

자연어 QA workshop

# 한국어 의미처리시스템

2015.08.21

 **KLPLAB** 울산대학교 한국어처리연구실

옥 철 영

울산대학교

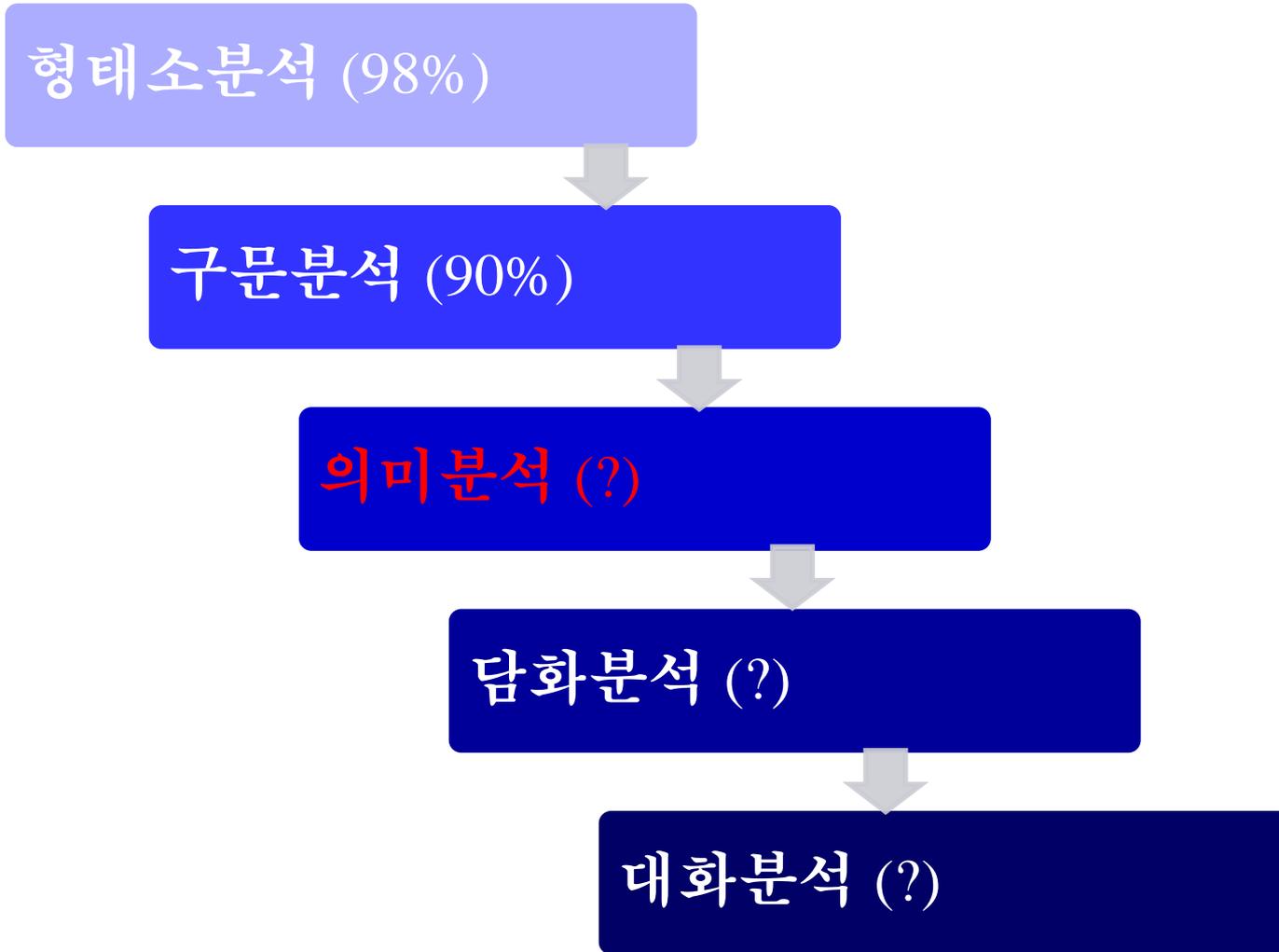
컴퓨터정보통신공학부/국어국문학과



- ❖ 한국어 의미자원
  - 어휘지도 (UWordMap)
  - 의미역 & UPropBank
- ❖ 한국어 의미처리시스템
  - 동형이의어 WSD (UTagger-HM)
  - 의존관계 분석 (UTagger-DP)
  - 다의어 WSD (UTagger-PS)
  - 의미역 태깅시스템 (UTagger-SR)
  - 개체명 인식기 (UTagger-NE)
- ❖ 시연 (한국어어휘지도, UTagger 시스템)



# 한국어 분석 단계 및 현 처리수준





## Lexical Semantics

### WSD, NER

- 하나의 형태소가 문맥에 따라 여러개의 의미로 해석될 때, 하나의 의미를 결정
- Homograph WSD
- Polysemy WSD
- Named Entity Recognition
- Lexical Semantic Network

## Sentence Semantics

### Parser, SRL

- 문장 내의 술어-논항(predicate-argument) 관계에 적합한 의미 관계(Semantic Role)를 결정
- Subcategorization
- Semantic Restriction
- Semantic Role Labeling
- Ontology & Inference



# 한국어 의미자원

- ✓ 한국어 어휘지도
- ✓ 의미역 & UPropBank
- ✓ 용언 의미 군집화 & 계층화

# 어휘의미망 필요성

- ❖ 어휘의미론(Lexical Semantics) 관점
  - 의미관계(상의어, 하의어, 반의어, 유의어, 부분/전체) 정립
  - 의미자질을 통한 개념화
  
- ❖ 문장의미론(Sentence Semantics) 관점
  - 통사구조와 의미구조 분석
  - 논항의 의미제약
  - 의미역(thematic roles) 결정
  
- ❖ Exo Brain의 기반 지식
  - QA에서의 질의유형, 정답유형 판단
  - 어휘의미의 개념화/범주화
  - 논항의 의미제약
  - 의미역 기반 triple 구조의 knowledge 표현
  - 다의어 WSD



## □ 부산대 윤애선교수 발표자료 (2008)

명칭	구축방식 (기반)	구축연도	구축기관	의미/개념 vs 어휘 수	구축 품사
<b>KorLex</b>	<b>참조(PWN)</b>	<b>2004-현재</b>	부산대학교	126,653s/143,479w	명, 동, 형, 부, 분류사
한국어 시소러스	참조(PWN)	1997-2000	포항공대	18,362s/ 21,390w	명
다국어 DB	참조(EWN)	2000-2006	고려대 민족문화연구소	5,500w	명
<b>CoreNet</b>	참조(NTT)	1995-2004	KAIST	2,938n/ 62,632w	명, 동, 형
<b>U-WIN</b>	<b>직접(표준)</b>	<b>2002-현재</b>	울산대학교	<b>57,792s/ 430,000w</b>	모든 품사
세종전자 사전	직접	1998-2007	서울대학교	581n/540,000w	모든 품사

❖ KorLex 규모 (2015. 5. 7 기준, ()는 2012. 3)

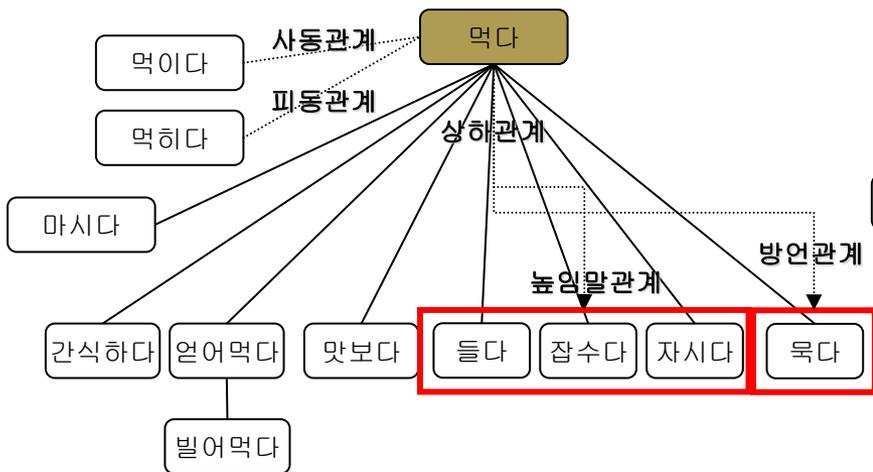
단위	Korean WordNet				
	명사	동사	형용사	부사	계
어형	106,294 (90,909)	17,962 (17,957)	41,107 (19,694)	7,806 (3,032)	174,350 (132,877)
신셋	101,869 (92,184)	17,075 (16,937)	18,582 (18,560)	3,668 (3,651)	142,571 (132,943)
어의	121,216 (104,417)	20,346 (20,151)	51,896 (20,897)	9,047 (3,123)	203,882 (150,199)

❖ KorLex & 표준국어대사전 mapping

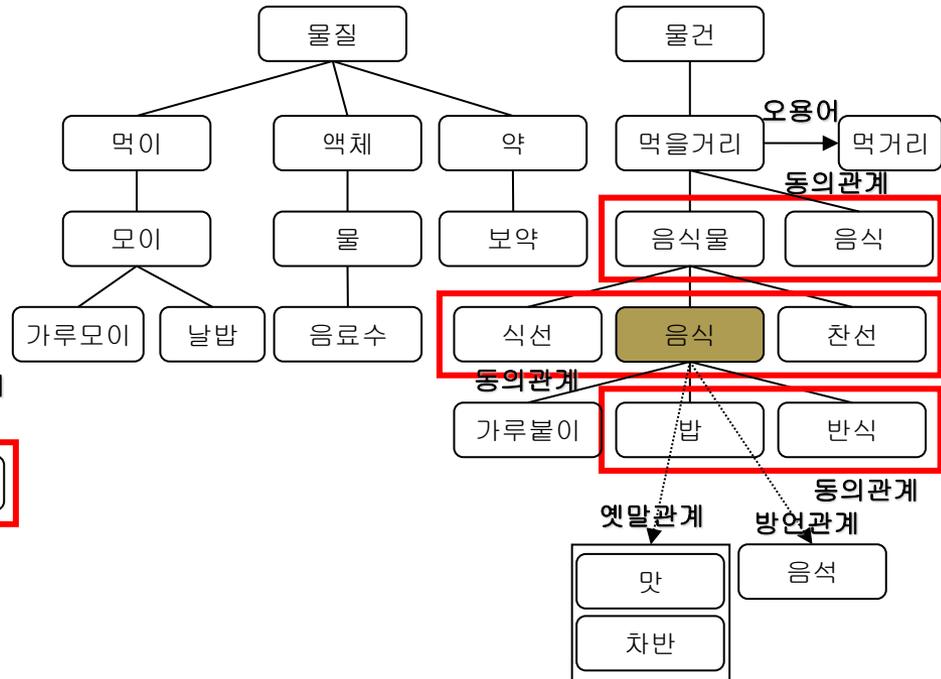
정의문 연동	Korean WordNet 정의문 연동 정보				
	명사	동사	형용사	부사	계
표준국어 대사전	74,446 (67,938)	12,225 (9,635)	19,274 (17,639)	3474 (2,913)	109,419 (99,291)
기타 사전	363	226	2,540	73	3,202



