세종 구문 분석 말뭉치의 의존 구문 구조로의 변환

1. 세종 구문 분석 말뭉치 → 의존 구조

세종 구문 분석 말뭉치는 모든 구절(phrase)이 이진 규칙으로 구성되어 있다. 또한 한국어는 대표적인 head-final 언어이기 때문에 각 구절에 대응되는 의존 관계를 추출하는 것이 용이하다.

phrase structure		**dependency structure**
sub tree	phrasal rule	dependency relation
(X (Y) (Z))	X → Y **Z**	dependent: HW(Y) head: HW(Z) HW(Z) → HW(Y)

Head-final 언어이기 때문에 구문 규칙 [X → Y Z]에서 가장 오른쪽에 있는 구절 Z의 중심어가 X의 중심어가 된다. 이러한 특성을 이용하여 구문 규칙 [X → Y Z]에서 추출할 수 있는 의존 관계는 [HW(Z) → HW(Y)]이다. 'HW()'는 해당 구절의 중심어를 의미하며, head와 dependent 사이는 [H → D]로 표시한다.

다음은 예제 문장 "다른 원인이 있다."에 대한 구문 트리와 추출 가능한 의존 관계이다.

phrase structure		**dependency structure**
sub tree	phrasal rule	dependency relation
(S (NP_SBJ (DM 다른) (NP_SBJ 원인이)) (VP 있다))	NP_SBJ → DM NP_SBJ1 S → NP_SBJ2 VP HW(DM): 다른 HW(NP_SBJ1):원인이 HW(NP_SBJ2): 원인이 HW(VP): 있다	원인이 → 다른 있다 → 원인이 ROOT → 있다

2. Head initial 예외 규칙

Head-final 언어이기 때문에 각 구절의 중심어(head word)는 구절이 차지하는 범위의 가장 오른쪽 어절로 결정하였다. 그러나, 예제 문장 '있을 것 같다'의 동사 구절에서는 실질적인 중심어가 제일 마지막 단어인 '같다'가 아니라 '있을'이 되어야 한다.

예) 다른 원인이 (VP 있을 것 같다.)

한국어가 head-final 언어라 하더라도 중심어가 가장 뒤쪽이 아닌 앞쪽으로 이동되어야 하는 경우들이 발생한다. 이와 같이 중심어가 앞쪽에 있어야 하는 예외 규칙은 주로 본용언과 이를 따르는 보조용언에 의해 구성되는 동사 구절에 나타나며 이들 본용언과 보조용언은 연속적인 단어들 사

이에서 발생하는 경우가 대부분이다.

	Head Initial Rules & Conditions	예제
(1)	X → Y('ETM'으로 끝남) Z('NNB+VCP'로 시작)	역할을 수행하는 것이다.
(2)	X → Y('기/ETN'으로 끝남) Z('때문+VCP'로 시작)	타겟이 움직이기 때문이다.
(3)	X → Y('EC로' 끝남) Z('VX'로 시작)	편지를 읽어 보았다
(4)	X → Y('ETM'으로 끝남) W('수'포함) Z('있'으로 시작)	과거를 생각할 수 있을까?
(5)	X → Y('ETM'으로 끝남) W('것'포함) Z('같'로 시작)	웃음을 띠는 것 같았어요.

예외 규칙 (1)~(3)은 연속된 두 어절 사이의 규칙이고 (4)~(5)는 연속된 세 어절 사이의 규칙이다. 일반적으로 가장 오른쪽 구절인 Z의 중심어가 구절 X의 중심어가 되어야 한다. 그러나 예외 규칙의 경우에는 구절 X의 중심어는 가장 왼쪽 구절 Y의 중심어가 되어야 한다.

다음은 (1)의 예제 문장에 대한 구문 트리이다.

```
(S   (NP_OBJ 역할을)
     (VP  (VP_MOD 수행하는)  (VNP 것이다)))
```

동사 구절 VP "수행하는 것이다"의 실제 중심어는 오른쪽 어절 "것이다"가 아니라 "수행하는"이 되어야 한다. 그래야만 "역할을"의 중심어가 "것이다"가 아닌 "수행하는"이 될 수 있다.

