

- <https://github.com/yseokchoi/SejongTree2Dependency>
- <https://github.com/yseokchoi/KoreanDependencyParser>
- pytorch(0.4.1)
  - pip3 install [http://download.pytorch.org/whl/cpu/torch-0.4.1-cp36-cp36m-win\\_amd64.whl](http://download.pytorch.org/whl/cpu/torch-0.4.1-cp36-cp36m-win_amd64.whl) (python 3.6)
  - pip3 install [http://download.pytorch.org/whl/cpu/torch-0.4.1-cp37-cp37m-win\\_amd64.whl](http://download.pytorch.org/whl/cpu/torch-0.4.1-cp37-cp37m-win_amd64.whl) (python 3.7)
- numpy, gensim
  - pip3 install numpy
  - pip3 install gensim

---

# Dependency Parsing을 위한 데이터 구축 실습

---

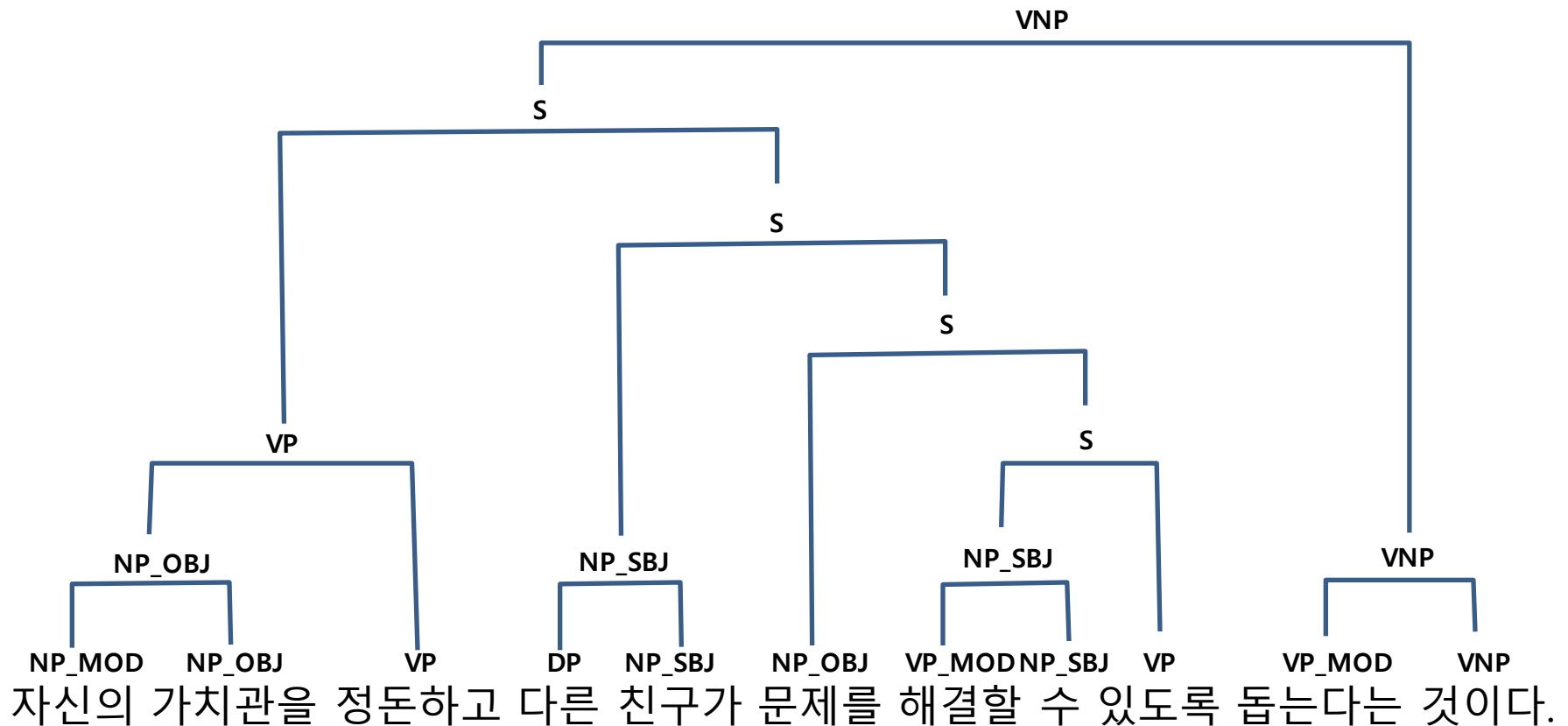
2018. 08.