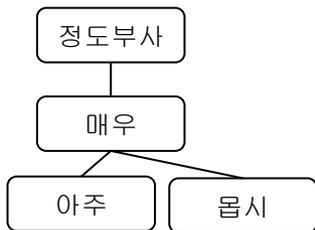
## U-WIN의 용언어휘망



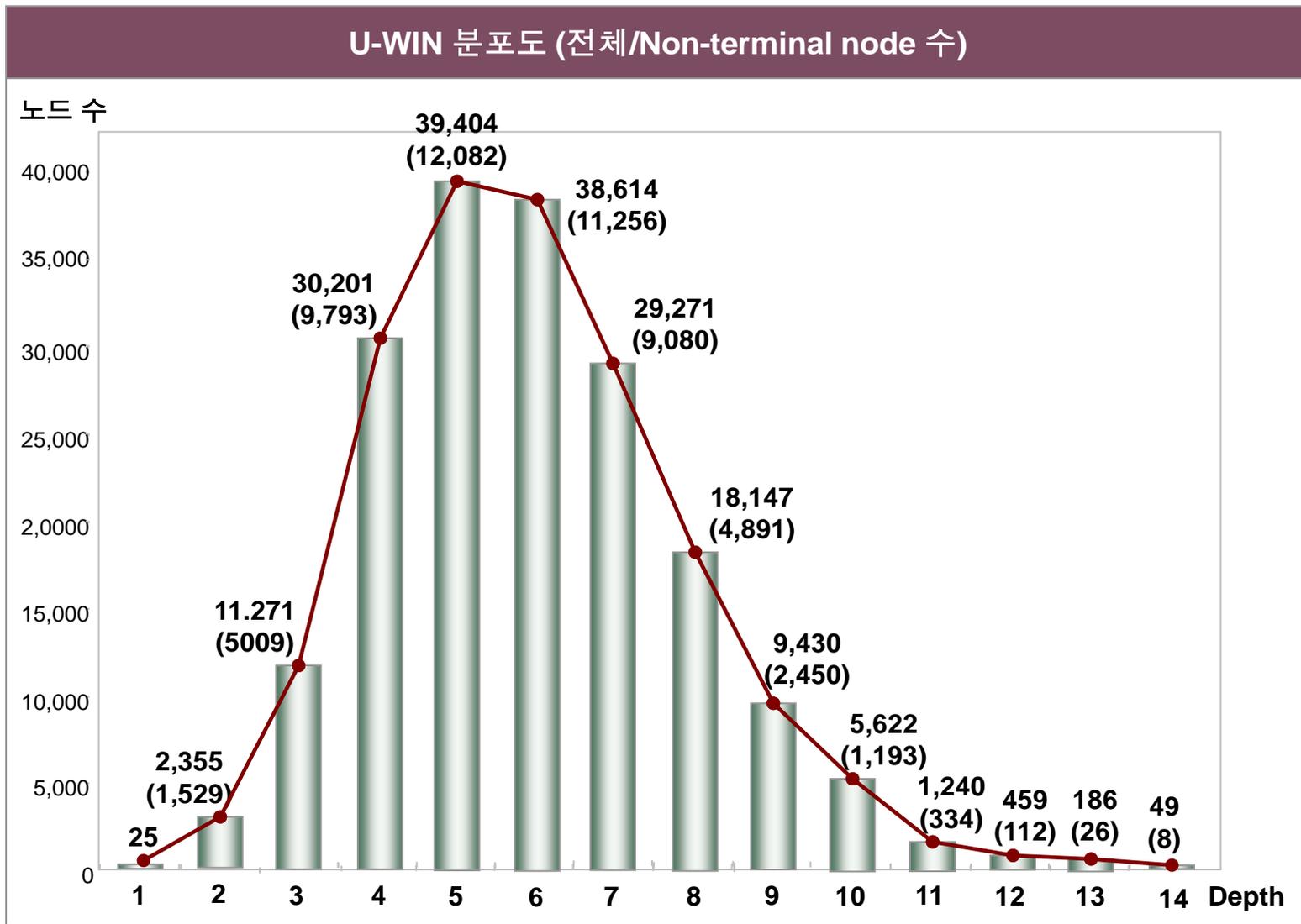
## U-WIN의 명사어휘망



## U-WIN의 부사어휘망



- U-WIN ver.1.0(2002 ~ 2007): 160,000 어휘 (수작업)
- 표준국어대사전 뜻풀이/용례 다의어 태깅 (2008 ~2010)
- U-WIN ver.1.9 : 420,000 어휘 (다의어 태깅 사전, 2010)



## ❖ 어휘 의미 관계 (다의어 수준)

품사	동의어 (=, ≍)	반대말	비슷한 말	준말	본말	낮춤말	높임말	참고 어휘	합계
명사	114,715	4,513	13,583	496	20	382	378	28,474	162,561
동사	12,702	1,474	10,942	0	141	6	21	3,165	28,451
형용사	1,463	443	1,317	170	46	3	1	2,451	5,894
부사	2,284	68	393	842	117	1	0	11,722	15,427
합계	131,164	6,498	26,235	1,508	324	392	400	45,812	212,333



# 뜻풀이 및 용례 의미 태깅 (1)

## ❖ 표준국어대사전 뜻풀이 태깅

- 다의어 수준 뜻풀이 : 587,833
- 다의어/동형이의어 태깅 대상 어휘 : 107,306
- 뜻풀이 태깅 어휘 수 : 3,610,106

## ❖ 표준국어대사전 용례 태깅

- 다의어/동형이의어 태깅 대상 어휘 : 78,083
- 용례 태깅 어휘 수 : 1,121,759

## ❖ 작업기간 : 2008 ~ 2010

## ❖ 사전(어휘지식) 해독하는 컴퓨터

- 동형이의어/다의어 WSD
- 의미역 triple 추출
- 개념화 => 어휘지도

38	가격_010000	5	1
39	가격_020000	1	
40	가격_030000	679	154
41	가결_010000	6	2
42	가결_030001	2	
43	가결_030002	1	
45	가결의_000001	2	
46	가결의_000002	2	
48	가정_040002	1	1
49	가정_080000	1	
50	가계_010001	1	
51	가계_010002	1	
52	가계_030000	2	1
53	가계_040000	1	
54	가계_050000	1	
55	가계_060000	23	3
56	가계_080001	21	13
57	가계_080002	1	1
58	가계_100000	1	
59	가곡_010001	101	1
60	가곡_010002	65	8
61	가공_010001	493	15
62	가공_010002	2	2
63	가공_020000	1	
64	가공_040001	12	
65	가공_040002	2	2
70	가공하다_010001	404	13
71	가공하다_030000	1	7
72	가과_010000	1	

# 뜻풀이 및 용례 의미 태깅 (2)

표준국어대사전 의미태깅 [작업장] - Windows Internet Explorer

http://nlplab.ulsan.ac.kr:5900/tagging\_def/default.aspx

파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)

☆ UWIN 표준국어대사... 국립국어원 표... 국립국어원:어...

작업자명 : 옥철영    작업날짜 : 2009-09-13    수정창    작업량 확인    로그아웃

동음이의어 리스트    시작페이지    현재페이지    마지막페이지    정보란

페이지번호/동음이의어    1    1    2

전체선택

하다\_\_001022+어 대다\_\_001019+는 소리\_\_001003

누르다\_\_001001+어 대다\_\_001003+는 나무\_\_001002

흐르다\_\_000001+어 대다\_\_001019+는 흐사위

회오리바람 [명사] {예술}

전라도 지방의 무당춤에서, 팔을 성난 파도처럼 흔들며 대는 춤사위.

전라도\_\_000002 지방\_\_005001+의 무당춤\_\_000001+에서 +, 팔\_\_001001+을 성난다\_\_000001+나 파도\_\_000001+처럼 흔들다\_\_000001+어 대다\_\_001019+는 춤사위

끝다\_\_000008+어다 대다\_\_001014+모+

끝다\_\_000008+어다 대다\_\_001015+는 모양\_\_002003

끝다\_\_000009+어다 대다\_\_001011+는 도량\_\_001000

검색    동음이의어만 검색     저장시 자동 페이지 넘기

대다\_999\_999 : 뜻풀이가 없는 경우 (고유명사:인명, 지명, 나라명 등) 선택  
 대다\_888\_888 : 뜻풀이가 없는 경우 (일반명사) 선택  
 대다\_777\_777 : 형태소 분석 오류

대다\_001\_001 [명사] [ ] : 전해진 시간에 맞거나 맞춘다.  
 대다\_001\_002 [명사] [ ] : 어떤 것을 목표로 삼거나 향하다.  
 대다\_001\_003 [명사] [ ] : 무엇에 어디에 닿게 하다.  
 대다\_001\_004 [명사] [ ] : 어떤 도구나 물건을 써서 일을 하다.  
 대다\_001\_005 [명사] [ ] : 차, 배 따위의 앞부분을 멈추어 서게 하다.  
 대다\_001\_006 [명사] [ ] : 차, 배 따위의 앞부분을 마련하여 주다.  
 대다\_001\_007 [명사] [ ] : 무엇이나 어떤 대거나 귀에 맞춘다.  
 대다\_001\_008 [명사] [ ] : 어떤 것을 목표로 하여 총, 호스 따위를 겨냥하다.  
 대다\_001\_009 [명사] [ ] : 내기 따위에서 돈이나 물건을 걸다.  
 대다\_001\_010 [명사] [ ] : 사람에게 구해서 소개해 주다.  
 대다\_001\_011 [명사] [ ] : 어떤 곳에 물을 끌어 들이다.  
 대다\_001\_012 [명사] [ ] : 사람에게 하거나 인양하다.  
 대다\_001\_013 [명사] [ ] : 다른 사람과 신체의 일부분을 닿게 하다.  
 대다\_001\_014 [명사] [ ] : 서양과 중국을 비교하다.  
 대다\_001\_015 [명사] [ ] : 이구나 구슬을 들어 보이다.

\* 뜻풀이가 없는 경우 고유명사인 경우와 아닌 경우를 나누어서 태깅하여 주십시오  
 \* 형태소 분석이 잘못된 경우 '형태소 분석 오류'를 선택해 주십시오

적용    저장

**뜻풀이**  
 어떤 도구나 물건을 써서 일을 하다.

**용례**  
 그림에 붓을 대다. 그는 기계에 공구를 대고 무언가를 열심히 고치고 있다. 아무리 급해도 어른보 먼저 음식에 숟가락을 대는 게 아니다.

페이지에 오류가 있습니다.    인터넷    100%

# 뜻풀이 및 용례 의미 태깅 (3)



제목 없음 - Windows Internet Explorer  
 http://nlplab.ulsan.ac.kr:5900/tagging\_def/Manager/Edit.aspx

표제어:  KeyID:

표제어 검색:  KeyID 검색:

명사 검색:

KeyID:  13180002  먹다  수정

뜻풀이

음식 따위를 입을 통하여 배 속에 들여보낸다.

음식\_000001/NNG 따위\_000001/NNB+을/JKO 입\_000001/NNG+을/JKO 통하다\_000702/VV+아/EC 배\_010001/NNG 속\_010002/NNG+에/JKB 들여보내다\_000101/VV+다/EF+./SF

<font color=red>음식\_000001</FONT> <font color=red>따위\_000001</FONT>+을 <font color=red>입\_000001</FONT>+을 <font color=red>통하다\_000702</FONT>+아 <font color=red>배\_010001</FONT> <font color=red>속\_010002</FONT>+에 <font color=red>들여보내다\_000101</FONT>+다+.

용례

밥을 먹다. 술을 먹다. 약을 먹다. 물을 먹다. 음식을 배불리 먹다. 닭이 모이를 먹다. 몸이 약해진 누나는 보약을 몇 차례나 먹어도 늘 골골거렸다.

밥\_010002/NNG+을/JKO 먹다\_020101/VV+다/EF+./SF 술\_010000/NNG+을/JKO 먹다\_020101/VV+다/EF+./SF 약\_070001/NNG+을/JKO 먹다\_020101/VV+다/EF+./SF 물\_010001/NNG+을/JKO 먹다\_020101/VV+다/EF+./SF 음식\_000001/NNG+을/JKO 배불리/MAG 먹다\_020101/VV+다/EF+./SF 닭/NNG+이/JKS 모이\_010000/NNG+을/JKO 먹다\_020101/VV+다/EF+./SF 몸\_010001/NNG+이/JKS 약하다\_010102/VA+아/EC지다\_040103/VX+L/ETM 누나\_010001/NNG+는/JX 보약/NNG+을/JKO 몇\_000101/MM 차례\_010003/NNG+나/JX 먹다\_020101/VV+다/EF+./SF

<font color=red>밥\_010002</FONT>+을 <font color=red>먹다\_020101</FONT>+다+. <font color=red>술\_010000</FONT>+을 <font color=red>먹다\_020101</FONT>+다+. <font color=red>약\_070001</FONT>+을 <font color=red>먹다\_020101</FONT>+다+. <font color=red>물\_010001</FONT>+을 <font color=red>먹다\_020101</FONT>+다+. <font color=red>음식\_000001</FONT>+을 배불리 <font color=red>먹다\_020101</FONT>+다+. 닭+이 <font color=red>몸\_010001</FONT>+을 <font color=red>약하다\_010102</FONT>+아 <font color=red>몇\_000101</FONT>+을 <font color=red>차례\_010003</FONT>+나 <font color=red>먹다\_020101</FONT>+다+.