Head-initial 예외 규칙에 대한 처리 방식에 따라 두 개의 다른 의존 트리 말뭉치로 변환을 하였다.

3. 의존 구문 구조로 변환

3.1 엄격한 head-final 준수 (Rigid Head Finality)

Head-final 속성을 철저히 준수하기 위하여 head-initial인 예외 규칙에 대해서도 중심어는 항상 가장 오른쪽 노드로 설정한다. 그렇기 때문에 문장의 최종 중심어는 항상 제일 마지막 어절이 된다. 대신, head-initial에 의해 구성되는 구절이 다른 구절과 결합할 경우에는 head-final 원칙을 준수하면서 최소한의 수정을 가하여 보조용언이 아닌 본용언과의 의존 관계를 설정할 수 있도록 하였다.

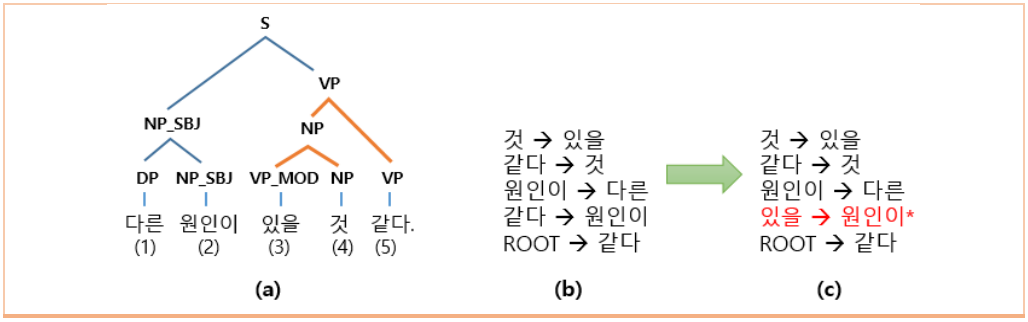
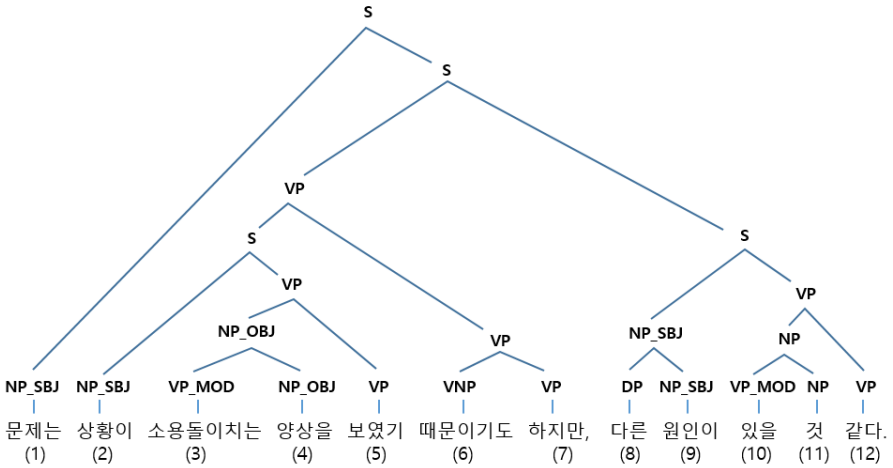
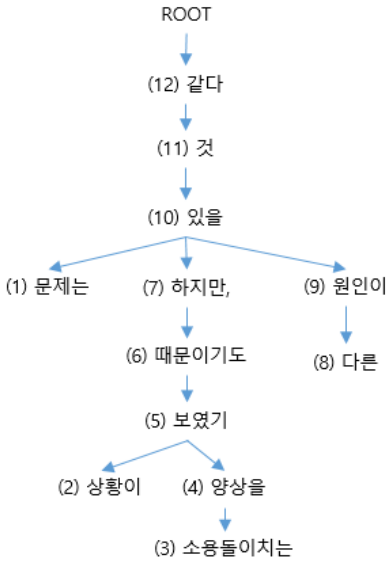
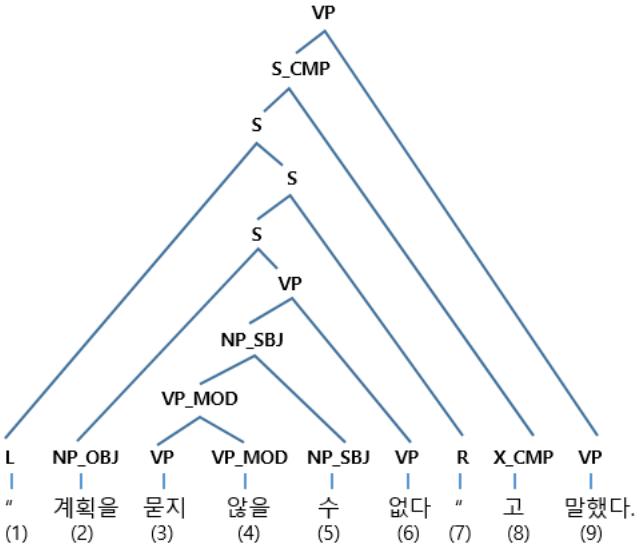
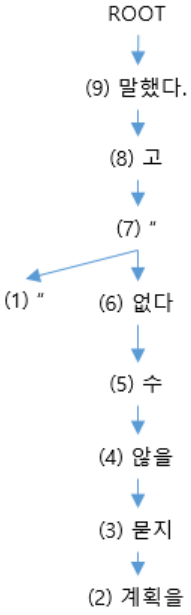