최용석  
충남대학교  
정보검색 및 지식공학 연구실

- 세종 구문 코퍼스 변환: SejongTree2Dependency
- Transition-based Dependency Parser를 위한 데이터 만들기
  - ARC STANDARD + forward
  - ARC EAGER + backward

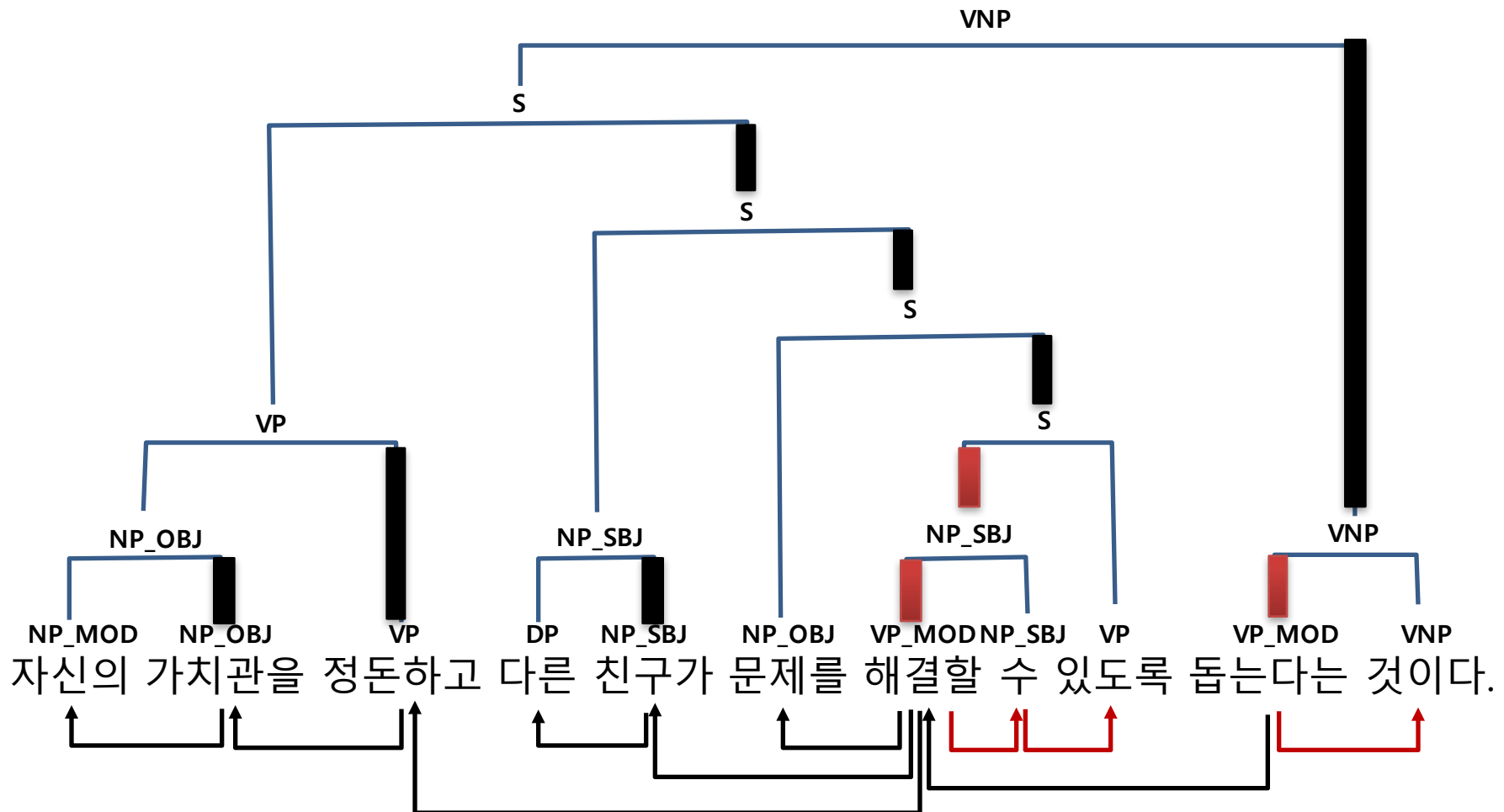
# 세종 구문 코퍼스 변환

Sejong Phrase Structure



# 세종 구문 코퍼스 변환

Sejong Phrase Structure & Dependency Structure



# 세종 구문 코퍼스 변환

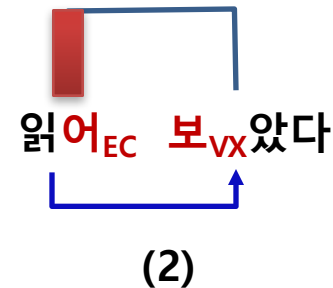
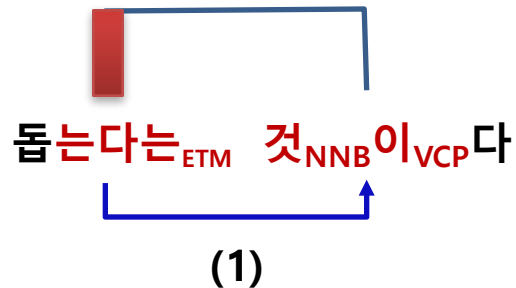
Head Initial 규칙

| Head Initial 규칙 |  | 예제                                   |
|-----------------|--|--------------------------------------|
| (1)             | $X \rightarrow Y$ ('ETM'으로 끝남) Z ('NNB+VCP'로 시작)       | 서로 돕 <u>는</u> <u>다</u> 는 <u>것</u> 이다 |
| (2)             | $X \rightarrow Y$ ('EC로' 끝남) Z ('VX'로 시작)              | 편지를 읽 <u>어</u> <u>보</u> 았다           |