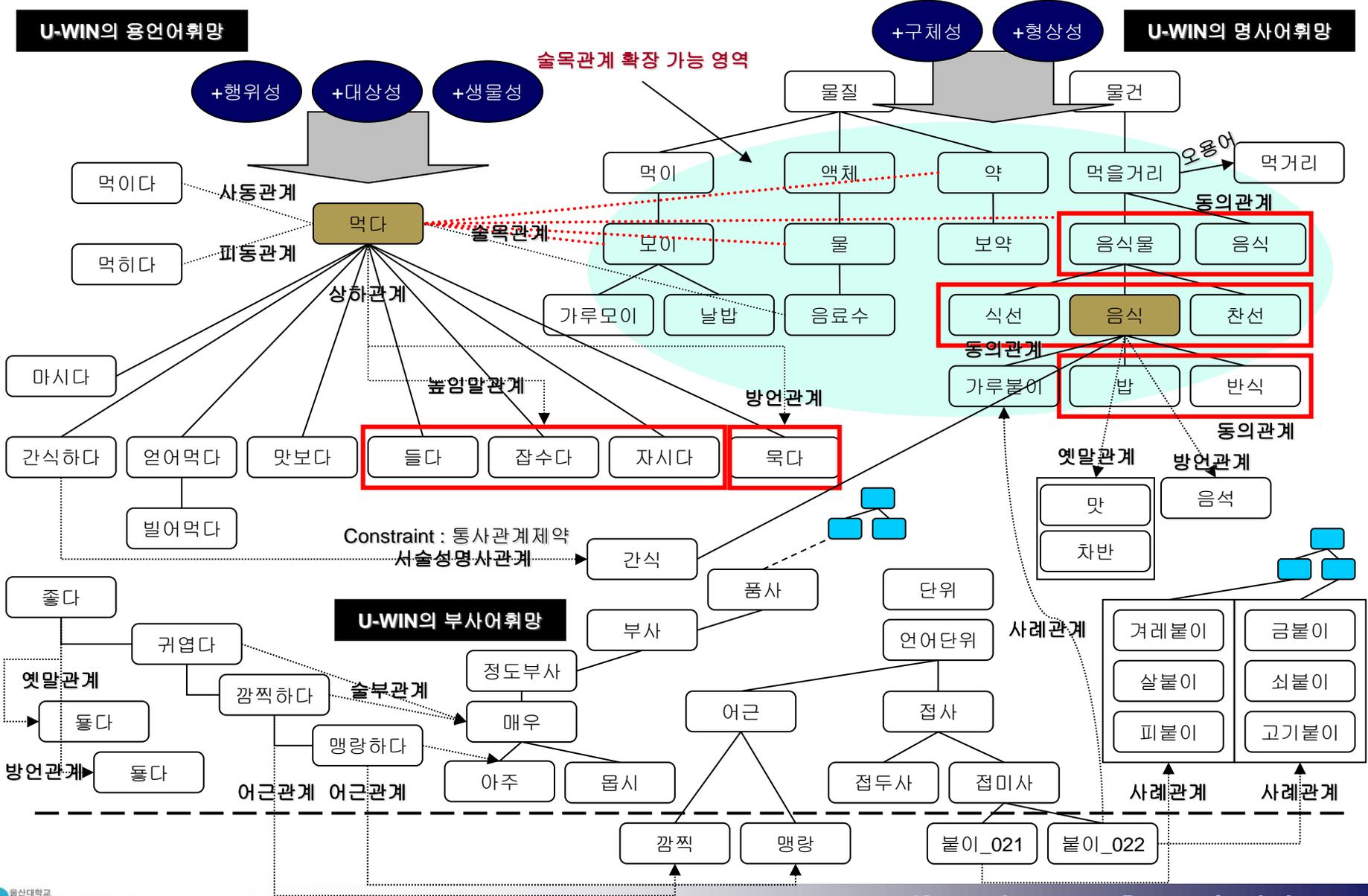
13179900 (01\_00\_00) : 귀나 코가 막혀서 제 기능을 하지 못함  
 13180002 (02\_01\_01) : 음식 따위를 입을 통하여 배 속에 들여보낸다.  
 13180003 (02\_01\_02) : 담배나 마편 따위를 피우다.  
 13180004 (02\_01\_03) : 연기나 가스 따위를 들이마시다.  
 13180005 (02\_01\_04) : 어떤 마음이나 감정을 품다.  
 13180006 (02\_01\_05) : 일정한 나이에 이르거나 나이를 더하다.  
 13180007 (02\_01\_06) : 욕, 핀잔 따위를 듣거나 당하다.  
 13180008 (02\_01\_07) : (속되게) 뇌물을 받아 가지다.  
 13180009 (02\_01\_08) : 수익이나 이문을 차지하여 가지다.  
 13180010 (02\_01\_09) : 물이나 습기 따위를 빨아들이다.  
 13180011 (02\_01\_10) : 어떤 등급을 차지하거나 점수를 따다.  
 13180012 (02\_01\_11) : 구기 경기에서, 점수를 잃다.  
 13180013 (02\_01\_12) : (속되게) 여자의 정조를 유린하다.  
 13180014 (02\_01\_13) : 매 따위를 맞다.  
 13180015 (02\_01\_14) : 남의 재물을 다루거나 맡은 사람이 그 재물을 잃다.  
 13180017 (02\_02\_01) : 날이 있는 도구가 소재를 깎거나 자르거나  
 13180018 (02\_02\_02) : 바르는 물질이 배어들거나 고루 퍼지다.  
 13180019 (02\_02\_03) : 벌레, 균 따위가 파 들어가거나 퍼지다.  
 13180020 (02\_02\_04) : 손이나 물자 따위가 들거나 쓰이다.  
 13180021 (02\_03\_00) : 앞말이 뜻하는 행동을 강조하는 말 주사어  
 44059415 (02\_01\_15) : 겁, 충격 따위를 느끼게 된다.

완료  신뢰할 수 있는 사이트 105%





# 한국어 어휘지도 (UWordMap) (1)





# 한국어 어휘지도 (UWordMap) (2)



품사	표준국어 대사전	U-WIN (계층관계)	UWordMap (Exo Brain Project)		
			1차년도 (14.03.24)	2차년도 (15.01.20)	3차년도 (15.06.19)
명사	377,281	365,774	LCS 72,020	97,410	98,264
동사	90,237	73,694	29,345	46,410	51,336
형용사	21,618	16,853	4,653	7,709	12,438
부사	25,178	17,697	6,186	6,187	6,187
합계	514,314	474,018	123,823	157,718	168,225



# 한국어 어휘지도 Browser (1)



한국어 어휘 지도 v2.0

검색어 목록 | 표제어 목록 | 계층 구조

표제어 검색

KLPLAB 울산대학교 한국어처리연구실

계층 구조

- 음언(사태 부류)
  - 사태
    - 행위
      - 물리적행위
        - 단독행위
          - 설위행위
            - 먹다\_02\_01\_01
              - 내먹다\_02\_00\_00

프로그램 정보 | 관련 정보

어휘 정보 | 설정 정보

어휘관계	유/무	색깔
상위어	<input checked="" type="checkbox"/>	■
하위어	<input checked="" type="checkbox"/>	■
반의어	<input checked="" type="checkbox"/>	■
동의어	<input checked="" type="checkbox"/>	■
~이	<input checked="" type="checkbox"/>	■
~이가	<input checked="" type="checkbox"/>	■
~을	<input checked="" type="checkbox"/>	■
~에	<input checked="" type="checkbox"/>	■

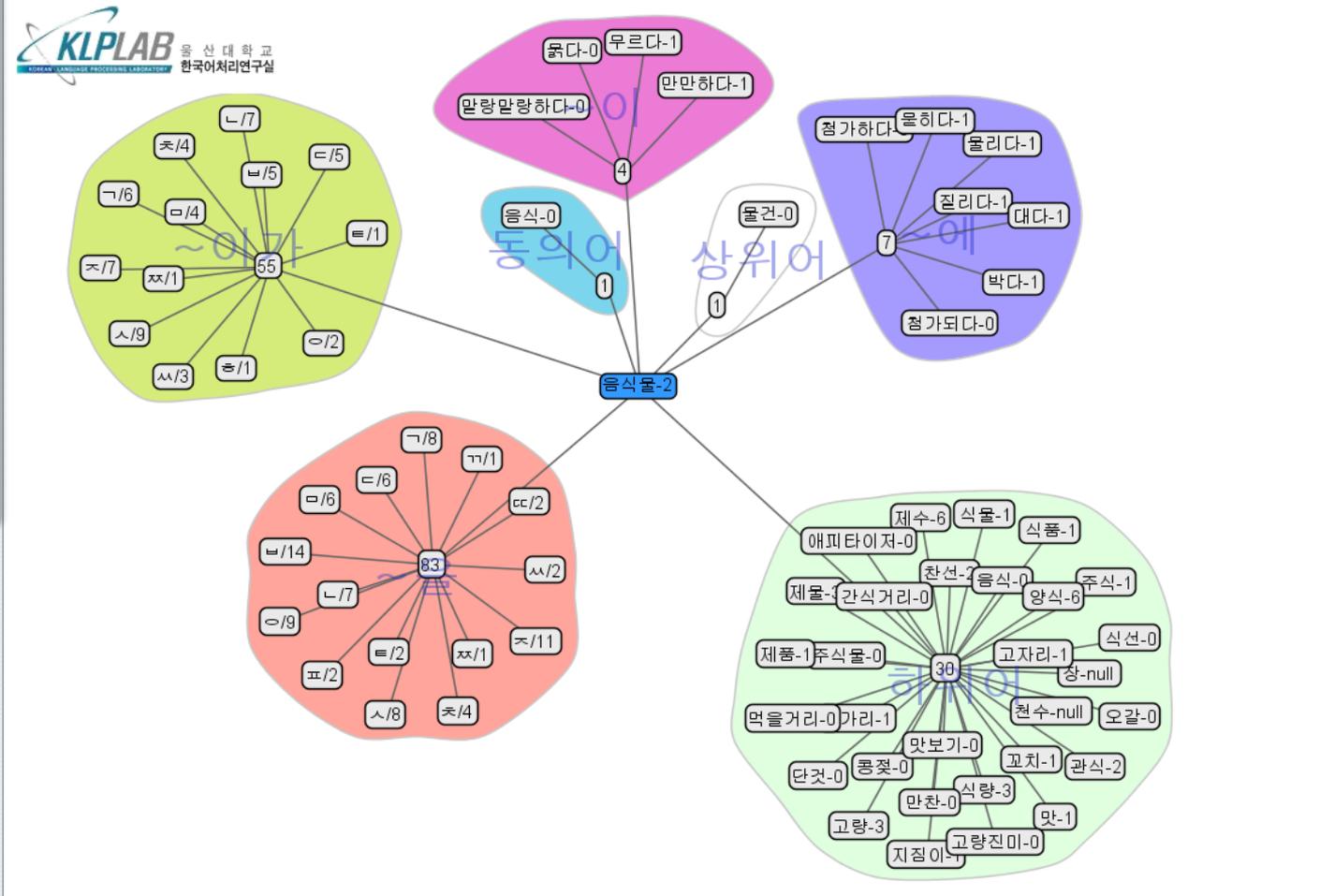
node 개수 30  
history 개수 10  
전체선택 전체해제 완료 취소

# 한국어 어휘지도 Browser (2)

한국어 어휘 지도 v2.0
표제어 검색
음식물\_02\_00\_00

계층 구조

- 음식물\_02\_00\_00
  - 간식거리\_00\_00\_00
  - 결밭\_00\_00\_00
  - 결들이\_00\_00\_01
  - 고량진미\_00\_00\_00
  - 공수\_10\_00\_00
  - 관식\_02\_00\_01
  - 구식\_02\_00\_00
  - 꼬치\_01\_00\_01
  - 녹미\_02\_00\_02
  - 단것\_00\_00\_00
  - 달걀\_00\_00\_00
  - 돈제우주\_00\_00\_00
  - 락식\_00\_00\_00
  - 육진해미\_00\_00\_00
  - 막말미\_00\_00\_00
  - 만나\_01\_00\_00



음식물-2

프로그램 정보		관련 정보	
어휘 정보		설정 정보	
어휘관계	유/무	색깔	
상위어	<input checked="" type="checkbox"/>	<span style="background-color: #90EE90; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
하위어	<input checked="" type="checkbox"/>	<span style="background-color: #CCCCFF; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
반의어	<input checked="" type="checkbox"/>	<span style="background-color: #00BFFF; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
동의어	<input checked="" type="checkbox"/>	<span style="background-color: #FF00FF; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
~이	<input checked="" type="checkbox"/>	<span style="background-color: #90EE90; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
~이가	<input checked="" type="checkbox"/>	<span style="background-color: #FF4500; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
~을	<input checked="" type="checkbox"/>	<span style="background-color: #32CD32; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
~에	<input checked="" type="checkbox"/>	<span style="background-color: #4169E1; border: 1px solid black; display: inline-block; width: 15px; height: 10px;"></span>	
node 개수	30		
history 개수	10		
전체선택	<input type="checkbox"/>	전체해제	<input type="checkbox"/>
	<input type="checkbox"/>	완료	<input type="checkbox"/>
	<input type="checkbox"/>	취소	<input type="checkbox"/>

