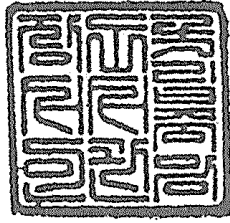


21세기 세종계획
구문분석말뭉치
종합관리도구
사용자 설명서

문 화 관 광 부
국립국어연구원

21세기 세종계획
구문분석말뭉치
종합관리도구
사용자 설명서

문 화 관 광 부
국립국어연구원



2988114

구문분석 말뭉치 종합관리도구 사용자 설명서

목차

1. 설치방법

2. 시스템 구성

- 가. 개요
- 나. 말뭉치' 폴더
- 다. '구문분석말뭉치종합관리도구' 폴더
- 라. '사용자 설명서' 폴더
- 마. '프로그램 소스' 폴더

3. 사용 방법

- 가. 구문분석 말뭉치 종합관리도구를 처음 실행했을 때의 화면
- 나. 구문분석 말뭉치 변환도구
- 다. 구문분석 말뭉치 검색도구
- 라. 구문분석 말뭉치 수정도구

4. 부록

- 가. 구문분석 말뭉치 검색도구에 사용되는 다단계 검색식 사용법

1. 설치 방법

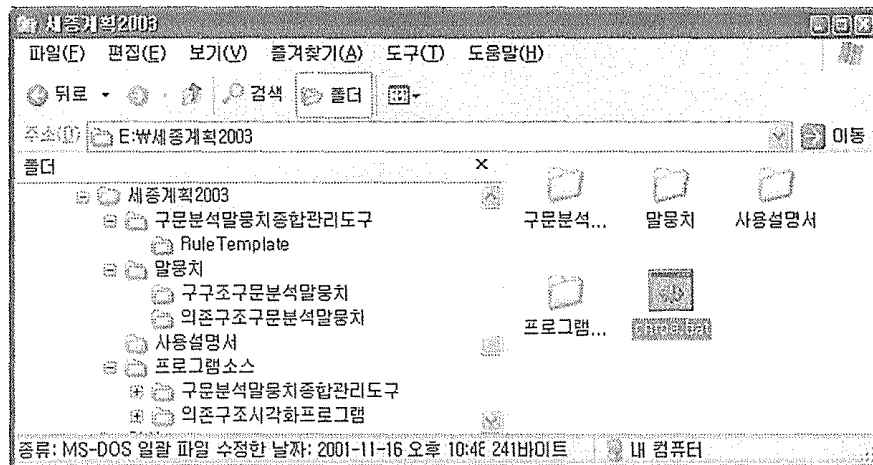
가. 설치방법

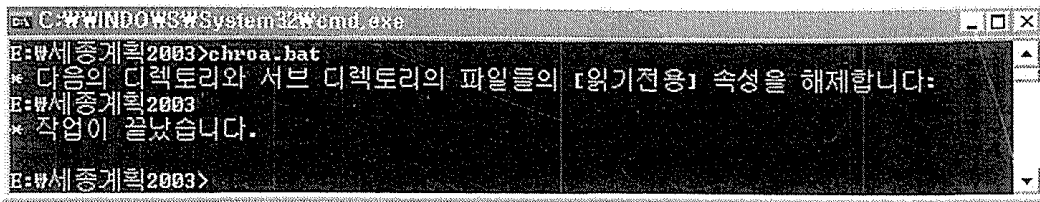
본 시스템은 21세기 세종계획에서 구축한 이진 구구조 구문분석 말뭉치를 관리하는 작업을 지원하는 도구이다. 본 시스템은 별도의 설치/제거 프로그램이 없다. 따라서 하드디스크의 원하는 폴더에 CD의 내용을 그대로 복사하여 사용하면 된다. 제거하려면, 해당 폴더를 삭제하면 된다.

나. 설치시 주의사항

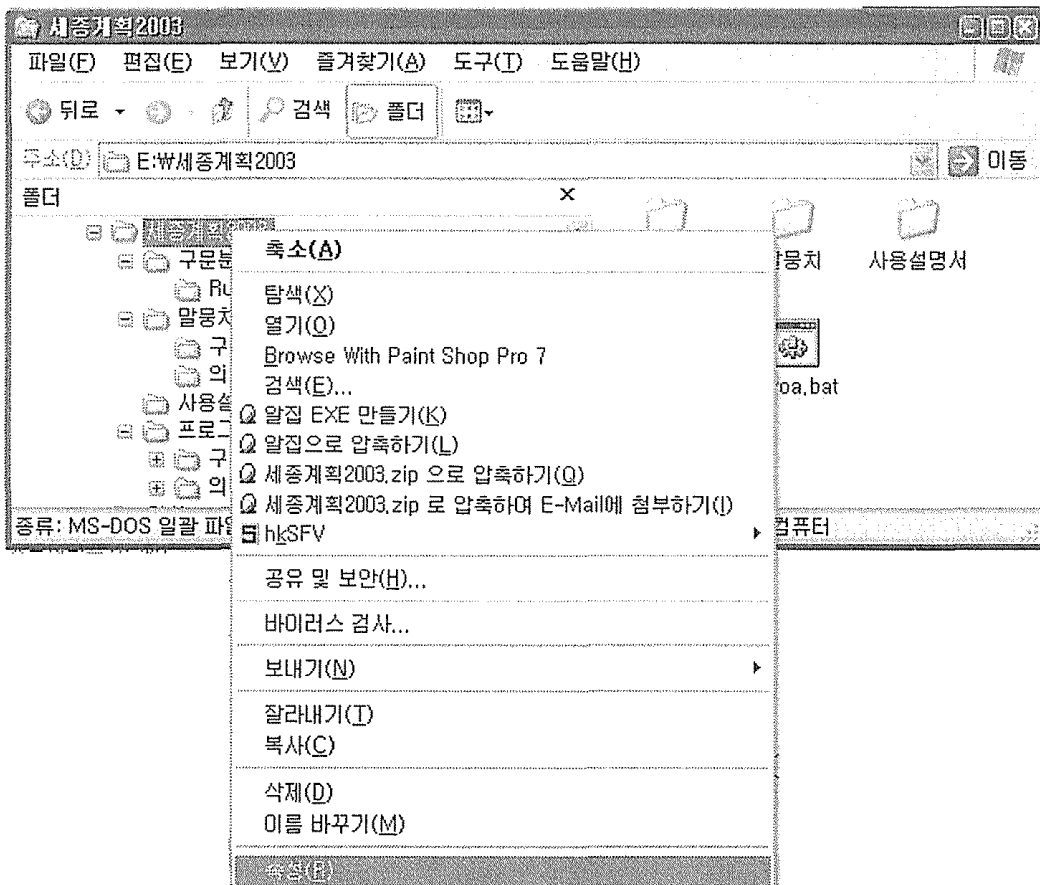
- ① 본 시스템은 Windows 98이상의 운영체제가 설치된 펜티엄급 PC에서 잘 동작한다.
- ② 본 시스템은 반드시 구문분석 말뭉치 종합관리도구의 최상위 폴더 전체를 하드디스크에 복사한 후에 사용해야 한다.
- ③ CD-ROM에 있는 파일은 [읽기전용] 특성이 있는데, 이러한 [읽기전용] 특성은 하드디스크에 복사하여도 그대로 유지된다. 그런데, 반자동 구문분석 말뭉치 구축도구가 제대로 작동하기 위해서는 [읽기전용] 특성을 해제하여야 한다. 이를 위해 [방법1]이나 [방법2]에서 하나를 선택하여 [읽기전용] 특성을 해제할 수 있다.