그림 (a)에서 붉은색으로 표시된 규칙이 head-initial 예외 규칙이다. (b)는 head-final 원칙에 의해 추출한 의존관계이다. 이진 트리에서 중심어는 항상 오른쪽 구절이 된다는 원칙을 그대로 적용한 결과이다. 이와 같이 head-final 원칙에 의해 의존 관계를 추출한 후, [S → NP_SBJ VP]의 규칙 적용 시, NP_SBJ의 '원인이'의 중심어를 '같다'에서 본용언 '있을'로 수정한다. 수정한 결과가 그림 (c)이다.

	예제
(1) Phrase Structure Tree	

(1) Dependency Tree	
(2) Phrase Structure Tree	
(2) Dependency Tree	

3.2 덜 엄격한 head-final 준수 (Non-Rigid Head Finality)

덜 엄격한 head-final 규칙 준수의 경우에는 head-initial 예외 규칙에 대해서는 중심어를 오른쪽 구절이 아닌 왼쪽 구절로 설정한다. 이렇게 할 경우, head-initial 예외 규칙에 대해서는 head-finality가 위배된다.

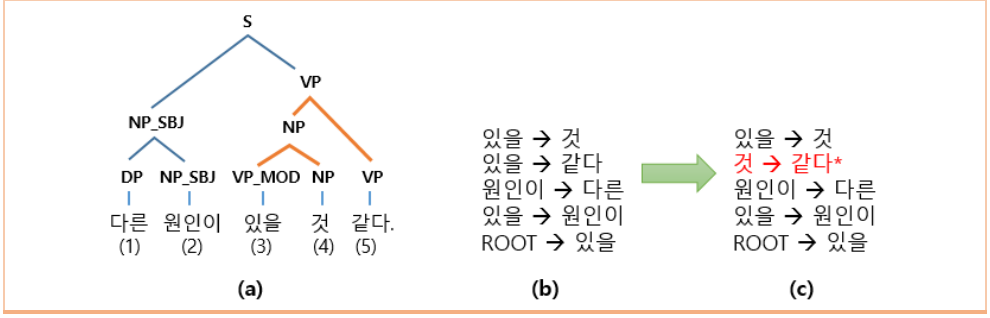
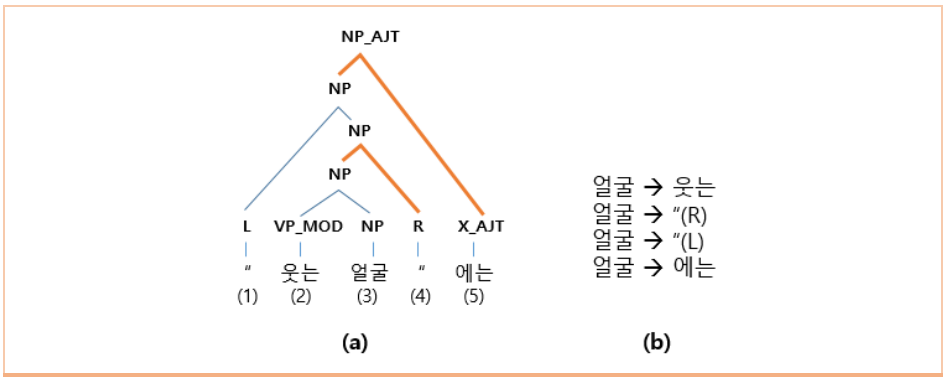
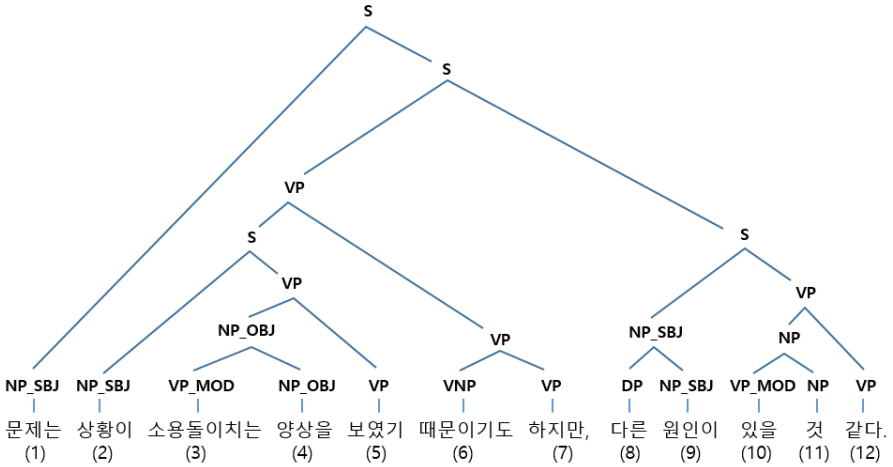
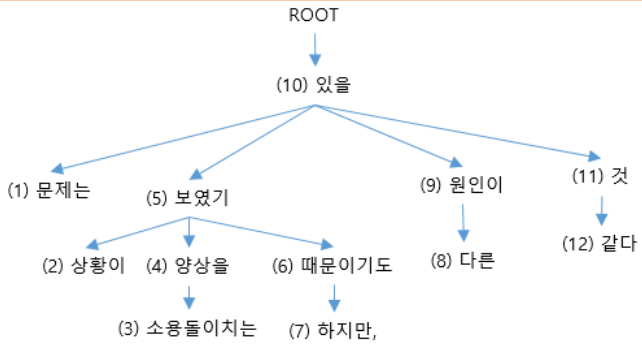
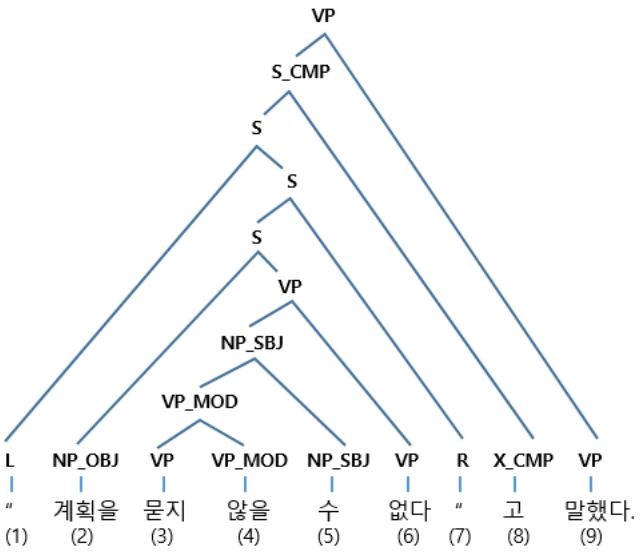