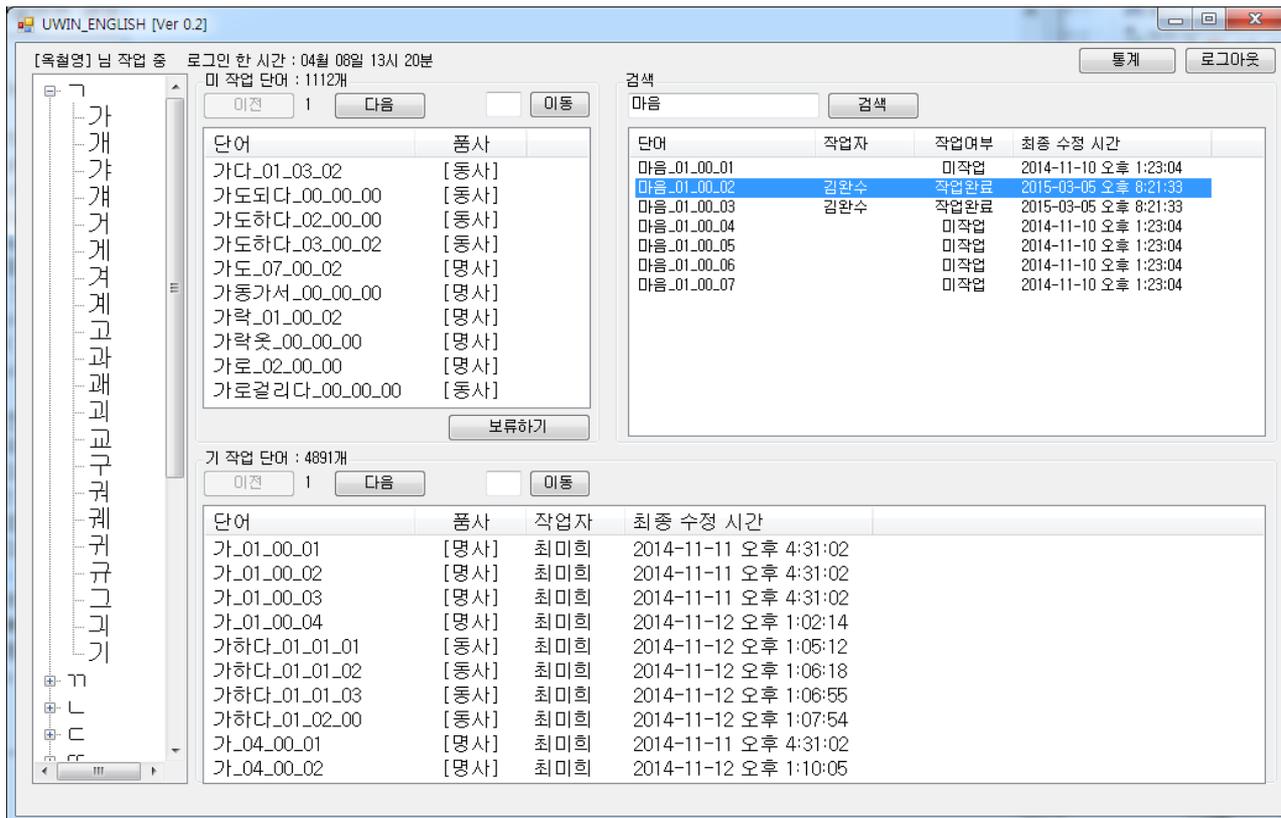


함수명	설명
GetPS	(단어 or 동형)의 다의어 받아오기
GetHyperWord	단어의 1레벨 위의 상위어 받아오기
GetHyperAllWord	다의어의 모든 상위어 받아오기
GetHypoWord	다의어의 1레벨 아래의 하위어 받아오기
GetNRelV	해당 용언과 논항(격조사)으로 관련된 명사 받아오기
GetVRelN	해당 명사를 논항(격조사)으로 가지는 용언 받아오기
GetRelSubCt	용언과 명사가 연결된 논항(격조사) 받아오기
GetSynSet	다의어의 동의어 받아오기
GetAntSet	다의어의 반의어 받아오기
GetDistance	다의어1과 다의어2의 거리 받아오기

- ❖ 부사 기준 API 추가 예정(2015)
- ❖ 어휘간 유사도 측정 (다양한 option 제공)

# Mapping UWordMap to WordNet(1)

- ❖ 표준국어대사전 다의어 의미를 WordNet Synset으로의 mapping
  - 전문용어, 방언 제외한 239,000 다의어 대상
  - KorLex-표준국어대사전 mapping 정보 활용(약 100,000 entry)
  - 다의어별 의미에 맞는 대역어(NAVER 사전) – WordNet synset
- ❖ Mapping 도구



# Mapping UWordMap to WordNet(2)

## ❖ 다의어별 의미에 맞는 대역어(NAVER 사전) – WordNet synset

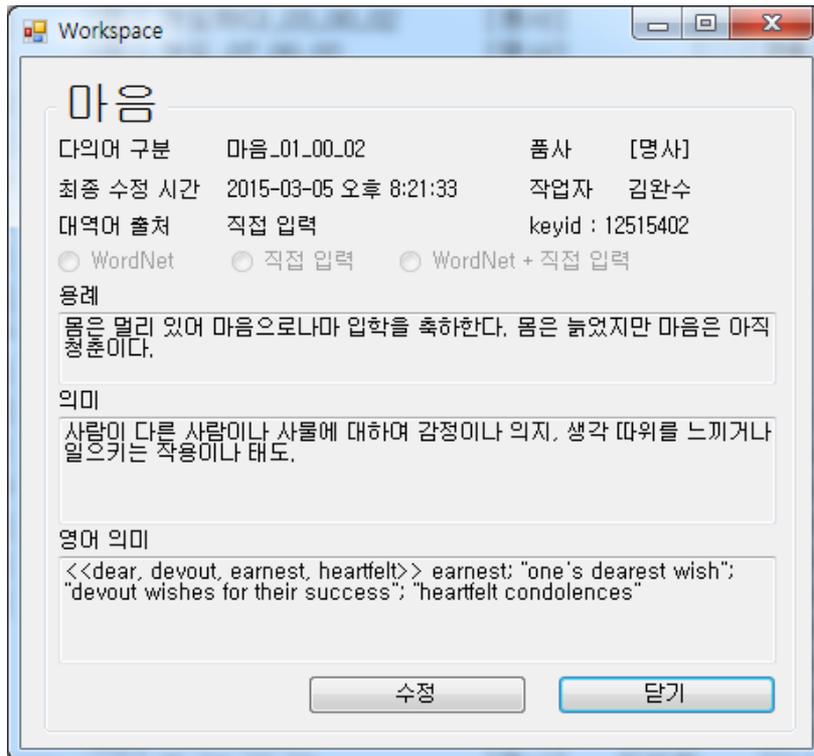
The image shows three screenshots illustrating the mapping of Korean words to WordNet synsets. The left screenshot shows a Korean dictionary entry for '사랑' (love) with various sub-entries and a 'Workspace' window. The middle screenshot shows a Naver search page for '사랑' with a dropdown menu and search results. The right screenshot shows a WordNet search page for 'love' with a list of synsets and their descriptions.

# Mapping UWordMap to WordNet(3)

## ❖ Mapping 도구 기능

- Mapping 방법: WordNet, 직접입력, WordNet+직접입력
- 작업자별 통계
- 보류
- 검색

(2015.08.18 기준)



**Workspace**

**마음**

대역어 구분	마음_01_00_02	품사	[명사]
최종 수정 시간	2015-03-05 오후 8:21:33	작업자	김완수
대역어 출처	직접 입력	keyid	: 12515402

WordNet   
  직접 입력   
  WordNet + 직접 입력

**용례**

몸은 멀리 있어 마음으로나마 입학을 축하한다. 몸은 늙었지만 마음은 아직 청춘이다.

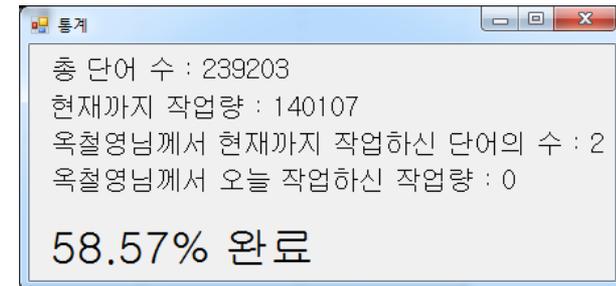
**의미**

사람이 다른 사람이나 사물에 대하여 감정이나 의지, 생각 따위를 느끼거나 일으키는 작용이나 태도.

**영어 의미**

<<dear, devout, earnest, heartfelt>> earnest: "one's dearest wish"; "devout wishes for their success"; "heartfelt condolences"

수정      닫기



**통계**

총 단어 수 : 239203  
 현재까지 작업량 : 140107  
 옥철영님께서 현재까지 작업하신 단어의 수 : 2  
 옥철영님께서 오늘 작업하신 작업량 : 0

**58.57% 완료**

# 의미역 (Semantic Role) (1)

## ❖ 세종 의미역 15개

### ❖ PropBank

- ArgN 최대 6개
- ArgM 13개

### ❖ ExoBrain 22개 (2013 결정)

- 격조사 의미분석
- 남승호 기준(16개) + 추가 (6개)
  - 시간격
  - MANNER
  - 목적
  - 내용
  - 정도
  - 재료

세종전자사전	PropBank	Exo-Brain
행위주	Agent	행동주
경험주		경험주
심리경험주		경험주
동반주		동반주
대상격	Theme	대상
처소격	LOC	처소
방향격	DIR	방향
도착격		착점
결과상태		착점
출발격		기점
도구격	INS	도구
영향주	Patient	피동주
기준치		비교기준
목적		목적
내용		내용
...	...	...
		자격
	TMP	시간
	MNR	방법
15개		22개



# 의미역 (Semantic Role) (2)



의미역	이,가, 는,은	께서, (이)라서, 서	을, 를, 르	더러, 보고	에	에게, 한테	에게로, 에게를, 한테로	에서	에게서, 에서부터, (으)로부터, 한테서	에다가	에를, 에의, 에야	(으)로	(으) 로서	(으) 로써	보다, 처럼, 마따나	같이	와,과, 하고, (이)랑	고, (이)라 고
행동주 AGT	o	o			o	o		o				o						
대상 THM	o		o		o													
경험주 EXP	o	o	o															
피동주 PAT	o	o	o	o		o												
자극 STM	o		o		o							o						
처소 LOC	o		o		o			o		o								
기점 SRC	o		o		o	o		o	o	o			o					
착점 GOL	o		o	o	o	o	o				o	o						
원인 CAU	o				o			o				o						
수혜자 BEN	o		o	o	o	o	o											
경로 ROU			o									o						
방향 DIR			o									o						
목적 PUR			o															
정도 DGR			o															
재료 MAT			o									o		o				
도구 INS			o		o							o		o				
시간 TMP					o							o		o		o		
비교기준 CRT					o			o							o	o	o	
동반주 COM												o					o	
내용 CNT												o						o
자격 ROL					o							o	o					
방법 MNR												o						

# 의미역 (Semantic Role) (3)

- ❖ UPropBank : 90,090개(다의어, 동사, 형용사)의 필수논항에 의미역 부여
  - 문형이 있는 용언: 41,518 (2013년)
  - 문형이 없는 용언(자동사, 형용사) : 48,572 (2014년)

표제어	문형	뜻풀이	용례	의미역 부착 격틀
감금되다	◁ ...에 ▷	감금.	감옥에 감금되다. 독방에 감금되다. 저녁이 되자 두 죄인은 좁고 어두운 방에 감금되었다.	X:행동주 Z:처소-에
감금하다	◁ ...을 ...에 ▷	감금.	자식 놀음 공간에 감금해 놓고 아비 꼴이 어쩐가 ... 사장은 ... 운전수를 이상한 창고 같은 방 속에 감금해 버리고는 혼자 그 집 안으로 사라져 버린다.	X:행동주 Y:대상-을 Z:처소-에
감급되다	◁ ...으로 ▷	감급.		X:대상 Z:정도-으로, 착점
감급하다	◁ ...을 ...으로 ▷	감급.	직원들 봉급을 100만 원에서 90만 원으로 감급했다.	X:행동주 Y:대상-을 Z:착점-으로
감기다 020101	◁ ...에 ▷	'감다03[1](1)'의 피동사.	왕거미의 거미줄에 풍뎡이가 감기듯이 그가 그 검은 노끈에 감기고 있었다. 아이들은 뛰어가면서 발목에 감기는 개울물의 감촉을 미리부터 즐긴다.	X:피동주 Z:도구-에
감기다 020102	◁ ...에 ▷	옷 따위가 몸을 친친 감듯 달라붙다.	젖은 치맛자락이 맨살에 감기는 듯하다. 이불 속에 감겨 있던 몸이 훌쩍 문밖으로 나서니 추위는 살을 에는 듯하였다.	X:대상 Z:착점-에
감기다 020103	◁ ...에 ▷	음식 따위가 감칠맛이 있게 착착 달라붙다.	며칠 같이 유하면서 송도 특유의 맛깔스러운 음식과 집에서 담근 허에 감기는 약주술로 송별을 겸해 회포를 풀 기회를 갖고 싶어서 청해 본 거였다. 큰아들이 허에 착착 감기는 조청이라면 작은아들은 목구멍에 걸린 가시였다.	X:대상 Z:착점-에
감기다 020104	◁ ...에 ▷	사람이나 동물이 달라붙어서 떠나지 아니하다.	강아지가 내 옆에 감겨서 꿈쩍도 않는다. 손자가 할머니 다리에 감겨 떨어질 줄을 모른다.	X:대상 Z:착점-에
감기다 020105	◁ ...에 ▷	음식을 너무 먹어 몸을 가누지 못하다.	유복이가 여러 날 변변히 먹지 못하고 굶주린 끝에 배불리 먹고 음식에 감기어서 길 갈 기운이 없어졌다.	X:피동주 Z:원인-에
감기다 030000	◁ ...을 ▷	'감다01(1)'의 사동사.	눈을 감기다.	X:행동주 Y:대상-을
감기다 040000	◁ ...을 ▷	'감다02'의 사동사.	머리를 감기다. 어머니는 할머니의 손톱과 발톱도 깎아 드리고 머리도 감겨 드렸다. 어머니는 창포물에 언니의 머리를 감기면서, 여자란 머릿결이 고와야 용모가 아름다운 것이라고 말씀하였다.	X:행동주 Y:대상-을
감기다 050000	◁ ...에게 ...을 ▷	'감다03[1]'의 사동사.	어머니는 아들에게 형클어진 실을 감겼다.	X:행동주 Z:피동주 -에게 Y:대상-을

# 한국어 의미처리시스템 (UTagger)

- ✓ 동형이의어 WSD (UTagger-HG)
- ✓ 의존관계 분석 (UTagger-DP)
- ✓ 다의어 WSD (UTagger-PS)
- ✓ 의미역 태깅시스템 (UTagger-SR)
- ✓ 개체명 인식기 (UTagger-NE)

# 의미처리(어휘 WSD) 필요성

- ❖ 형태소 분석 오류 : 문맥 반영 못함
  - 과도한 음주로 인해 **간이** 좋지 않다/ **간이** 식당에서 요기를 했다
  - 논이 **걸어서** 벼가 잘 자란다/ 옷을 **걸어서** 보관하다/ **걸어서** 하늘까지
  - 그는 **칠** 주야를 걸어 서울에 왔다/ 헤엄을 **칠** 줄 아는 ...