| (3)             | $X \rightarrow Y$ ('ETM'으로 끝남) W ('수' 포함) Z ('있'으로 시작) | 과거를 생각 <u>할</u> <u>수</u> <u>있</u> 을까 |
| (4)             | $X \rightarrow Y$ ('ETM'으로 끝남) W ('것' 포함) Z ('같'로 시작)  | 웃음을 띠 <u>는</u> <u>것</u> <u>같</u> 았어요 |

# 세종 구문 코퍼스 변환

Head Initial 규칙

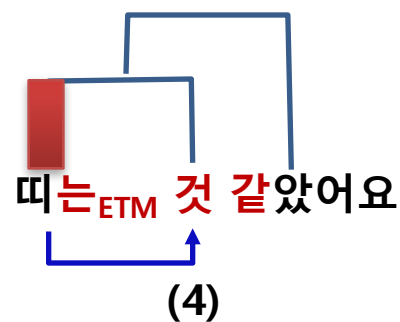
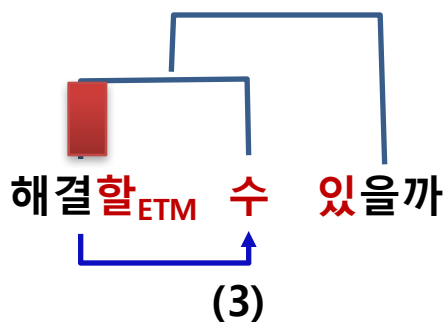
| Head Initial 규칙 |  | 예제                  |
|-----------------|--|---------------------|
| (1)             | $X \rightarrow Y$ ('ETM'으로 끝남) $Z$ ('NNB+VCP'로 시작)         | 서로 <u>돕는다는 것</u> 이다 |
| (2)             | $X \rightarrow Y$ ('EC로' 끝남) $Z$ ('VX'로 시작)                | 편지를 <u>읽어 보</u> 았다  |
| (3)             | $X \rightarrow Y$ ('ETM'으로 끝남) $W$ ('수' 포함) $Z$ ('있'으로 시작) | 과거를 생각할 수 있을까       |
| (4)             | $X \rightarrow Y$ ('ETM'으로 끝남) $W$ ('것' 포함) $Z$ ('같'로 시작)  | 웃음을 띠는 것 같았어요       |