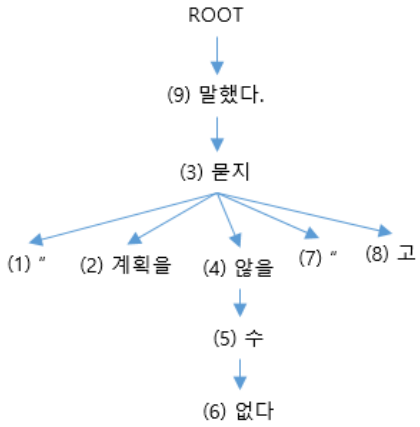
그림 (a)에서 붉은 색으로 표시된 구절이 head-initial에 의해 구성된 부분이다. 그렇기 때문에 어절 (3)~(5)의 VP 경우 '있을'이 중심어가 되며, (2)번 어절 '원인이'의 중심어도 '같다'가 아닌 '있을'로 설정된다. 결과 (b)가 head-initial 규칙을 그대로 적용하여 추출한 의존 관계이다. 다만, 세 어절 이상으로 구성된 head-initial 규칙의 경우, 보조용언 표현을 연속된 어절로 만들기 위해 (c)와 같은 수정을 가하였다.

또한 보조용언에 관련된 규칙 이외에도 심볼과 기능어에 대한 규칙을 head-initial 예외 규칙으로 추가하였다.

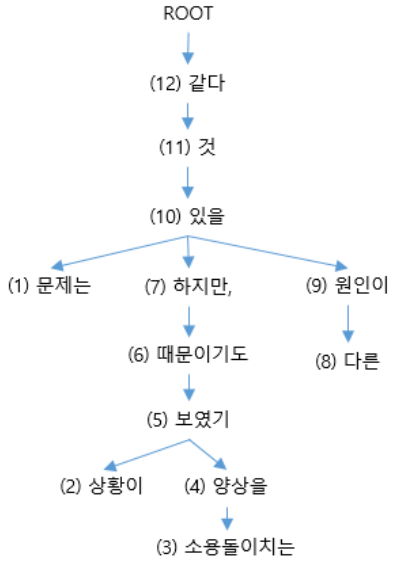
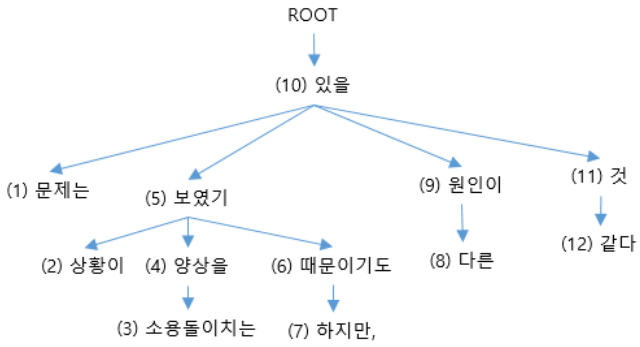
	Head initial Rules	예제
(1)	X → **Y** Z(punctuation으로만 구성)	" 웃는 얼굴 "
(2)	X → **Y** Z(조사나 어미로만 구성)	류창렬 씨(44세 회사원)는
(3)	X → **Y** Z(접미사로만 구성)	자료를 다운로드(down load)한다