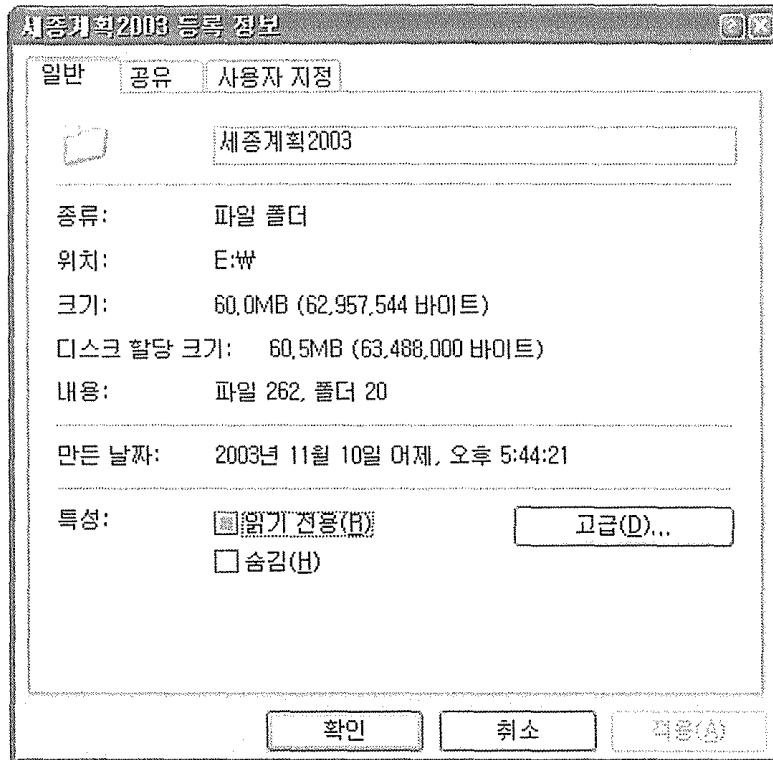
[방법1] '세종계획2003' 폴더에 있는 chroa.bat를 실행



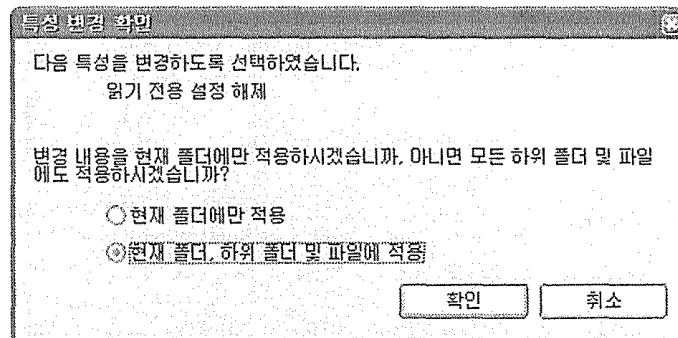


[방법2] '세종계획2003' 폴더를 선택하고 마우스 오른쪽버튼을 눌러 '등록정보' 혹은 '속성'을 선택한다. 등록정보의 [일반] 탭 하단에 있는 [읽기전용] 특성이 (V)(윈도우98/2000) 혹은 (O)(윈도우XP)로 설정되어 있으면, 클릭하여 □로 [읽기전용] 특성을 해제한다.