# 세종 구문 코퍼스 변환

Head Initial 규칙

| Head-final 예외 규칙 |   | 예제            |
|------------------|---|---------------|
| (1)              | $X \rightarrow Y$ ('ETM'으로 끝남) $Z$ ('NNB+VCP'로 시작)        | 서로 돕는다는 것이다   |
| (2)              | $X \rightarrow Y$ ('EC로' 끝남) $Z$ ('VX'로 시작)               | 편지를 읽어 보았다    |
| (3)              | $X \rightarrow Y$ ('ETM'으로 끝남) $W$ ('수'포함) $Z$ ('있'으로 시작) | 문제를 해결할 수 있을까 |
| (4)              | $X \rightarrow Y$ ('ETM'으로 끝남) $W$ ('것'포함) $Z$ ('같'로 시작)  | 웃음을 띠는 것 같았어요 |





# 세종 구문 코퍼스 변환

CONLL-U Format

- CONLL-U Format(<http://universaldependencies.org/format.html>)

| Field |         | DESC   |
|-------|---------|--|
| 1     | ID      | Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.   |
| 2     | FORM    | Word form or punctuation symbol.   |
| 3     | LEMMA   | Lemma or stem of word form.  |
| 4     | UPOSTAG | <a href="#">Universal part-of-speech tag</a> .   |
| 5     | XPOSTAG | Language-specific part-of-speech tag; underscore if not available.   |
| 6     | FEATS   | List of morphological features from the <a href="#">universal feature inventory</a> or from a defined <a href="#">language-specific extension</a> ; underscore if not available. |
| 7     | HEAD    | Head of the current word, which is either a value of ID or zero (0).   |
| 8     | DEPREL  | <a href="#">Universal dependency relation</a> to the HEAD ( <a href="#">root</a> if HEAD = 0) or a defined language-specific subtype of one.                                     |
| 9     | DEPS    | Enhanced dependency graph in the form of a list of head-deprel pairs.  |
| 10    | MISC    | Any other annotation.  |

# 세종 구문 코퍼스 변환

## CONLL-U Format 예제

#SENTID:1

#FILE:practice.txt

#ORGSENT: 자신의 가치관을 정돈하고 다른 친구가 문제를 해결할 수 있도록 돕는다는 것이다.

| #  | FORM | LEMMA   | UPOSTAG | XPOSTAG       | FEATS | HEAD | DEPREL* | DEPS | MISC |
|----|------|---------|---------|---------------|-------|------|---------|------|------|
| 1  | 자신의  | 자신 의    | NOUN    | NNG+JKG       | -     | 2    | nmod    | -    | -    |
| 2  | 가치관을 | 가치관 을   | NOUN    | NNG+JKO       | -     | 3    | obj     | -    | -    |
| 3  | 정돈하고 | 정돈 하 고  | VERB    | NNG+XSV+EC    | -     | 7    | advcl   | -    | -    |
| 4  | 다른   | 다른      | ADJ     | MM            | -     | 5    | amod    | -    | -    |
| 5  | 친구가  | 친구 가    | NOUN    | NNG+JKS       | -     | 7    | nsubj   | -    | -    |
| 6  | 문제를  | 문제 를    | NOUN    | NNG+JKO       | -     | 7    | obj     | -    | -    |
| 7  | 해결할  | 해결 하 ㄹ  | VERB    | NNG+XSV+ETM   | -     | 10   | advcl   | -    | -    |
| 8  | 수    | 수       | NOUN    | NNB           | -     | 7    | aux     | -    | -    |
| 9  | 있도록  | 있 도록    | VERB    | VV+EC         | -     | 8    | aux     | -    | -    |
| 10 | 돕는다는 | 돕 는다는   | VERB    | VV+ETM        | -     | 0    | root    | -    | -    |
| 11 | 것이다. | 것 이 다 . | ADJ     | NNB+VCP+EF+SF | -     | 10   | aux     | -    | -    |

\* 변경될 수 있음

# 세종 구문 코퍼스 변환

## CONLL-U Format 예제

#SENTID:6

#FILE:practice.txt

#ORGSENT: 웅가로는 “실내 장식품을 디자인할 때 옷을 만들 때와는 다른 해방감을 느낀다”고 말한다.