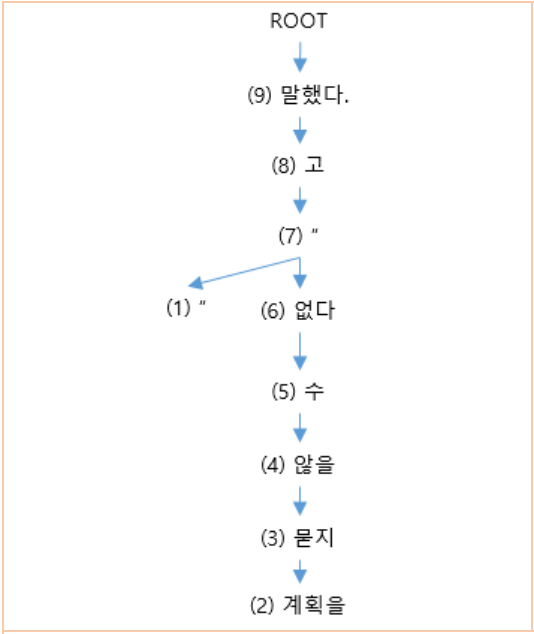


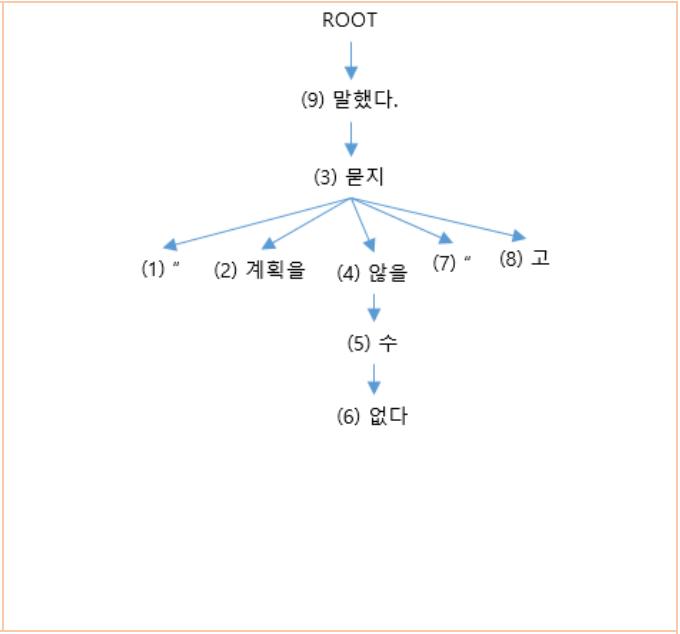
	예제
(1) Phrase Structure Tree	
(1) Dependency Tree	
(2) Phrase Structure Tree	

(2) Dependency Tree	

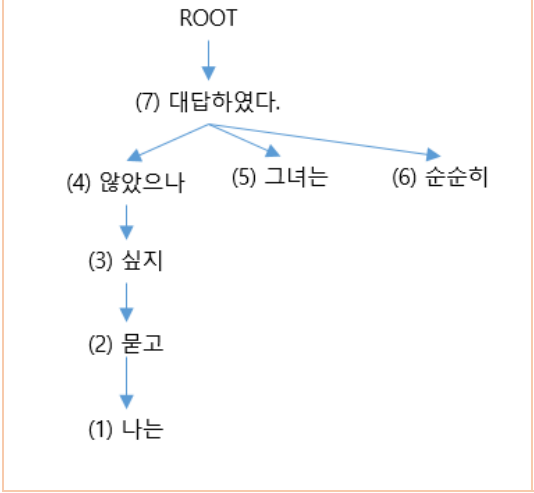
3.3 Rigid Head Final vs. Non-Rigid Head Final 구문 트리 비교

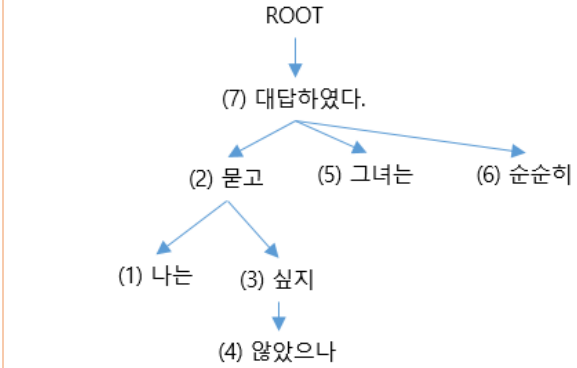
Rigid Head Final	Non-Rigid Head Final
문장: 이 문제는 상황이 소용돌이치는 양상을 보였기 때문이기도 하지만, 다른 원인이 있을 것 같다.	
	
문장: "계획을 묻지 않을 수 없다"고 말했다.	





문장: 나는 묻고 싶지 않았으나 그녀는 순순히 대답하였다.





4. CoNLL-U 형식

세종 구문 분석 말뭉치를 의존 구조로 변환하고 CoNLL-U의 형식으로 저장한다.

CoNLL-U는 다음과 같이 10개의 필드로 한 입력 토큰을 표시한다.

	Field	**설명**
1	ID	토큰(단어 또는 어절) 번호
2	FORM	어절의 word surface 형태
3	LEMMA	어절의 형태소 원형
4	UPOSTAG	Universal POS 태그
5	XPOSTAG	세종 코퍼스의 품사 정보
6	FEATS	형태소 과련 속성 정보
7	HEAD	Head 어절 번호, 문장의 ROOT인 경우에는 0

8	DEPREL	의존 관계의 관계 정보
9	DEPS	확장된 의존 관계(의존 그래프) 정보
10	MISC	기타 정보

예제:

엠마누엘 웅가로는 "실내 장식품을 디자인할 때 옷을 만들 때와는 다른 해방감을 느낀다"고 말한다.