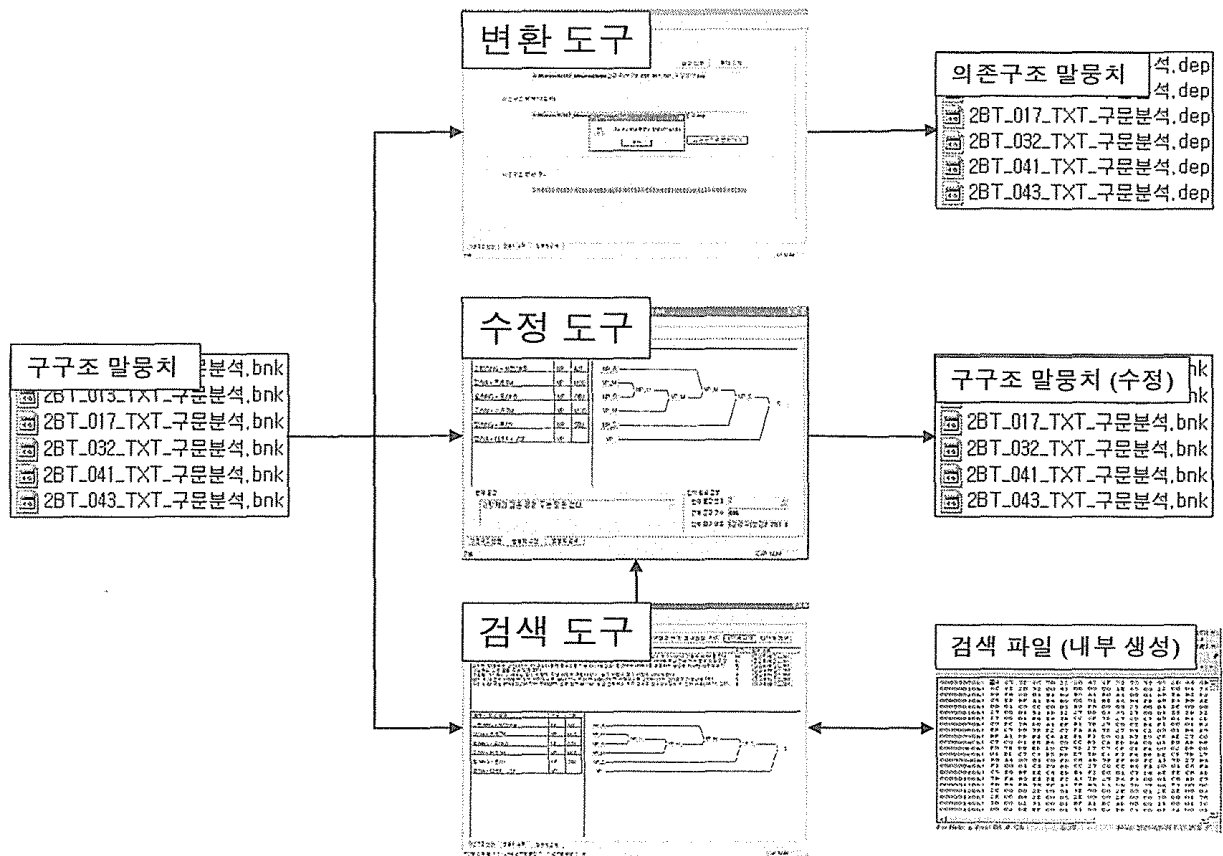
※ 윈도우 98에서는 모든 하위 폴더와 파일에 대해서도 위의 방법으로 [읽기 전용] 특성을 해제한다. 반면에, 윈도우 2000에서 [읽기전용] 특성을 설정하거나 해제하면, 다음과 같은 대화상자가 나오는데, [현재 폴더, 하위 폴더 및 파일에 적용]을 선택하여 [읽기전용] 특성을 해제한다.



2. 시스템 구성

가. 개요

구문분석 말뭉치 종합관리도구는 다음과 같이 말뭉치 변환도구, 말뭉치 검색도구, 말뭉치 수정도구가 하나의 프로그램 안에 합쳐져 있는 통합환경으로 구성되어 있다. 첫째, 말뭉치 변환도구는 이진 구구조 말뭉치(.bnk)를 입력하면, 의존구조가 부착된 별도의 말뭉치(.txt)를 출력한다. 둘째, 말뭉치 검색도구는 이진구구조 말뭉치를 대상으로 원하는 조건을 가진 문장과 구문 구조를 찾아준다. 셋째, 말뭉치 수정도구는 구문분석 말뭉치에 오류가 있을 경우, 이를 수정할 수 있는 기능을 제공하는 도구이다.



CD의 포함된 자료는 다음과 같은 폴더 구조에 저장되어 있다. 최상위 폴더는 '세종계획2003'이다.

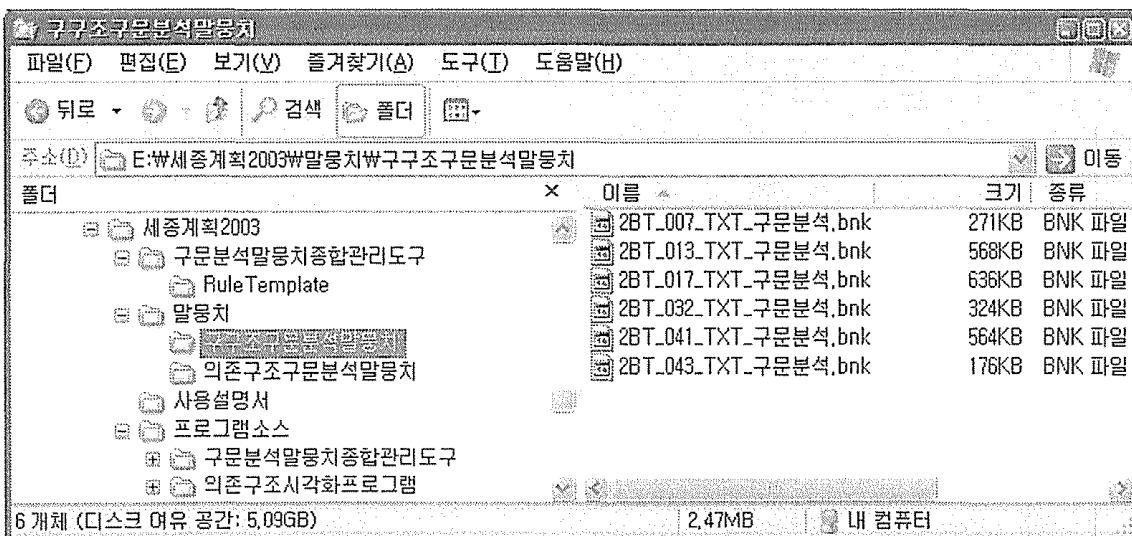
폴더 구조	내용물
<ul style="list-style-type: none"> ☐ 세종계획2003 <ul style="list-style-type: none"> ☐ 구문분석말뭉치종합관리도구 <ul style="list-style-type: none"> ☐ RuleTemplate ☐ 말뭉치 <ul style="list-style-type: none"> ☐ 구구조구문분석말뭉치 ☐ 의존구조구문분석말뭉치 ☐ 사용설명서 ☐ 프로그램소스 <ul style="list-style-type: none"> ☐ 구문분석말뭉치종합관리도구 프로그램 소스 ☐ 의존구조시각화프로그램 프로그램 소스 	<p>최상위 폴더</p> <p>프로그램 실행파일이 포함된 폴더</p> <p>구구조 수정에 필요한 규칙 템플릿이 들어 있는 폴더</p> <p>말뭉치가 포함된 폴더</p> <p>구구조구문분석말뭉치(세종계획2002년도 결과물)</p> <p>의존구조구문분석말뭉치(의존구조로 변환한 결과)</p> <p>사용자 설명서가 포함된 폴더</p> <p>프로그램 소스가 들어있는 폴더</p> <p>구문분석말뭉치종합관리도구 프로그램 소스</p> <p>의존구조시각화도구 프로그램 소스</p>

구문분석 말뭉치 종합관리도구의 전체 폴더 구조

나. ‘말뭉치’ 폴더

말뭉치 폴더는 다음과 같이 구구조 구문분석 말뭉치(.bnk), 의존구조 구문분석 말뭉치(.dep)로 구성되어 있다. 각 폴더는 해당 말뭉치 파일을 포함한다.