| #  | FORM | LEMMA     | UPOSTAG | XPOSTAG       | FEATS | HEAD | DEPREL* | DEPS | MISC          |
|----|------|-----------|---------|---------------|-------|------|---------|------|---------------|
| 1  | 웅가로는 | 웅가로 는     | PROPN   | NNP+JX        | -     | 15   | nsubj   | -    | -             |
| 2  | “    | “         | PUNCT   | SS            | -     | 12   | punct   | -    | SpaceAfter=No |
| 3  | 실내   | 실내        | NOUN    | NNG           | -     | 4    | nmod    | -    | -             |
| 4  | 장식품을 | 장식품 을     | NOUN    | NNG+JKO       | -     | 5    | obj     | -    | -             |
| 5  | 디자인할 | 디자인 하 ㄹ   | VERB    | NNG+XSV+ETM   | -     | 6    | acl     | -    | -             |
| 6  | 때    | 때         | NOUN    | NNG           | -     | 12   | obl     | -    | -             |
| 7  | 옷을   | 옷을        | NOUN    | NNG+JKO       | -     | 8    | obj     | -    | -             |
| 8  | 만들   | 만들 ㄹ      | VERB    | VV+ETM        | -     | 9    | acl     | -    | -             |
| 9  | 때와는  | 때 와 는     | NOUN    | NNG+JKB+JX    | -     | 10   | nsubj   | -    | -             |
| 10 | 다른   | 다르 ㄴ      | ADJ     | VA+ETM        | -     | 11   | acl     | -    | -             |
| 11 | 해방감을 | 해방감 을     | NOUN    | NNG+JKO       | -     | 12   | obj     | -    | -             |
| 12 | 느낀다  | 느끼 ㄴ 다    | VERB    | VV+EC         | -     | 15   | ccomp   | -    | SpaceAfter=No |
| 13 | ”    | ”         | PUNCT   | SS            | -     | 12   | punct   | -    | SpaceAfter=No |
| 14 | 고    | 고         | ADP     | JKQ           | -     | 12   | case    | -    | -             |
| 15 | 말한다. | 말 하 ㄴ 다 . | VERB    | NNG+XSV+EF+SF | -     | 0    | root    | -    | -             |

\* 변경될 수 있음

# 세종 구문 코퍼스 변환

Sejong POS Tag to Universal POS Tag

| 대분류 | 세종 품사 태그 | UPOSTAG |
|-----|----------|---------|
| 체언  | NNG NNB  | NOUN    |
|     | NNP      | PROPN   |
|     | NR       | NUM     |
|     | NP       | PRON    |
| 용언  | VV       | VERB    |
|     | VA       | ADJ     |
|     | VX       | AUX     |
|     | VCP VCN  | ADJ     |
| 접두사 | XPN      | PART    |
| 접미사 | XSN      | PART    |
|     | XSV      | VERB    |
|     | XSA      | ADJ     |

| 대분류 | 세종 품사 태그 | UPOSTAG        |
|-----|----------|----------------|
| 관형사 | MM       | DET ADJ<br>NUM |
| 부사  | MAG      | ADV            |
|     | MAJ      | CCONJ<br>SCONJ |
| 감탄사 | IC       | INJT           |
| 조사  | J*       | ADP            |
|     | JC       | CCONJ          |
| 어미  | E*       | PART           |
| 어근  | XR       | NOUN           |

# 세종 구문 코퍼스로부터 의존 구문 구조로 변환해보기

## ■ Parameters

- **root\_dir**: 세종 코퍼스 폴더 위치
- **file\_name(optional)**: 세종 코퍼스 파일 이름(하나의 세종 코퍼스 파일만 읽고자 할때)
- **save\_file**: 변환한 구문 구조를 저장할 파일 이름
- **head\_initial\_file**: head-final 예외 규칙 파일명
- **head\_final**: 1 if *head-final* 0 else (default: 0)

# 세종 구문 코퍼스로부터 의존 구문 구조로 변환해보기

- 실행 방법

```
python SejongToDependency.py -root_dir ./Corpus -save_file sejong.conll -  
head_initial_file ./Rules/linear_rules.txt
```



# 세종 구문 코퍼스 변환

Head Initial 규칙 정의

## ▪ Head Initial 규칙 정의(두 어절)

(1) parent label

(2) right most node label(RMN Label)