Rigid Head-final 버전

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	엠마누엘	엠마누엘	PROPN	NNP	-	**2**	NP	-	-
2	웅가로는	웅가로 는	PROPN	NNP JX	-	**16**	NP_SBJ	-	-
4	"	"	PUNCT	SS	-	**14**	L	-	SpaceAfter=No
4	실내	실내	NOUN	NNG	-	**5**	NP	-	-
5	장식품을	장식품 을	NOUN	NNG JKO	-	**6**	NP_OBJ	-	-
6	디자인할	디자인 하 ㄹ	VERB	NNG XSV ETM	-	**7**	VP_MOD	-	-
7	때	때	NOUN	NNG	-	**13**	NP_AJT	-	-
8	옷을	옷 을	NOUN	NNG JKO	-	**9**	NP_OBJ	-	-
9	만들	만들 ㄹ	VERB	VV ETM	-	**10**	VP_MOD	-	-
10	때와는	때 와 는	NOUN	NNG JKB JX	-	**11**	NP_SBJ	-	-
11	다른	다르 ㄴ	ADJ	VA ETM	-	**12**	VP_MOD	-	-
12	해방감을	해방감 을	NOUN	NNG JKO	-	**13**	NP_OBJ	-	-
13	느낀다	느끼 ㄴ다	VERB	VV EC	-	**14**	VP	-	SpaceAfter=No
14	"	"	PUNCT	SS	-	**15**	VP	-	SpaceAfter=No
15	고	고	ADP	JKQ	-	**16**	VP_CMP	-	-
16	말한다.	말 하 ㄴ다 .	VERB	NNG XSV EF SF	-	**0**	ROOT	-	-

Non-rigid Head final 버전

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	엠마누엘	엠마누엘	PROPN	NNP	-	**2**	NP	-	-
2	웅가로는	웅가로 는	PROPN	NNP JX	-	**16**	NP_SBJ	-	-
3	"	"	PUNCT	SS	-	**13**	L	-	SpaceAfter=No
4	실내	실내	NOUN	NNG	-	**5**	NP	-	-
5	장식품을	장식품 을	NOUN	NNG JKO	-	**6**	NP_OBJ	-	-
6	디자인할	디자인 하 ㄹ	VERB	NNG XSV ETM	-	**7**	VP_MOD	-	-
7	때	때	NOUN	NNG	-	**13**	NP_AJT	-	-
8	옷을	옷 을	NOUN	NNG JKO	-	**9**	NP_OBJ	-	-
9	만들	만들 ㄹ	VERB	VV ETM	-	**10**	VP_MOD	-	-
10	때와는	때 와 는	NOUN	NNG JKB JX	-	**11**	NP_SBJ	-	-
11	다른	다르 ㄴ	ADJ	VA ETM	-	**12**	VP_MOD	-	-
12	해방감을	해방감 을	NOUN	NNG JKO	-	**13**	NP_OBJ	-	-

13	느낀다	느끼 ㄴ다	VERB	VV EC	-	**16**	VP_CMP	-	SpaceAfter=No
14	"	"	PUNCT	SS	-	**13**	R	-	SpaceAfter=No
15	고	고	ADP	JKQ	-	**13**	X_CMP	-	-
16	말한다.	말 하 ㄴ다 .	VERB	NNG XSV EF SF	-	**0**	ROOT	-	-

'LEMMA'는 입력 어절(FORM)의 형태소 분석 결과 중 형태소 정보만을 space단위로 분리하여 저장한다. 'XPOSTAG'는 입력 어절(FORM)의 형태소 분석 결과 중 품사 정보만을 space단위로 분리하여 저장한다.

'DEPREL'은 세종 구문 분석 말뭉치의 dependent 구절의 label 정보를 그대로 저장하였다.

세종 코퍼스의 품사 태그와 'UPOSTAG' 사이의 품사 매핑은 다음 표와 같다. 한 어절에 대한 UPOSTAG는 어절을 구성하는 가장 핵심 형태소의 UPOSTAG로 결정하였다.

