disfluency Detection using Auto-Correlational Neural Networks”, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
[4] Godfrey, John J et al. “Switchboard: Telephone speech corpus for research and development”, *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 1992.
[5] Meeter, Marie, et al. “Dysfluency annotation stylebook for the Switchboard corpus”, Linguistic Data Consortium. Retrieved from ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps, 1995.
[6] Marcus, Mitchell P, et al. “Building a large annotated corpus of English: The Penn Treebank”, *Computational Linguistics, 19(2), 313-330*, 1993.
[7] Rehbein, Ines et al. “Annnotating Spoken Languagage” in Haugh, M. et al. “Best Practices for spoken Corpora in Linguistic Research”, Newcastle upon Tyne: Cambridge Scholars Publishing, 2014.
[8] 고경재, “한국어 담화표지에 대한 고찰 -담화표지의 정체 및 형성의 문제와 관련하여”, 한국어학회, 한국어학 83, pp.97-128, 2019.
[9] 국립국어원, “21세기 세종계획 국어 특수자료 구축”, 21세기 세종 계획 국어 특수 자료 구축 분과 보고서, 2007.
[10] 한국전자통신연구원, 의존 구문 분석 말뭉치 구축을 위한 의존 관계 태그 세트 및 의존 관계설정 방법, 정보통신단체표준(구문표준), TTAK.KO-10/0852, 2015.
[11] Universl Dependencies version2, http://universaldependencies.org
[12] ETRI api, http://aiopen.etri.re.kr/demo_nlu.php