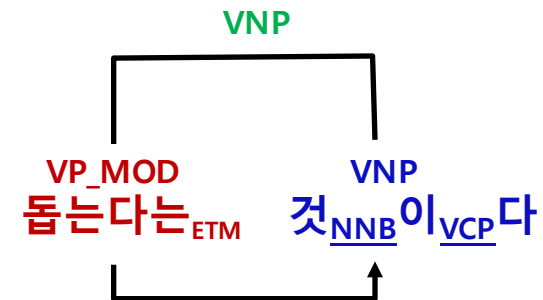
(3) right most node(RMN)

(4) left most node label(LMN Label)

(5) left most node(LMN)

(6) left context

(7) right context



| Parent Label | RMN Label | RMN | LMN Label | LMN     | Left Context | Right Context |
|--------------|-----------|-----|-----------|---------|--------------|---------------|
| VNP*         | VP_MOD    | -   | VNP       | NNB+VCP | -            | -             |



# 세종 구문 코퍼스 변환

Head Initial 규칙 정의

## ▪ Head Initial 규칙 정의(세 어절)

(1) parent label

(2) right most node label(RMN Label)

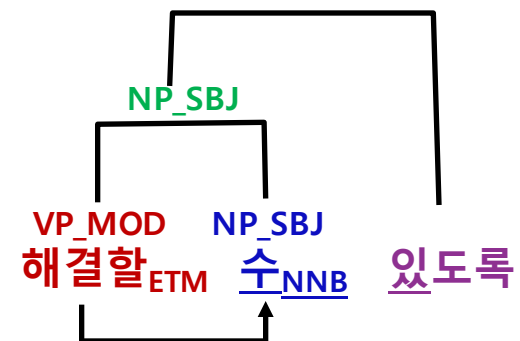
(3) right most node(RMN)

(4) left most node label(LMN Label)

(5) left most node(LMN)

(6) left context

(7) right context



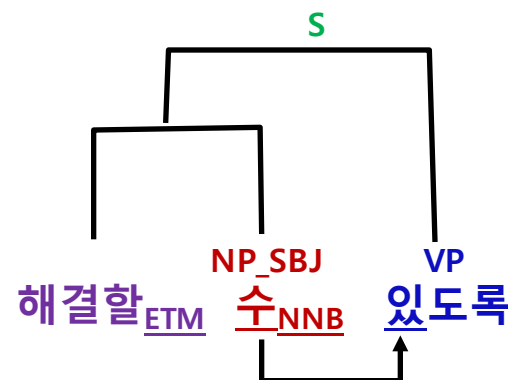
| Parent Label | RMN Label | RMN | LMN Label | LMN   | Left Context | Right Context |
|--------------|-----------|-----|-----------|-------|--------------|---------------|
| NP_SBJ       | VP_MOD    | -   | NP_SBJ    | 수/NNB | -            | 있             |

# 세종 구문 코퍼스 변환

Head Initial 규칙 정의

## Head Initial 규칙 정의(세 어절)

- (1) parent label
- (2) right most node label(RMN Label)
- (3) right most node(RMN)
- (4) left most node label(LMN Label)
- (5) left most node(LMN)
- (6) left context
- (7) right context



| Parent Label | RMN Label | RMN   | LMN Label | LMN | Left Context | Right Context |
|--------------|-----------|-------|-----------|-----|--------------|---------------|
| S*           | NP_SBJ    | 수/NNB | VP*       | 있   | ETM          | -             |

# Transition-based Dependency Parser를 위한 데이터 만들기

## ■ Parameters

- **root\_dir**: 세종 코퍼스 폴더 위치
- **file\_name(optional)**: 세종 코퍼스 파일 이름(하나의 세종 코퍼스 파일만 읽고자 할때)
- **save\_file**: 변환한 구문 구조를 저장할 파일 이름
- **arc\_standard**: ARC-Standard를 이용한 데이터 변환

# Transition-based Dependency Parser를 위한 데이터 만들기

- 실행 방법

- **ARC-STANDARD**

```
python MakeTransitionCorpus.py -root_dir ./ -file_name sejong.conll -  
save_file sejong_ARC_STANDARD.txt -arc_standard
```