- ❖ 구문/의존관계 분석 시 Case frame (용언) 제공
  - 주한 미군 **철수** 맨 일 재무장 초래
  - 중국과 깊이 **관계**
  - 너 기저귀 **차고** 놀 때 나는 공 **차고** 놀았다

**차다02** [전체 보기](#) [차, 차니]

「동사」  
[...을]

「1」 발로 내어 지르거나 발아 올린다.  
「2」 발을 힘껏 뻗어 사람을 치다.  
「3」 허끝을 입천장 앞쪽에 붙였다가 떼어 소리를 내다.  
「4」 발로 힘 있게 밀어젖히다.  
「5」 (속되게) 주로 남녀 관계에서 일방적으로 관계를 끊다.  
「6」 날쌔게 빼앗거나 움켜 가지다.

**차다03** [전체 보기](#) [차, 차니]

「동사」  
[1] [...에 ...을]

「1」 물건을 몸의 한 부분에 달아매거나 끼워서 지니다.  
「2」 수갑이나 차고 따위를 팔목이나 발목에 끼우다.  
[2] [...을]

(속되게) 매인으로 삼아 데리고 다니다.

- ❖ 전문용어/개체명, 합성어, 파생어의 의미 분석
  - “한국전자통신연구원”, “한국 전자통신 연구원”
  - **전기**기록 : 前期+기록, 電氣+기록, 轉機+기록
  - 내각**제**(制), 추모**제**(祭), 미국**제**(製), 위염**제**(劑)

❖ 표준국어대사전 표제어 507,100 개 중 125,152 개 동형이의어(25%)

# UTagger 특징 (1)

## ❖ 학습말뭉치 기반

- 세종형태의미부착 말뭉치 : 11,116,320 어절(339파일)
- 세종원시말뭉치 태깅 : 52,338,450 어절(1,770 파일)

## ❖ 학습사전

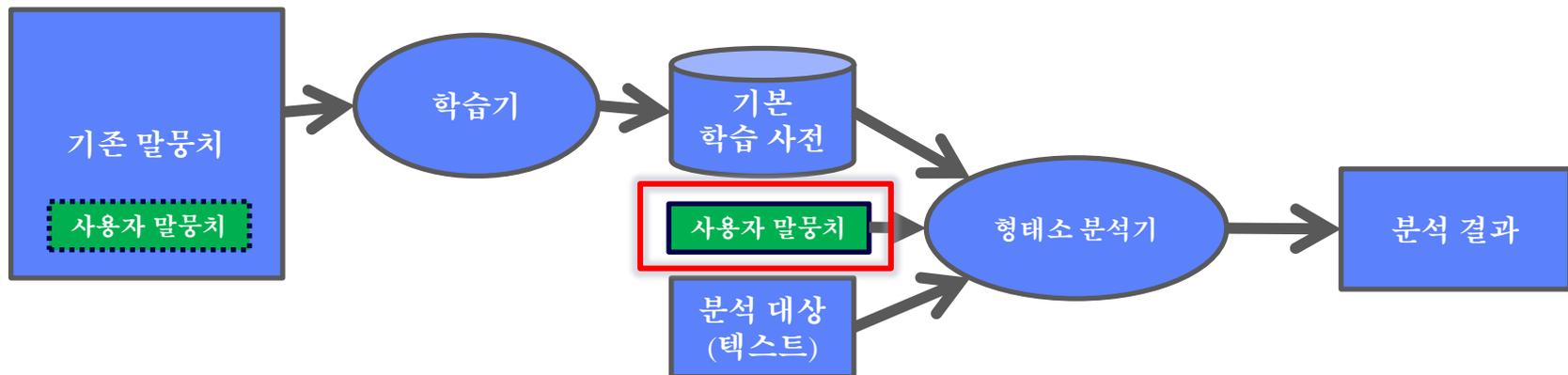
- 자체 개발 파일DB 사용
- CKMA : 형태소분석용, 인접형태소/품사 전이확률
- 품사/동형이의어 태깅용, 인접 두 어절 간의 전이 : biAF, biEF, biFF

## ❖ CKMA (Corpus-based Korean Morphology Analysis)

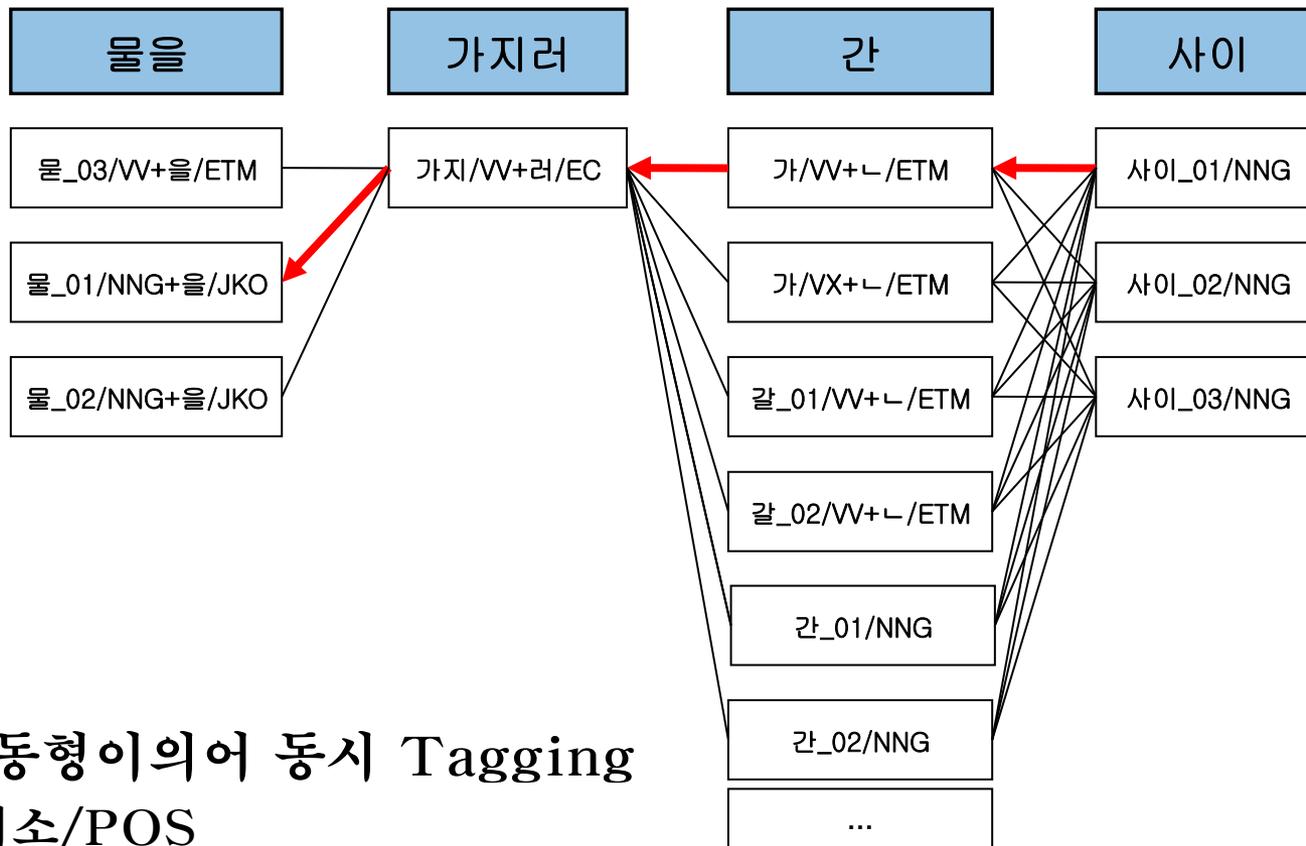
- 형태소분석 : 기분석 어절의 형태소간 전이 확률 이용 (재현율 99.9%)
- 불규칙용언에 대한 분석/변형/복원 규칙 없음
- 체언류/용언류 형태소분석 루틴 불필요
- 띄어쓰기 오류 어절 분석(주민들사이에, 그리기시작했으며, 박전대표위원)

# UTagger 특징 (2)

- ❖ 품사/동형이의어 태깅
  - 형태의미 주석말뭉치에서 bigram 간의 전이 확률 학습(단계별 적용)
  - HMM 모델(2012)
  - SCP(부분어절 조건부확률) 모델(2013)
  - 복합명사/접미사 동형이의어 태깅
- ❖ 미등록어/신조어 추정
  - CKMA 기반
  - 사용자사전 지원
- ❖ 사용자말뭉치 문맥 학습(incremental learning)



# 단계별 전이 모델 & Viterbi



## ❖ POS/동형이의어 동시 Tagging

- 형태소/POS
- biAA : 앞서질 전체 형태소, 뒤어질 전체 형태소
- biAD : 앞서질 전체 형태소, 뒤어질 첫 형태소
- biLD : 앞서질 마지막 형태소, 뒤어질 첫 형태소
- 논문 참조 : 정보과학회논문지 39/5, 39/11 (2012)

# UTagger option (대화형 사용자용)

**설정**

학습사전용 메모리 사용량  확인

---

분석 파일  
 태깅 말뭉치  원시 말뭉치  사용자 사전 취소

서술형 명사 용언 분석  
 어간형  어근형 품사 태그 세트  
 세종  ETRI

추가 출력 형식  
 없음  한 줄 한 어절

형태소만  
 형태소만

동형미의어 분별정보  
 출력 없음  의미분별(한자 사전뜻풀이 일부)  
 의미범주(명사어휘망)

한자 변환 옵션  
 한자 변환 안함  한자 변환  한글(한자) 병기

한자 급수  
 8급  7급  6급  5급  4급  3급  2급  1급

복합어 분해  
 분해안함  사전기반  사전기반+알고리즘  
음절수(3미상적용)   
구성명사최소빈도(1~5권장)

입력 한자 처리  
 한자 그대로  한글로 변환

사용자 사전 내용

샤아머니즘  
샤아머니즘/NNG  
졸  
졸/IC  
난센  
난센/NNP  
무와탈리슈  
무와탈리슈/NNP  
진카쿠지  
진카쿠지/NNP  
체르마트  
체르마트/NNP  
생모리츠  
생모리츠/NNP  
슬레이  
슬레이/NNP  
하호마  
하호마/NNP  
크사바루나  
크사바루나/NNP  
아나히타  
아나히타/NNP  
마코마나  
마코마나/NNP  
수스키노  
수스키노/NNP  
합스부르크  
합스부르크/NNP  
그리신  
그리신/NNP  
허치슨  
허치슨/NNP  
날래  
나\_01/VV+e 래/EF  
아멘호텔  
아멘호텔/NNP  
호코피자  
호코피자/NNG  
바로셀로나  
바로셀로나/NNP  
한울림  
한울림/NNP  
사과를 말했다.  
사과\_08/NNG+를/JKO 말하+VV+었/EP+다/EF+./SF