(1) ‘구구조구문분석말뭉치’ 폴더



구문분석 말뭉치는 이진 분할로 상위 노드와 하위 노드가 대치되어 있는 구조로 되어 있다. 본 시스템에서는 기본적으로 세종계획 형태소 분석 말뭉치의 형식을 바탕으로 자동으로 지정되도록 만들어 졌다. 그러나, 구문 표지는 사용자가 지정하여 수정이 가능하도록 구성되어 있다. 구문태그 및 분석방법은 고려대학교 민족문화연구원 전자텍스트연구소에서 지정한 태그 및 분석방침을 따른다. 구문분석 말뭉치의 예는 다음과 같다.

```

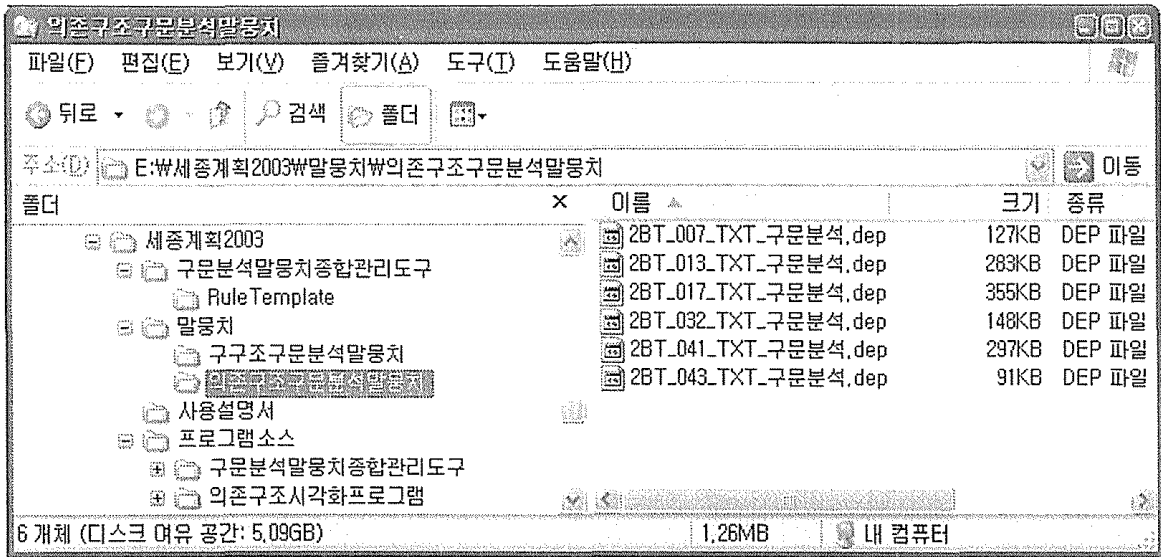
; 고향의 품
(NP      (NP_MOD 고향/NNG + 의/JKG)
         (NP 품/NNG))

; 고향처럼 깊은 꿈을 주는 말은 없다.
(S (NP_SBJ (VP_MOD (NP_AJT 고향/NNG + 처럼/JKB)
                  (VP_MOD? (NP_OBJ (VP_MOD 깊/VA + 은/ETM)
                                   (NP_OBJ 꿈/NNG + 을/JKO))
                               (VP_MOD 주/VV + 는/ETM)))
      (NP_SBJ 말/NNG + 은/JX))
  (VP 없/VA + 다/EF + ./SF))

; 우리가 자라난 곳이 어머니의 품이라면, 고향은 어머니나 아버지까지를 포함한 더욱더 큰 품이랄 수
있다.
(S (S (NP_SBJ (S_MOD (NP_SBJ 우리/NP + 가/JKS)
                    (VP_MOD 자라나/VV + ㄴ/ETM))
      (NP_SBJ 곳/NNG + 이/JKS))
  (VNP (NP_MOD 어머니/NNG + 의/JKG)
      (VNP 품/NNG + 이/VCP + 라면/EC + ,/SP)))
  (S (NP_SBJ 고향/NNG + 은/JX)
  ...

```

(2) '의존구조구문분석말뭉치' 폴더



의존구조 구문분석 말뭉치는 본 관리도구를 사용하여 변환한 결과를 모아놓은 것이다. 어절번호 n_{dep} , 지배소번호 n_{head} , 기능태그, (어절별) 형태소분석결과의 순서로 표기되어 있으며, 그 의미는 “어절 n_{dep} 는 어절 n_{head} 의 기능(태그)이다”로 해석한다. 의존구조 구문분석 말뭉치에 나타난 예문 및 구구조 부착 결과는 다음과 같다. 예를 들어, “2 3 MOD”는 “어절2(깊은)은 어절3(꿈을)의 MOD(체언수식어)이다”라는 것을 의미한다. 기능태그의 경우, 이진 구구조 말뭉치에 부착된 기능태그는 그대로 계승하지만, 기능태그가 없으면 자동으로 추정하고 기능태그의 끝에 “*”를 부착한다. 기능태그의 부착 방식은 고려대학교 민족문화연구원 전자텍스트연구소와의 협의를 거쳐 결정된 것이다. 변환되는 의존구조 말뭉치에서 사용되는 기능태그는 다음과 같다.

기존 기능태그	SBJ(주어), OBJ(목적어), CMP(보어), MOD(체언수식어), AJT(용언수식어), CNJ(체언 접속어), INT(독립어), PRN(삽입어구)
추가된 기능태그 (자동으로 추정)	<ul style="list-style-type: none"> · SYM(기호): “, ”, /, ‘, ’와 같은 기호 · NOM(명사부류): 기능어가 없는 명사나 명사구 · LNK(연결서술어): 연결형 어말어미로 끝난 서술어 · PRD(종결서술어): 종결형 어말어미로 끝난 서술어

앞에서 본 이진 구구조 구문분석 말뭉치에 대한 의존구조 말뭉치는 다음과 같다.

; 고향의 품

1 2 MOD 고향/NNG + 의/JKG
2 0 NOM* 품/NNG

; 고향처럼 깊은 꿈을 주는 말은 없다.