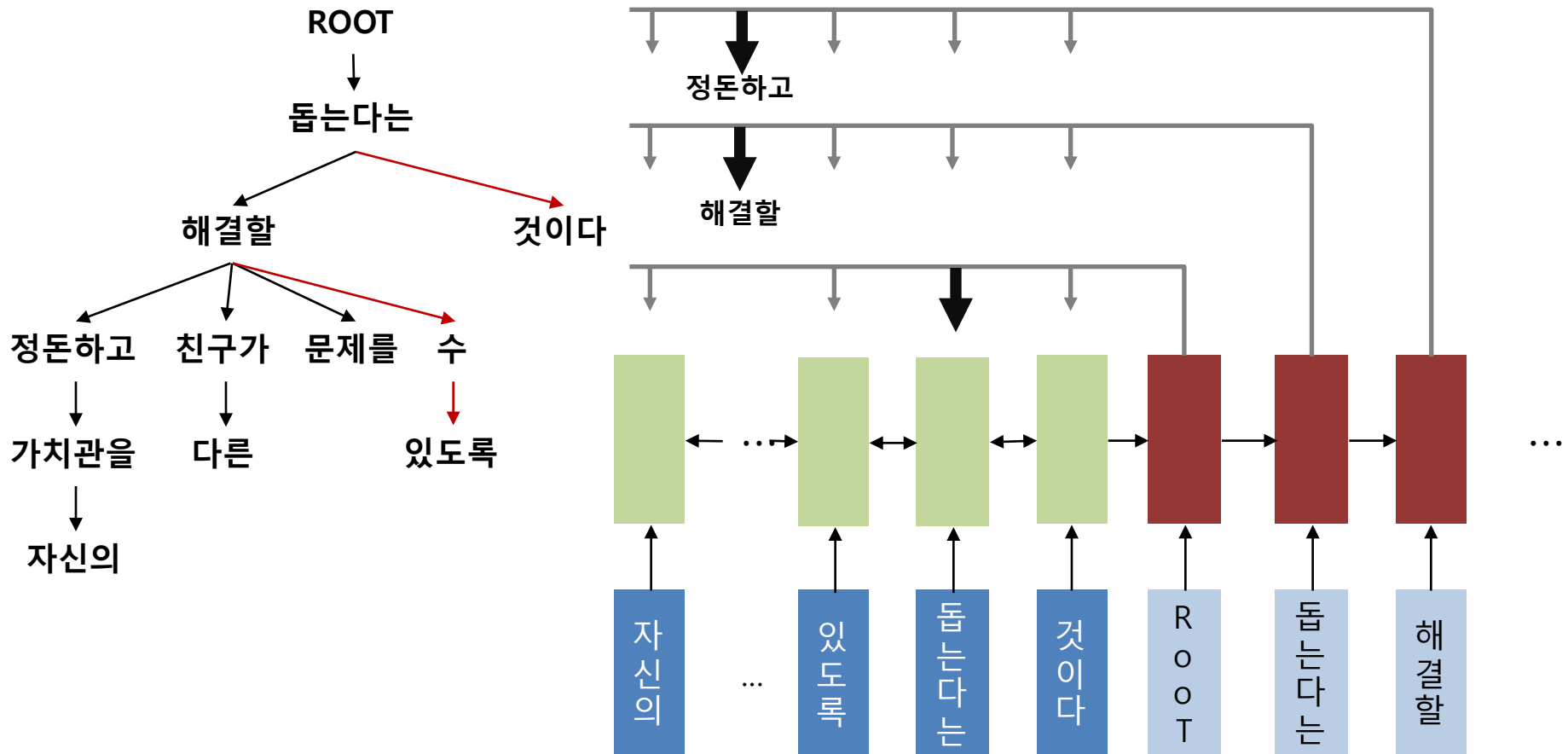
- **ARC-EAGER**

```
python MakeTransitionCorpus.py -root_dir ./ -file_name sejong.conll -  
save_file sejong_ARC_EAGER.txt
```

# 의존 구문 파서

## Stack-Pointer Networks

SENT: 자신의 가치관을 정돈하고 다른 친구가 문제를 해결할 수 있도록 돕는다는 것이다.



# 의존 구문 파서

실습

---

## ■ Parameters

- **model\_path**: 저장한 모델 폴더 위치
- **model\_name**: 저장한 모델 이름
- **output\_path**: 결과 파일 저장할 폴더 위치
- **test**: test file 위치+파일명

## ■ 실행 방법

```
python StackPointerParser_test.py --model_path ./models/ --model_name  
models.pt --output_path ./output/ --test ./test/sejong.conll
```

---

The End  
Thank you

---