대분류	세종 품사 태그	UPOSTAG
체언	NNG	NOUN
	NNP	PROPN
	NNB	NOUN
	NR	NOUN
	NP	PRON
용언	VV	VERB
	VA	ADJ
	VX	AUX
	VCP	AUX
	VCN	AUX
관형사	MM	DET
부사	MAG	ADV
	MAJ	CCONJ
감탄사	IC	INJT
조사	JKS	ADP
	JKC	ADP
	JKG	ADP
	JKO	ADP
	JKB	ADP
	JKV	ADP

	JKQ	ADP
	JX	ADP
	JC	ADP
선어말 어미	EP	PART
어말 어미	EF	PART
	EC	PART
	ETN	PART
	ETM	PART
접두사	XPN	PART
접미사	XSN	PART
	XSV	AUX
	XSA	AUX
어근	XR	NOUN
부호	SF	PUNCT
	SP	PUNCT
	SS	PUNCT
	SE	PUNCT
	SO	PUNCT
	SW	SYM
분석 불능	NF	X
	NV	X
	NA	X
한글 이외	SL	X
	SH	X
	SN	NUM

Head-initial Rules

	parent_label	ightmostnode_labe	rightmostnode_eojul	leftmostnode_label	leftmostnode_eojul	left context	right context
2				수정 버전 - 180314			
3	VP* VNP* S*	VP_MOD VNP_MOD	-	VNP	NNB+VCP 정도+VCP 노릇+VCP 지경+VCP 탓+VCP	-	-
4	VP* VNP* S*	VP_MOD VNP_MOD	-	VP*	만/NNB+하 듯/NNB+하 체/NNB+하 뻔/NNB+하 성/NNB+싶	-	-
5	VP* VNP* S*	VP* S* VNP*	기/ETN	VNP	때문+VCP 일쑤+VCP 십상+VCP 마련+VCP	-	-
6	VP* VNP* S*	VP* S* VNP*	기/ETN+JX 기/ETN+로/JKB	VP*	하	-	-
7	VP* VNP* S*	VP* VNP*	EC	VP* S*	VX	-	-
8	VP* VNP* S*	VP* VNP*	게/EC	VP* VNP* S*	마련+VCP	-	-
9	VP* VNP* S*	VP* VNP*	도록/EC 으면/EC 면/EC	VP* VNP* S*	되/VV	-	-
10	VP* VNP* S*	VP* VNP*	으면/EC 면/EC	VP* S*	안+되	-	-
11	VP* VNP* S*	VP* VNP*	어/EC 아/EC	VP* S*	보/VV 오/VV	-	-
12	NP_SBJ	VP_MOD	-	NP_SBJ	수/NNB 리/NNB 바/NNB 탓/NNB 적/NNB	-	있 없
13	VP* VNP* S*	NP_SBJ	수/NNB 리/NNB 바/NNB 탓/NNB 적/NNB 수/NNB+가/JKS 리/NNB+가/JKS 바/NNB+가/JKS 적/NNB+이/JKS	VP*	있 없	ETM	-
14	NP*	VP_MOD	-	NP	것/NNB 거/NNB	-	같 뿐+VCP
15	VP* VNP* S*	NP	것/NNB 거/NNB	VP* VNP*	같 뿐+VCP	ETM	-
16	NP*	VP_MOD	-	NP*	뿐/NNB	-	아니/VCN
17	VP* VNP* S*	NP*	뿐/NNB	VP*	아니/VCN	ETM	-
18	NP*	VP_MOD	-	NP*	듯/NNB	-	하
19	VP* VNP* S*	NP*	듯/NNB	VP*	하	ETM	-
20	VP* VNP* S*	VP* VNP*	으면/EC 면/EC	AP	안/MAG	-	되
21	VP* VNP* S*	AP	안/MAG	VP*	되	으면/EC 면/EC	-

	parent_label	ightmostnode_labe	rightmostnode_eojul	leftmostnode_label	leftmostnode_eojul	left context	right context
2	-	-	-	X*	SS SP	-	-
3	-	-	-	R	SS SP	-	-
4	-	-	-	X*	J* E* XP* XS*	-	-