1 4 AJT 고향/NNG + 처럼/JKB
2 3 MOD 깊/VA + 은/ETM
3 4 OBJ 꿈/NNG + 을/JKO
4 5 MOD 주/VV + 는/ETM
5 6 SBJ 말/NNG + 은/JX
6 0 PRD* 없/VA + 다/EF + ./SF

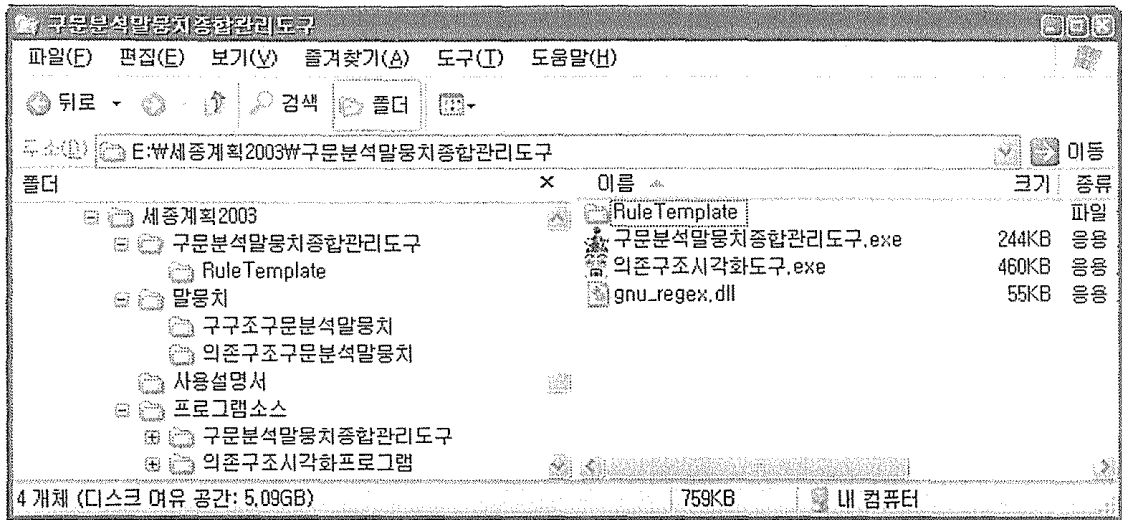
; 우리가 자라난 곳이 어머니의 품이라면, 고향은 어머니나 아버지까지를 포함한 더욱더 큰 품이랄 수 있다.

1 2 SBJ 우리/NP + 가/JKS
2 3 MOD 자라나/VV + ㄴ/ETM
3 5 SBJ 곳/NNG + 이/JKS
4 5 MOD 어머니/NNG + 의/JKG
5 14 LNK* 품/NNG + 이/VCP + 라면/EC + ./SP
6 14 SBJ 고향/NNG + 은/JX

...

다. ‘구문분석말뭉치종합관리도구’ 폴더

구문분석말뭉치종합관리도구 폴더에는 구문분석 말뭉치종합관리도구를 구동하는데에 필요한 실행파일과 구성되어 있다. 두 개의 실행파일이 들어 있는데, 하나는 구문분석말뭉치종합관리도구(구문분석말뭉치종합관리도구.exe)이고, 또 하나는 의존구조 시각화 도구(의존구조시각화도구.exe)이다. gnu_regex.dll은 종합관리도구 실행에 꼭 필요한 파일이므로, 같이 있어야만 한다.

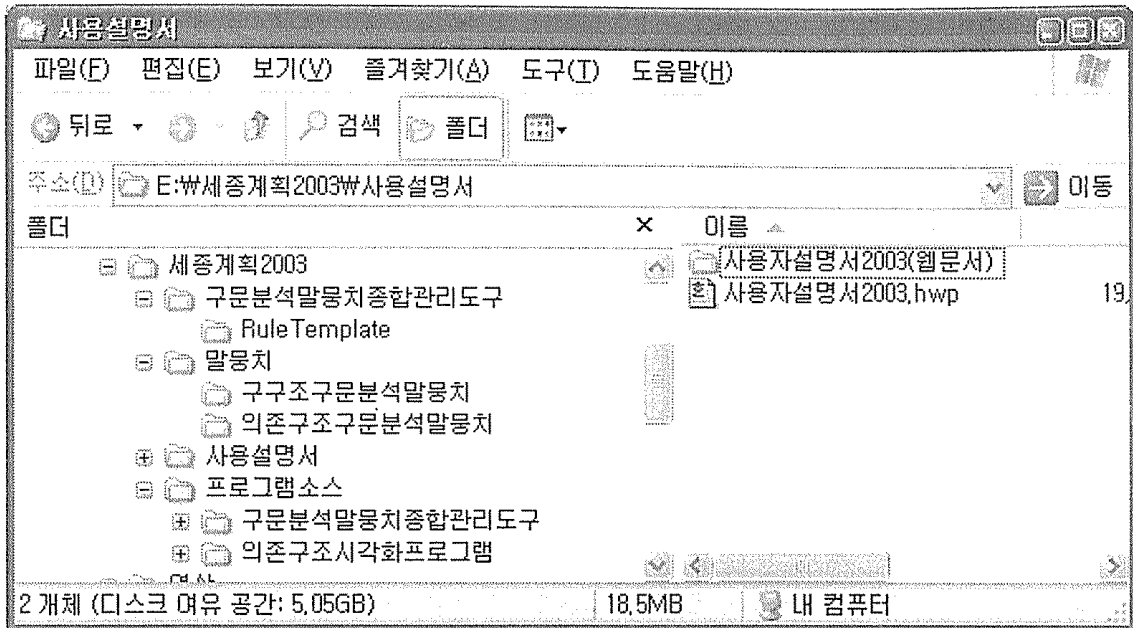


(1) 'RuleTemplate' 폴더

본 종합관리도구의 일부인 구문분석 말뭉치 수정도구가 사용하는 내부 규칙 템플릿(rule.list)을 담고 있는 폴더이다. 이 폴더에 들어 있는 규칙은 어절 합치기 혹은 구문구조 재생성시 사용되는데, 폴더가 없거나 이 폴더의 내용이 누락될 경우에는 말뭉치종합관리도구의 실행시 “규칙템플릿을 여는 데에 실패했습니다”라는 메시지가 나온다. 반드시 이 폴더는 실행파일이 있는 곳에 위치해 있어야 한다.