# UTagger option (분석 option)

MEMORY 2000	// 학습사전을 load할 memory 양(KByte), 많을수록 빠름
HLX_DIR ..\학습사전\	// 학습사전 폴더
cache 100000	// 캐쉬 사용량 조절. 윈도우는 100000 권장. 리눅스 유닉스는 0 권장
hanja_to_hangul 0	// 입력 한자 처리 한자 그대로(0), 한글로 변환(1)
depen	// 의존관계 분석시 용언의 의미제약(UWordMap) 사용
tag_poly_uwm 1	// 다의어 태깅
Noun_Attribute 0	// 복합명사 분해시 속성정보사용 여부 미사용(0), 사용(1)
analyzeMore 0	// 고유명사 분해할 최소 음절(3), 고유명사 추가분해 안함(0)
analyzeMoreNNG 1	// 고유명사 분해 시, 일반명사도 적용(분해할 최소 음절 적용)
analyzeMoreMiniFreq 5	// 복명명사 분해할 경우, 분해된 구성명사의 최소출현빈도
separate_compound 0	// 사전등재 복합명사(A^B, A-B 유형) 추가 분해 여부
Light 2	// CKMA 분석방법(정확률-속도): 가장느림(0), 약간느림(1), 빠름(2)
tagging 2	// 형태소분석(CKMA)만(0), HMM태깅(1), SCP태깅(2)
useAD 0	// HMM태깅: 앞 어절 전체형태소, 뒤 어절의 첫 형태소
useLD 0	// HMM태깅: 앞 어절 마지막형태소, 뒤 어절의 첫 형태소
useAF 1	// SCP태깅: 앞 어절전체, 뒤 어절의 첫 2개 음절
useEF 1	// SCP태깅: 앞 어절의 마지막 2개음절, 뒤 어절의 첫 2개 음절
recursive 1	// 띄어쓰기 오류 재귀분석
probability_equation 0	// 전이확률 계산시 확률식으로 계산 (정확률 조금 낮아짐)



# UTagger option (출력 option)



TAG_STYLE 0	// 울산대/세종 태그(0), ETRI 태그(1)
hadaVerb 0	// 서술형 명사 용언 어간형(0), 어근형(1)
print_sense_num 1	// 태깅 결과 출력 시 어깨번호 출력 안함(0), 동형이의어(1)
hangul_to_hanja 0	// 한자변환 안함(0), 변환(1), 병기(2)
hanjaLevel 0 1 2 3 4 5 6 7 8	// 한자 변환/병기 시 출력할 한자능력검정 급수
one_length_hanja_word_no_trans 1	// 1음절 한자어 변환여부 변환(0), 미변환(1)
ucs2le 1	// 한자 출력시 유니코드 출력 안함(0), 출력(1)
hanja_UCS2 1	// ucs2le=1 일 때 나라별 한자 한국ANSI(0), 한국(1), 대만(2), 중국(3), 일본(4)
CATE 1	// 의미매핑정보출력(대화식) 없음(0), 한자-뜻풀이(1), 의미범주/상위어(2)
print_end_empty_line 1	// 줄 단위로 모든 출력이 끝나면 마지막에 빈 줄을 출력한다.
print_original_sentence 0	// 입력 문장을 출력 여부. 안함(0), 출력(1)
print_one_line_sentence 1	// 태깅 결과를 한 줄로 출력. 안함(0), 출력(1)
print_ex 3	// 태깅결과 출력 안함(0), 한줄에 한어절(3)
print_depen 1	// 의존관계 출력 여부 안함(0), 출력(1), 규칙포함(2)
print_guess_line 0	// 미학습어절 별도 출력여부 안함(0), 출력(1)
preserve_splitter 1	// 입력 문장에서 어절 사이에 띄어쓰기 모양 유지 여부 안함(0), 유지(1)
preserve_newLine 1	// 입력 빈줄 출력 여부 안함(0), 유지(1)
debug_msg 1	// 콘솔로 실행시 각종 디버깅용 메시지를 출력. 0안함. 1사용



# UTagger 성능 평가 (1)

	2010년	2012년	2013년
형태소분석	규칙기반 (icma)	학습기반 (CKMA)	학습기반 (CKMA)
tagging	HMM	HMM (AD,LD)	SCP (AF,EF,FF)
정확률	88.93%	96.49%	96.37% (2013) 96.53% (2015)
속도(full option)	180sec	42.6sec	25.7sec
공유메모리	지원안함	지원	지원
띄어쓰기 오류 처리	처리 못함	기분석어절포함 된 경우 분석	기분석어절포함 된 경우 분석
복합명사 의미분석	못함	분석(접사 포함)	분석(접사 포함)
Code (한자)	KS5601	KS5601	Unicode

❖ 학습사전 : 세종형태의미말뭉치 1천만 어절 대상, 약 1G



# UTagger-HM 성능 평가 (2)

단어	동형이의어/POS	학습사전 출현빈도	10% 정답	UTagger 태깅	정답률
눈	눈__01/NNG	12,455	1,236	1,244	100.56%
	눈__04/NNG	1,379	141	133	94.33%
	소계	13,834	1,377	1,377	100.0%
손	손__01/NNG	9,209	925	928	100.32%
	손__05/NNB	3	1	0	0.00%
	손__08/NNP	188	18	18	100.0%
	손__09/NNG	27	1	2	20.0%
	소계	9,427	945	948	100.32%
말	말__01/NNG	34,136	3,337	3,337	100.0%
	말__03/NNB	118	7	6	85.71%
	말__05/NNG	584	57	58	101.75%
	말__07/NNG	2	1	1	100.0%
	말__11/NNB	2,723	219	262	119.63%
	소계	37,563	3,621	3664	101.19%
거리	거리__01/NNG	2,056	220	210	99.55%
	거리__02/NNB	339	31	31	100.0%
	거리__04/NNB	14	1	1	100.0%
	거리__08/NNG	1,491	152	153	100.66%
	소계	3,900	404	395	97.77%
바람	바람__01/NNB	1,222	128	125	97.66%
	바람__01/NNG	3,468	357	363	101.68%
	바람__02/NNG	120	10	7	70.0%
	소계	4,810	495	495	100.0%

자리	자리__01/NNG	7,589	717	723	100.84%
	자리__02/NNG	94	12	6	50.0%
	소계	7,683	729	729	100.0%
의사	의사__02/NNG	1,018	108	107	99.07%
	의사__03/NNG	66	6	6	100.0%
	의사__04/NNG	1	1	1	100.0%
	의사__09/NNG	3	1	1	100.0%
	의사__11/NNG	6	2	2	100.0%
	의사__12/NNG	1,754	183	184	100.55%
	의사__14/NNG	24	1	1	100.0%
	소계	2,872	302	302	100.0%
점	점__02/NNG	7	1	1	100.0%
	점__03/NNG	86	9	9	100.0%
	점__10/NNB	1,557	237	154	64.98%
	점__10/NNG	9,861	942	1,025	108.81%
	소계	11,491	1,189	1,189	100.0%
밤	밤__01/NNG	5,388	494	494	100.0%
	밤__02/NNG	124	17	17	100.0%
	소계	5,512	511	511	100.0%
목	목__01/NNG	1,968	188	189	100.53%
	목__09/NNG	3	1	1	100.0%
	목__10/NNG	65	7	7	100.0%
	목__12/NNG	3	1	1	100.0%
	목__13/NNG	9	1	1	100.0%
	목__14/NNG	13	2	0	0.0%
소계	2,061	200	199	99.5%	

# UTagger-DP (의존관계 분석) (1)

- 문장 구성성분의 의존 관계 분석
- UTagger-HM의 분석결과 이용 : 동형이의어 분별

- 동형이의어 용언의 문형 정보 활용

- 차\_01/VV : <...에> <...으로> “독에 물이 차다”
- 차\_02/VV : <...을> “공을 차다/혀를 차다”
- 차\_03/VV : <...에 ...을> “허리에 칼을 차다”
- 차\_04/VA : “성격이 차고 매섭다 / 바람이 차다”

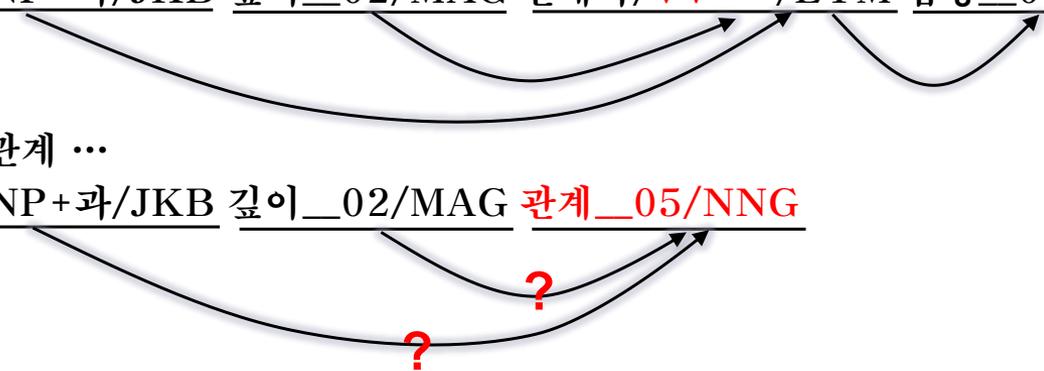
- 서술성 명사

- 중국과 깊이 관계한 협상은 ...

중국\_01/NNP+과/JKB 깊이\_02/MAG 관계하/VV+ㄴ/ETM 협상\_01/NNG+은/JX

- 중국과 깊이 관계 ...

중국\_01/NNP+과/JKB 깊이\_02/MAG 관계\_05/NNG



- 규칙기반 정확률 : 85.53% (의존관계로 변환된 세종구문분석말뭉치 대상)
- 기계학습기반 의존관계분석 (2015) : 정확률 ??.% > 87%



# UTagger-DP (의존관계 분석) (2)

## ❖ UTagger-DP (ver. 0.9)

UTagger (품사 & 동형의외어 태깅 시스템)

UI 보이기 | 파일 열기 | 문장 분석 | 연속 분석 | 저장 | 파일 분석 | 폴더 분석 | 기타 기능 | 옵션 설정 | 종료

직접 입력 분석 | 어절 단위 보기 | 문장 단위 보기

로 s건설의 중구 봉래동 현장에서는 본사 부장급인 현장 소장보다 월급을 더 많이 받는 기능공도 적지 않게 있다.

입력분석

//////////태깅 결과//////////

실제로/MAG S/SL+건설/NNG+의/JKG 중\_\_04/NNG+구\_\_15/NNG 봉래동/NNP 현장\_\_03/NNG+에서/JKB+는/JX 본사\_\_03/NNG 부장\_\_07/NNG+급\_\_04/NNG+이/VCP+ㄴ/ETM 현장\_\_03/NNG 소장\_\_08/NNG+보다/JKB 월급/NNG+을/JKO 더\_\_01/MAG 많이/MAG 받\_\_01/VV+는/ETM 기능공/NNG+도/JX 적\_\_02/VA+지/EC 많/VX+게/EC 있\_\_01/VX+다/EF+./SF

1	13	24	실제로	실제로/MAG
2	3	4	s건설의	S/SL + 건설/NNG + 의/JKG
3	4	9	중구	중__04/NNG + 구__15/NNG
4	5	9	봉래동	봉래동/NNP
5	13	26	현장에서는	현장__03/NNG + 에서/JKB + 는/JX
6	7	9	본사	본사__03/NNG
7	8	4	부장급인	부장__07/NNG + 급__04/NNG + 이/VCP + ㄴ/ETM
8	9	9	현장	현장__03/NNG
9	12	20	소장보다	소장__08/NNG + 보다/JKB
10	13	36	월급을	월급/NNG + 을/JKO
11	12	7	더	더__01/MAG
12	13	6	많이	많이/MAG
13	14	4	받는	받__01/VV + 는/ETM
14	15	14	기능공도	기능공/NNG + 도/JX
15	16	2	적지	적__02/VA + 지/EC
16	17	2	않게	않/VX + 게/EC
17			있다.	있__01/VX + 다/EF + ./SF