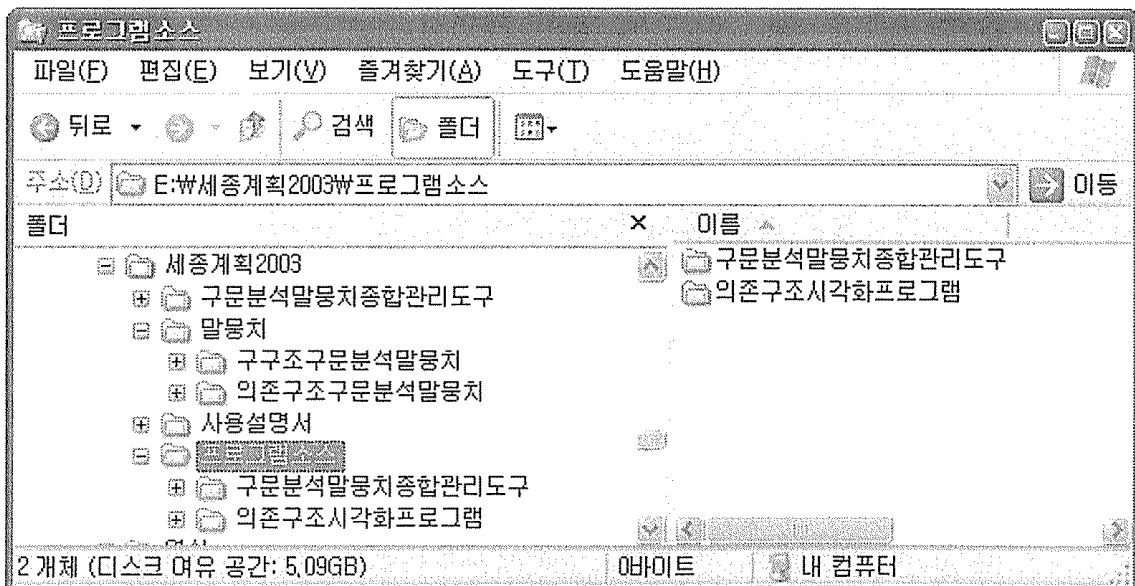
라. '사용 설명서' 폴더

'사용 설명서' 폴더에는 구문분석 말뭉치 종합관리도구의 사용방법을 자세히 기술한 사용설명서가 포함되어 있다. 사용설명서는 한글 97에서 읽을 수 있는 사용자설명서.hwp와 웹브라우저에서 읽을 수 있는 사용자설명서.htm로 구성되어 있으며, 두 파일의 내용은 동일하다.



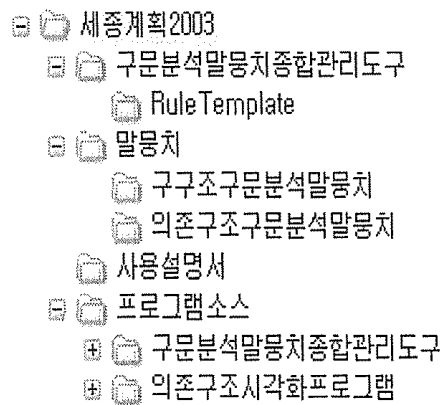
마. '프로그램 소스' 폴더

구문분석 말뭉치 종합관리도구에서 사용하는 구문분석말뭉치종합관리도구.exe와 의존구조시각화도구.exe에 관련된 프로그램 소스 파일과 프로젝트 파일을 포함하고 있다.



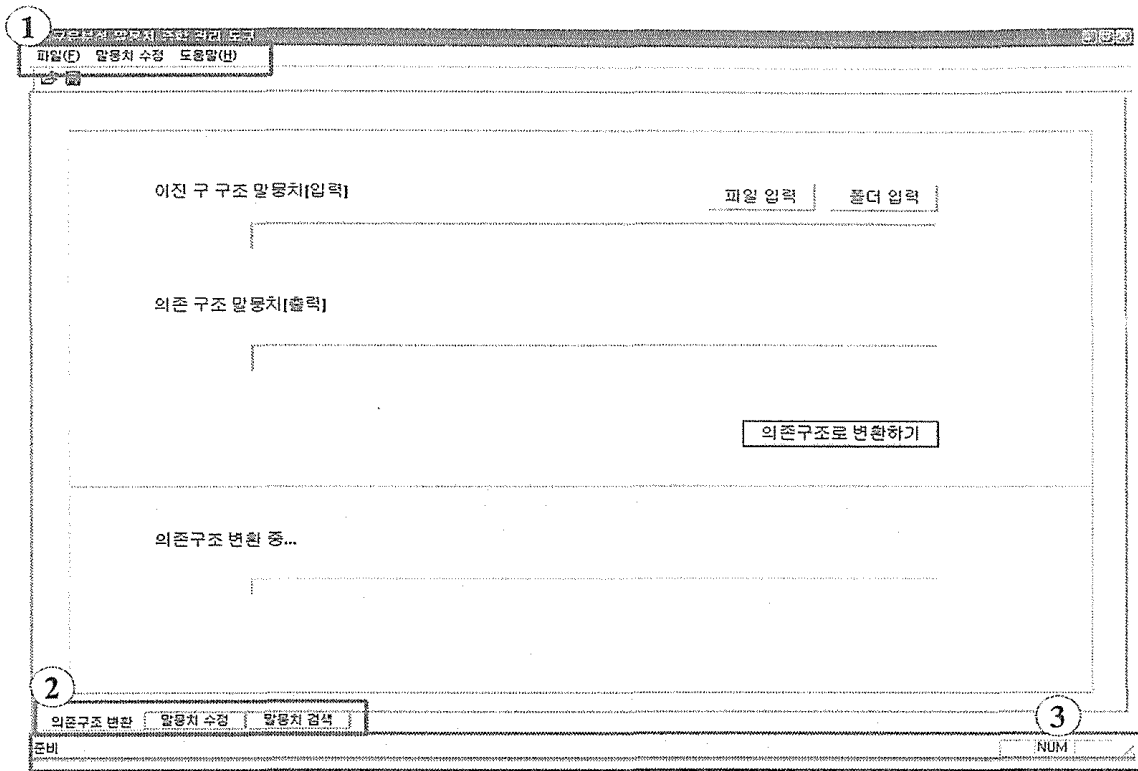
3. 사용 방법

구문분석 말뭉치 종합관리도구를 사용하기 전에 각각의 폴더 구조를 다음과 동일하게 조정한다. 이 폴더 구조는 반자동 구문분석 말뭉치 구축도구의 CD에 있는 폴더 구조와 동일하다. 그리고, 각 실행파일이 필요로 하는 폴더나 파일(예: RuleTemplate 폴더와 그 폴더 내의 내용이 제대로 구성되어 있는지를 확인한 후에 사용해야 한다.