형태소 수정 반영



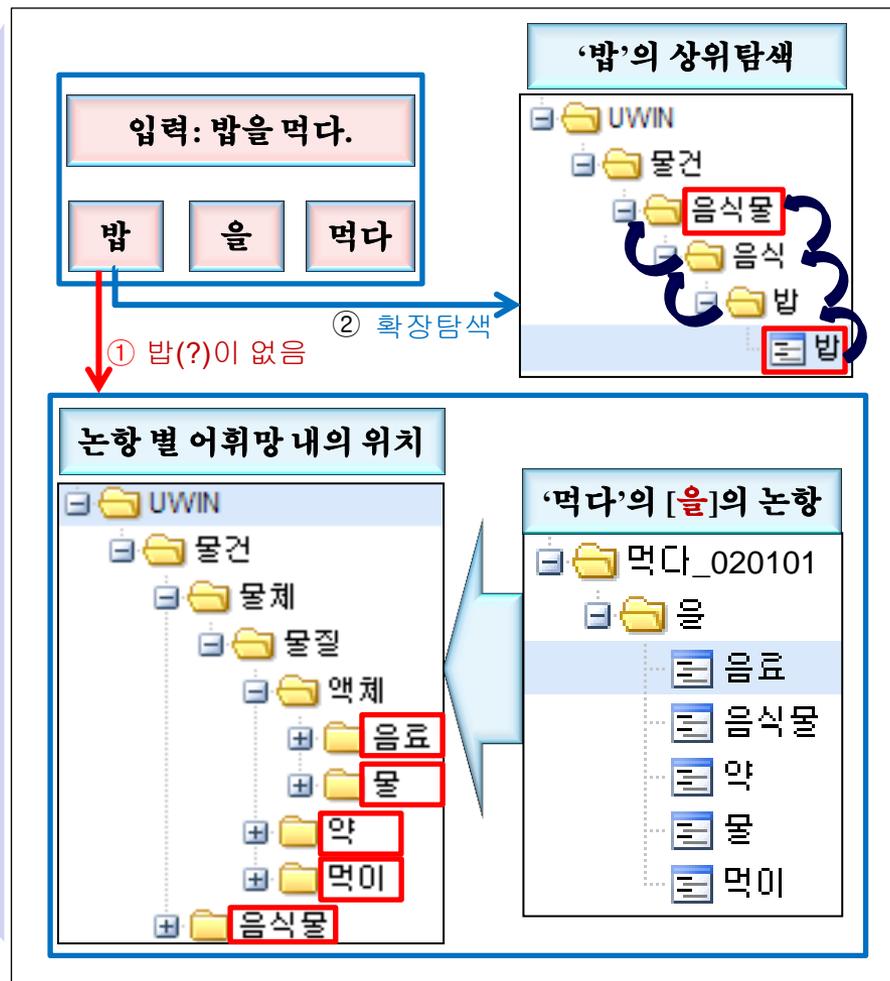
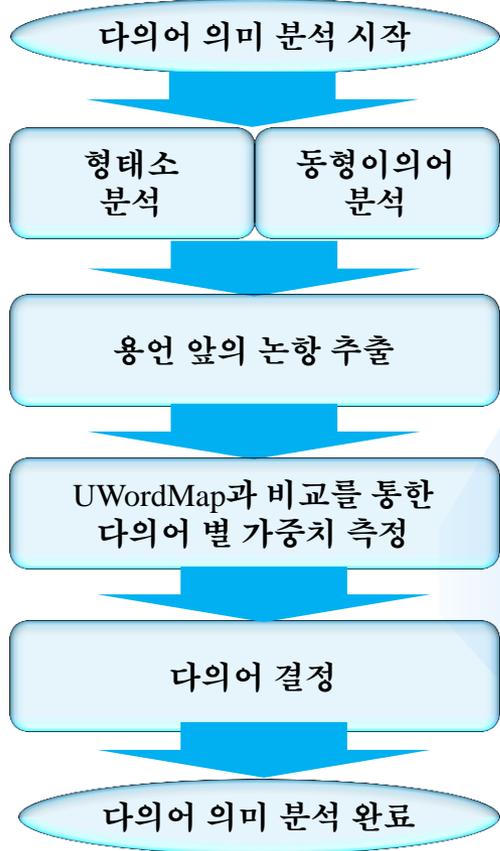
# UTagger-DP (의존관계 분석) (3)

- 규칙기반 정확률 : 85.53% (의존관계로 변환된 세종구문분석말뭉치)

규칙번호	정답	오답	규칙적용	전체비율	정확률
rule_1	3,801	1	3,802	1.02	99.97
rule_2	46,134	380	46,514	12.49	99.18
rule_3	2,816	516	3,332	0.89	84.51
rule_4	55,565	10,330	65,895	17.70	84.32
rule_5	7,685	2,360	10,045	2.70	76.51
rule_6	8,832	192	9,024	2.42	97.87
rule_7	392	4	396	0.11	98.99
rule_8	454	21	475	0.13	95.58
rule_9	34,008	5,106	39,114	10.51	86.95
rule_10	518	0	518	0.14	100.00
rule_11	2,124	131	2,255	0.61	94.19
rule_12	494	497	991	0.27	49.85
rule_13	47,941	579	48,520	13.03	98.81
rule_14	16,089	2,876	18,965	5.09	84.84
rule_15	29,330	2,392	31,722	8.52	92.46
rule_16	6,435	892	7,327	1.97	87.83
rule_17	211	2	213	0.06	99.06
rule_18	2,597	472	3,069	0.82	84.62
rule_19	0	0			
rule_20	402	200	602	0.16	66.78
rule_21	2,455	691	3,146	0.84	78.04

규칙번호	정답	오답	규칙적용	전체비율	정확률
rule_22	3,203	1,575	4,778	1.28	67.04
rule_23	695	107	802	0.22	86.66
rule_24	783	742	1,525	0.41	51.34
rule_25	468	152	620	0.17	75.48
rule_26	1,236	496	1,732	0.47	71.36
rule_27	10,145	3,919	14,064	3.78	72.13
rule_28	11	2	13	0.00	84.62
rule_29	8,334	7,380	15,714	4.22	53.04
rule_30	2,470	577	3,047	0.82	81.06
rule_31	510	108	618	0.17	82.52
rule_32	1,599	1,021	2,620	0.70	61.03
rule_33	2,048	569	2,617	0.70	78.26
rule_34	508	348	856	0.23	59.35
rule_35	691	167	858	0.23	80.54
rule_36	3,001	1,301	4,302	1.16	69.76
rule_37	0	0			
rule_38	82	47	129	0.03	63.57
rule_39	491	101	592	0.16	82.94
rule_40	13,904	7,602	21,506	5.78	64.65
<b>total</b>	<b>318,462</b>	<b>53,856</b>	<b>372,318</b>	<b>100.00</b>	<b>85.53</b>
UTagger 오류	3,605				

## ❖ 의미 분석 과정





# UTagger-PS (다의어 WSD) (2)

## ❖ UTagger-PS (ver. 0.5)

- 동형이의어분별 후 어휘지도 상에서 “용언-명사” 발견
- 의존관계 사용하지 않음
- 용언의 왼쪽, 오른쪽 어절에 대해 다른 가중치 적용

UTagger (품사 & 동형이의어 태깅 시스템)

UI 보이기    기타 기능    옵션 설정    종료

직접 입력 분석 | 어절 단위 보기 | 문장 단위 보기

너 기저귀 차고 놀 때 나는 공 차고 놀았다.

입력분석

//////////태깅 결과//////////

너\_010000/NP 기저귀\_000000/NNG 차\_030101/VV+고/EC 놀\_010103/VV+ㄹ/ETM 때\_010001/NNG 나\_030100/NP+는/JX  
 공\_010001/NNG 자\_020001/VV+고/EC 놀\_010101/VV+았/EP+다/EF+./SF

1 2 9 너    너\_010000/NP  
 2 3 36 기저귀    기저귀\_000000/NNG  
 3 4 40 차고    차\_030101/VV + 고/EC  
 4 5 4 놀    놀\_010103/VV + ㄹ/ETM  
 5 6 9 때    때\_010001/NNG  
 6 8 36 나는    나\_030100/NP + 는/JX  
 7 8 36 공    공\_010001/NNG  
 8 9 40 차고    차\_020001/VV + 고/EC  
 9    놀았다.    놀\_010101/VV + 았/EP + 다/EF + ./SF

//////////형태소 분석 결과//////////

너 FWD no  
 [ 1 ] 너\_01[이인칭\_대명사]/NP --- 19284546.000000000  
 [ 2 ] 너\_02[첫임]/MM --- 100002.000000000  
 [ 3 ] 너\_88/NNG --- 50000.000000000  
 [ 4 ] 너\_01[수지에서\_벗어나\_지나다]/VV --- 5.000000000  
 [ 5 ] 너\_01[위하여\_풀쳐\_놀다]/VV --- 5.000000000

기저귀 FWD AD  
 [ 1 ] 기저귀/NNG --- 590147.000000000

차고 FWD AD  
 [ 1 ] 차\_03[달아매거나\_끼워서\_지니다]/VV+고/EC --- 2090006.000000000  
 [ 2 ] 차\_04[대기의\_온도가\_낮다]/VA+고/EC --- 1380000.000000000  
 [ 3 ] 차\_02[내지르거나\_받아\_올리다]/VV+고/EC --- 1160007.000000000  
 [ 4 ] 차\_01[가득하게\_되다]/VV+고/EC --- 1040010.000000000

px

수정창    작업량 확인    로그아웃

▶ 정보란

차다    검색    동형이의어만 검색     저장시 자동 페이지 넘기기

차다\_010203[동사] [] : 머연 풀이다 한도네 미르는 상태가 되다.  
 차다\_010301[동사] [] : 정한 수량, 나이, 기간 따위가 다 되다.  
 차다\_010302[동사] [] : 이지러진 데가 없이 아주 온전하다.  
 차다\_020001[동사] [] : 발로 내지르거나 받아 올린다.  
 차다\_020002[동사] [] : 발을 힘껏 뺨어 사람을 치다.  
 차다\_020003[동사] [] : 허골을 입천장 앞쪽에 붙였다가 떼어 소리를 내다.  
 차다\_020004[동사] [] : 발로 힘있게 밀어젖히다.  
 차다\_020005[동사] [] : (속되게) 주로 남녀 관계에서 일반적으로 관계를 끊다.  
 차다\_020006[동사] [] : 날채게 빼앗거나 움켜 가지다.  
 차다\_020007[동사] [] : 자기에게 베풀어지거나 차례가 오는 것을 받아들이지 않다  
 차다\_030101[동사] [] : 물건을 몸의 한 부분에 달아매거나 끼워서 지니다.  
 차다\_030102[동사] [] : 수갑이나 차고 따위를 팔목이나 발목에 끼우다.  
 차다\_030200[동사] [] : (속되게) 애인으로 삼아 데리고 다니다.  
 차다\_040101[형용사] [] : 몸에 달은 물체나 대기의 온도가 낮다.  
 차다\_040102[형용사] [] : 인정이 없고 쌀쌀하다.  
 차다\_040103[형용사] [] {한의학} : 약재(藥材)나 약제(藥劑)의 성질이 차서 몸의

※ 뜻풀이가 없는 경우 고유명사인 경우와 마닌 경우를 나누어서 태깅하여 주십시오  
 ※ 형태소 분석이 잘못된 경우 '형태소 분석 오류'를 선택해 주십시오

적용    저장

## ❖ 전체 다의어 대상 실험 (2015.03.05)

- 표준국어대사전에서 용언을 포함하는 용례 대상 : 210,426문장
- 현재의 어휘지도 구축 현황 (18쪽 참고)

	명사	부사	동사	형용사	합계
정답	572,182	74,292	337,613	88,567	1,072,654
Base line (첫번째다의어)	430,660 (75.27%)	48,170 (64.84%)	175,440 (51.96%)	40,307 (45.51%)	694,577 (64.75%)
어휘지도를 이용한 다의어 WSD	418,873 (73.21%)	47,586 (64.05%)	191,799 (56.81%)	43,640 (49.27%)	710,898 (65.44%)



# U-Tagger-SR (UPropBank 기반 의미역 부착 도구) (1)

## ❖ 의미역 부착 말뭉치 구축(31개 파일, 2014)

- 세종구구조부착말뭉치를 의존구조로 변경
- 수작업으로 의미역 부착
- 개별 의미역 통계
- 용언별로 격조사-의미역 통계

## ❖ UPropBank

- 다의어 수준, 90,090 용언
- 동형이의어 수준으로 의미역 통합

전체 서술어 개수			240,747
의미역 부착 서술어 개수			181,404
행동주 (AGT)	48,150	경험주 (EXP)	3,085
피동주 (PAT)	5,095	동반자 (COM)	3,073
대상 (THM)	129,919	기점 (SRC)	8,200
착점 (GOL)	20,718	처소 (LOC)	12,197
자극 (STM)	412	원인 (CAU)	3,519
비교기준 (CRT)	6,404	시간 (TMP)	6,946
정도 (DGR)	3,583	방법 (MNR)	7,652
자격 (ROL)	1,741	재료 (MAT)	341
도구 (INS)	2,975	경로 (ROU)	250
방향 (DIR)	2,291	수혜자 (BEN)	1,289
내용 (CNT)	7,910	목적 (PUR)	548



# UTagger-SR (UPropBank 기반 의미역 부착 도구) (2)

UTaggerSR 1.0

파일 이름: C:\Users\Wsk\Desktop\W김윤정\작업\일지\부어개번호미부착\_ETRI\_인코딩수정\WBGAD 열기(O)    파일 형식: 일반 텍스트    작업 파일 열기

문장: 대통령 취임 후에도 그는 `연설`, `한 여름의 회상`을 출판하는 등 격무 속에서도 저술 활동을 계속하고 있다.

의미역부착 형태소:의존관계

순서	의존	어절	9 출판하_02	15 계속하_03
1	2	대통령/NNG		
2	3	취임/NNG		
3	15	후_08/NNG+에//JKB+도//JX	TMP	TMP
4	15	그_01/NP+는//JX	AGT	AGT
5	9	`/SW+연설_02/NNG+`/SS+`//SP		
6	7	`/SW+한_01/MM		
7	8	여름_01/NNG+의//JKG		
8	9	회상_02/NNG+`/SS+을//JKO	THM	
9	10	출판하_02/VV+는//ETM		
10	11	등_05/NNB		
11	12	격무/NNG		
12	15	속_01/NNG+에서//JKB+도//JX		
13	14	저술/NNG		
14	15	활동_02/NNG+을//JKO		THM
15	16	계속하_03/VV+고//EC		
16	16	있_01/VX+다//EF+`//SF		

문장 목록: 1 / 416    이전(P)    다음(N)    ▶

대통령 취임 후에도 그는 `연설`, `한 여름의 회상`을 출판하는 등 격무 속에서도 저술 활동을 계속하고 있다.

이 같은 문장 활동에 대한 평가와 인권운동가로서의 활약으로 하벨 대통령은 에라스무스상, 자유의 상, 유네스코 인권상 등 국제적 ...

대통령 재임 2년 동안 그에 대한 평가가 슬로바키아쪽에선 일부 부정적으로 나오고 있긴 하다.

사전 검색

문형

계속하다\_030001[동]  
계속하다\_030002[동]

\* ...을

용법

날말 뜻

계속04[I](1). 끊이지 않고 이어 나감.

예문

- \* 두 달 동안 아침 운동을 계속하다.
- \* 공부를 하루 종일 계속하다.
- \* 민중은 끈질긴 항쟁을 계속했다.
- \* 사람들의 만세 소리가 여기저기에서 계속하여 울려 퍼졌다.

서술어 선택: 15 계속하\_03

의미역 선택 기준: 문형, 용언 참고

행동주 AGT, 대상 THM, 자극 STM, 시간 TMP, 재료 MAT, 수혜자 BEN, 삭제(D)

경험주 EXP, 기점 SRC, 원인 CAU, 정도 DGR, 도구 INS, 내용 CNT, 되돌리기

피동주 PAT, 착점 GOL, 비교기준 CRT, 방법 MNR, 경로 ROU, 목적 PUR, 저장(S)

동반자 COM, 처소 LOC, 자격 ROL, 방향 DIR

## ❖ 실험 결과

- 65,529개 문장
- 59,257개 문장 격조사-용언별 의미역 통계
- 6,272개 문장 실험

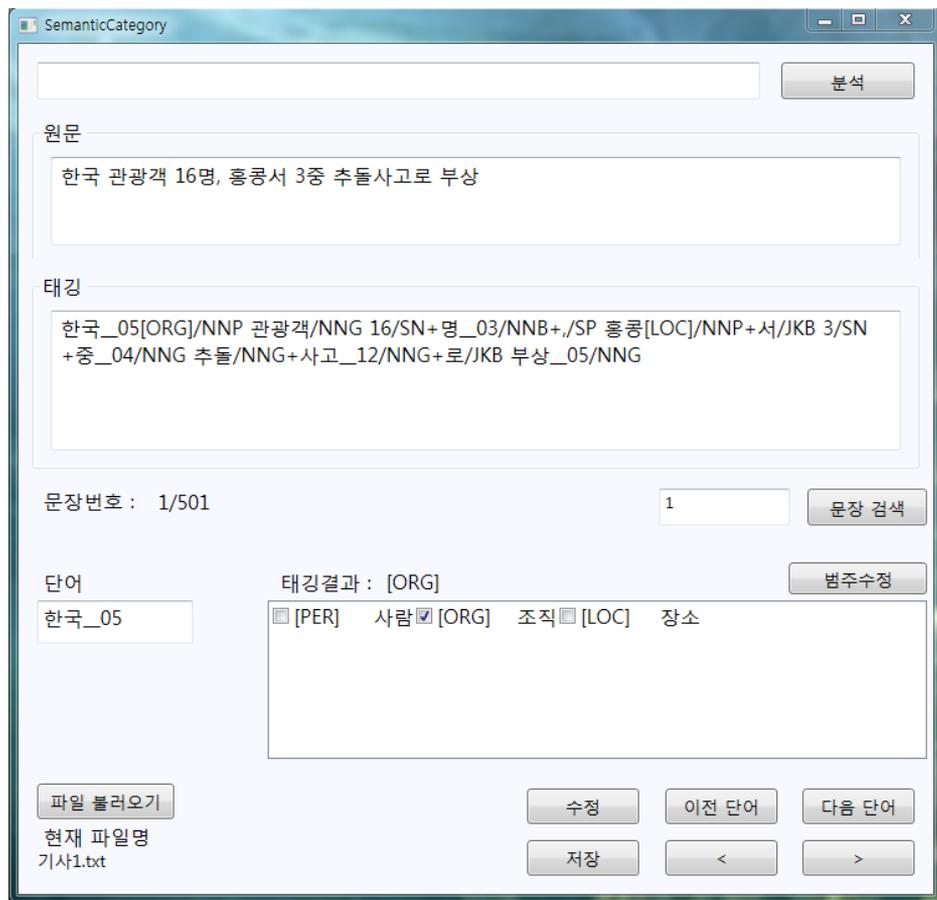
방법	태깅 정답수	의미역 수	정확률(%)
UPropBank & 후보 1개	5,548	7,106	78.08
후보 2개 이상 이거나 없을 때, 격조사별 의미역 빈도	7,390	10,668	69.27
후보 2개 이상 이거나 없을 때, 격조사-용언별 의미역 빈도	7,891	11,191	70.51
후보 2개 이상 이거나 없을 때, 뒤 서술어에 대해서 격조사-용언별 의미역 빈도	7,882	10,901	72.31

❖ 의미범주 : ETRI 대분류 + UWordMap의 상위 노드 참조 (13개 범주)

범주명	의미	범주명	의미
[PER]	사람	[ACC]	정도
[KNW]	지식	[ACT]	활동
[MAT]	물질	[ANM]	동물
[CHR]	성질	[PNT]	식물
[ORG]	조직	[REL]	관계
[SPC]	공간	[ETC]	기타
[TME]	시간		

# UTagger-NE (개체명 인식기) (2)

- ❖ UTagger-NE (ver. 1.0)
  - 사용자 의미범주 정의 기능
  - 신경회로망 기반 학습
  - 자동 태깅 및 수정



SemanticCategory

분석

원문

한국 관광객 16명, 홍콩서 3중 추돌사고로 부상

태깅

한국\_05[ORG]/NNP 관광객/NNG 16/SN+명\_03/NNB+./SP 홍콩[LOC]/NNP+서/JKB 3/SN +중\_04/NNG 추돌/NNG+사고\_12/NNG+로/JKB 부상\_05/NNG

문장번호 : 1/501      1      문장 검색

단어      태깅결과 : [ORG]      범주수정

한국\_05       [PER]     사람 [ORG]     조직 [LOC]    장소

파일 불러오기      수정    이전 단어    다음 단어

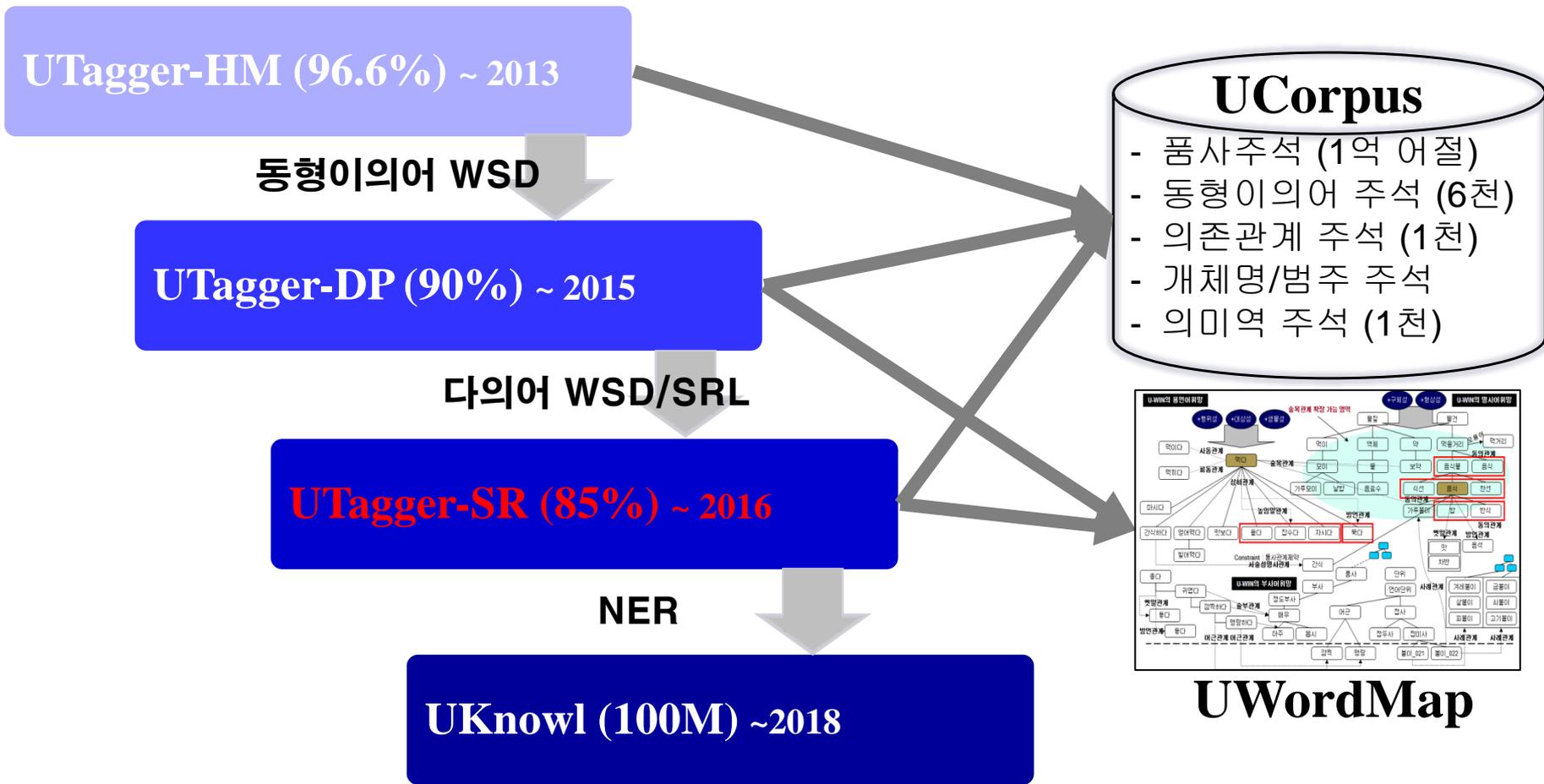
현재 파일명      저장    <    >

기사1.txt

**❖ 실험(PLO) 결과**

문서	PER	ORG	LOC	소계	정확률(%)
신문1	172/195	135/194	125/150	432/539	80.15
신문2	42/52	134/142	349/372	525/566	92.76
신문3	109/133	146/165	522/591	777/889	87.40
백과사전1	111/122	206/236	221/261	538/619	86.91
백과사전2	134/138	412/417	126/133	672/688	97.67
소설1	225/237	4/5	59/81	288/323	89.16
소설2	33/34	2/2	81/89	116/125	92.80
소설3	65/67	7/8	21/35	93/110	84.55
합계	891/978	1,046/1,169	1,504/1,712	3,441/3,859	
정확률(%)	91.10	89.48	87.85	89.17	

- ❖ 향상된 의존관계파서를 적용
  - 용언과 논항의 관계를 보다 정확히 파악
  
- ❖ 어휘지도 활용
  - 미학습 동형이의어 태깅
  - 의존관계 분석 시 의미제약
  - 의미역 태깅
  - 의미범주
  
- ❖ 다의어 분석 대상 확대
  - (동사, 형용사) + 명사
  - 수의 논항에 의한 WSD 방법





## Lexical Semantics

### WSD, NER

- 하나의 형태소가 문맥에 따라 여러개의 의미로 해석될 때, 하나의 의미를 결정
- Homograph WSD
- Polysemy WSD
- Named Entity Recognition
- Lexical Semantic Network

## Sentence Semantics

### Parser, SRL

- 문장 내의 술어-논항(predicate-argument) 관계에 적합한 의미 관계(Semantic Role)를 결정
- Subcategorization
- Semantic Restriction
- Semantic Role Labeling
- Ontology & Inference

❖ 연구실 홈페이지 : <http://nlplab.ulsan.ac.kr> DEMO

- WordsMap Browser
  - [http://klplab.ulsan.ac.kr:8080/uwin\\_nrel.jnlp](http://klplab.ulsan.ac.kr:8080/uwin_nrel.jnlp) : UWIN전체 대상
  - <http://youtu.be/bE3GtuG6gN8>
- 형태소/동형이의어 태깅시스템 (UTagger-HM)
  - <http://nlplab.ulsan.ac.kr:8080/KMAClient/KMAClient.jsp> (WEB용)
  - 한자자동변환시스템 <http://hanjaro.juntong.or.kr/>
  - 다의어 WSD [http://youtu.be/g1V2DNq\\_xt8](http://youtu.be/g1V2DNq_xt8)
  - 아래한글에 UTagger-HM addon <http://youtu.be/Ibhy7okBO98>
- UTagger-SR (격틀사전 기반 의미역 부착도구)
  - <https://youtu.be/8AgQnrE71n4>
- UTagger-NE (개체명인식기)
  - [https://www.youtube.com/watch?v=BgxoxN\\_a4n8&feature=youtu.be](https://www.youtube.com/watch?v=BgxoxN_a4n8&feature=youtu.be)

# Q & A

[okcy@ulsan.ac.kr](mailto:okcy@ulsan.ac.kr)

052-259-2222 (010-2561-5830)

울산대학교 한국어처리연구실/지능형컴퓨터연구실

<http://nlplab.ulsan.ac.kr>