가. 구문분석 말뭉치 종합관리도구를 처음 실행했을 때의 화면

‘구문분석말뭉치종합관리도구’ 폴더에 있는 구문분석말뭉치종합관리도구.exe를 실행시키면 다음과 같은 화면이 나타난다. (의존구조 시각화도구의 설명은 “3.2 구문분석 말뭉치 변환도구”에서 설명한다.)



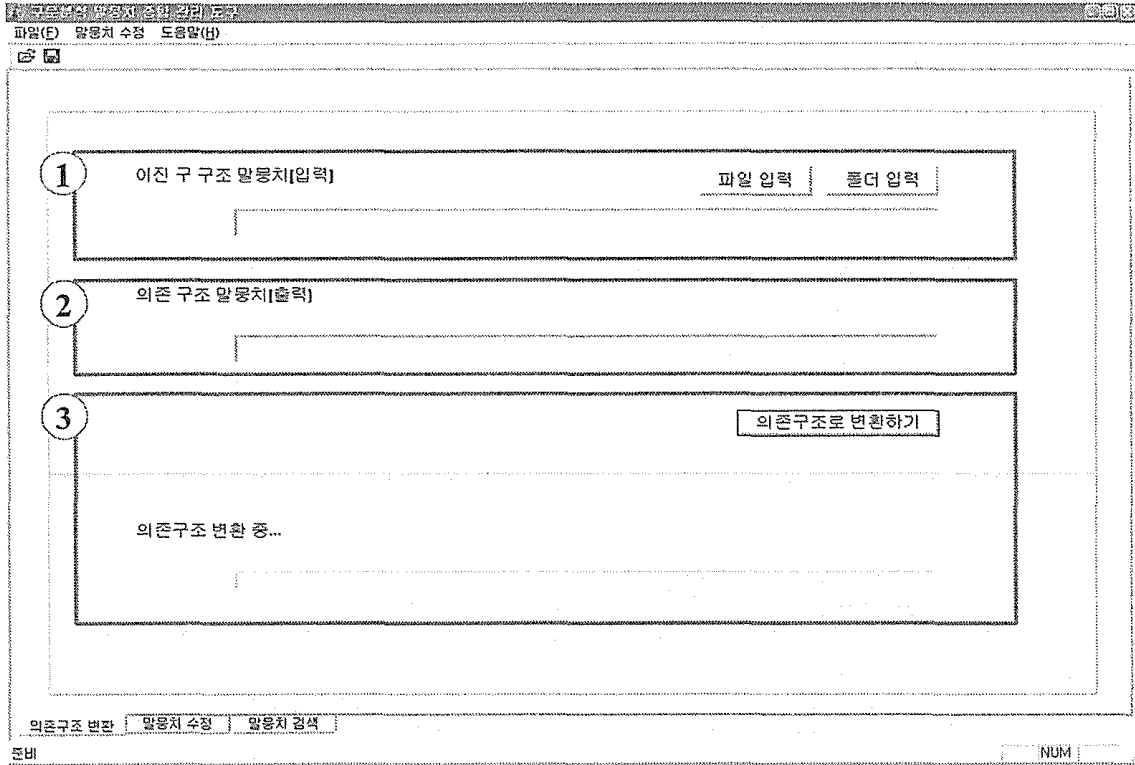
프로그램을 구성하는 각 구역의 기능을 간단하게 설명하면 다음과 같다.

- ❶ 메뉴바 : 이진 구구조 말뭉치 파일을 읽어들이거나 말뭉치 수정작업을 행할 때 사용하는 메뉴 항목들이 포함되어 있다. 둘째 부메뉴인 ‘말뭉치 수정’ 메뉴는 말뭉치 수정도구를 사용하고 있을 때에만 활성화된다.
- ❷ 작업 도구 변경 탭 : 변환도구/검색도구/수정도구 중 하나를 선택할 때 사용하는 탭으로서, 마우스버튼을 클릭하면 각각의 도구로 자유롭게 전환할 수 있다.
- ❸ 상태 표시 줄: 작업도구에 따라 사용자에게 보여주어야 하는 정보를 출력하는 부분이다. 검색도구를 예로 들면, 검색된 문장의 개수와 검색된 구문구조의 개수가 여기에 표시된다.

나. 구문분석 말뭉치 변환도구

구문분석 말뭉치 변환도구는 이진 구구조 형식의 구문분석 말뭉치를 입력으로

받아, 의존구조 형식의 구문분석 말뭉치로 변환하고, 그 결과를 구문분석 말뭉치 파일에 출력하는 실행 프로그램이다. 본 도구의 화면은 다음과 같고, 각 구역의 기능을 간단하게 설명하면 다음과 같다.



1 이진 구구조 말뭉치 입력부 : 의존구조로 변환하기를 원하는 이진 구구조 말뭉치가 있는 파일이나 폴더를 선택할 때 사용한다.

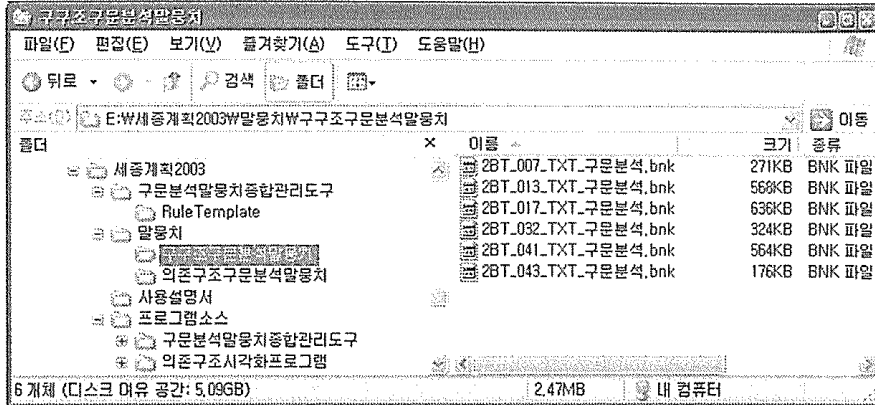
2 의존구조 말뭉치 출력부 : 이진 구구조 말뭉치로부터 변환된 의존구조 말뭉치의 위치와 이름을 나타내는 부분이다. 이진 구구조 말뭉치와 동일한 폴더에 의존구조 말뭉치가 생성된다.

3 의존구조 변환부 : 이진 구구조 말뭉치를 의존구조 말뭉치로 변환하는 부분이다. 그리고, 현재 이진 구구조 말뭉치가 의존구조 말뭉치로 얼마나 변환되었는지를 보여주는 기능도 제공한다.

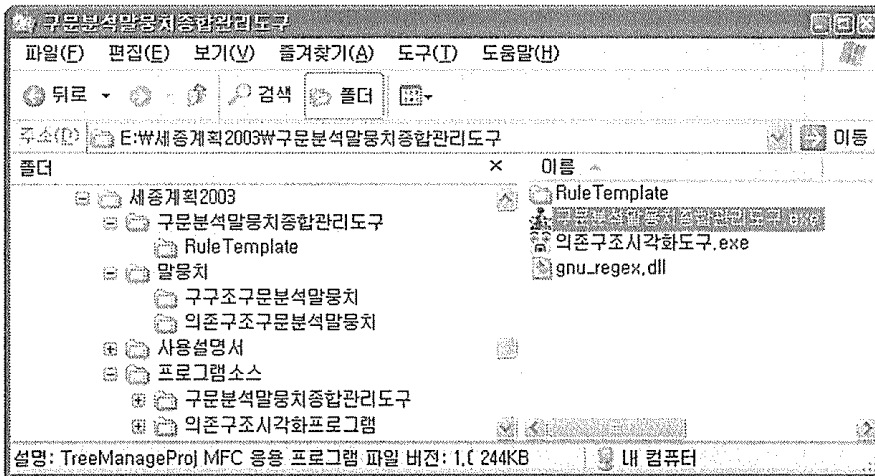
(1) 의존구조 말뭉치로의 변환과정

① '구구조구문분석말뭉치' 폴더에 다음과 같이 이진 구구조 말뭉치 파일(텍스

트 파일형식)이 있는지 확인한다. 없으면, 이진 구구조 말뭉치 파일을 찾아 가져다 놓는다.



② '구문분석말뭉치종합관리도구' 폴더에 있는 "말뭉치종합관리도구.exe"를 더블 클릭한다.

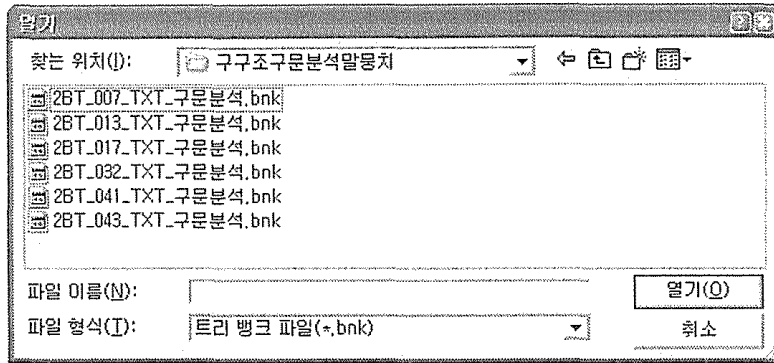


③ 말뭉치 종합 관리도구가 실행되면, 도구의 하단에 있는 "의존구조 변환" 탭을 클릭한다.

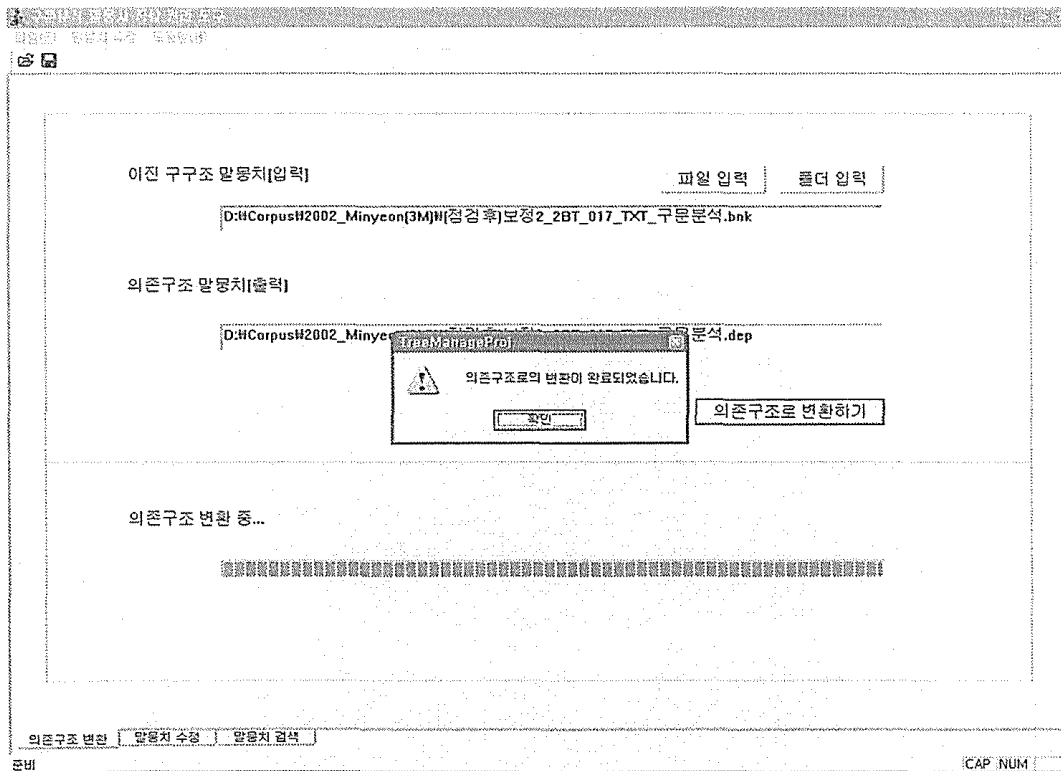


④ 의존구조 변환도구에서 이진 구구조 말뭉치[입력]에 있는 [파일 열기] 버튼을 클릭한다. 대화상자가 나타나면, 원하는 이진 구구조 말뭉치 파일을 클릭한다.

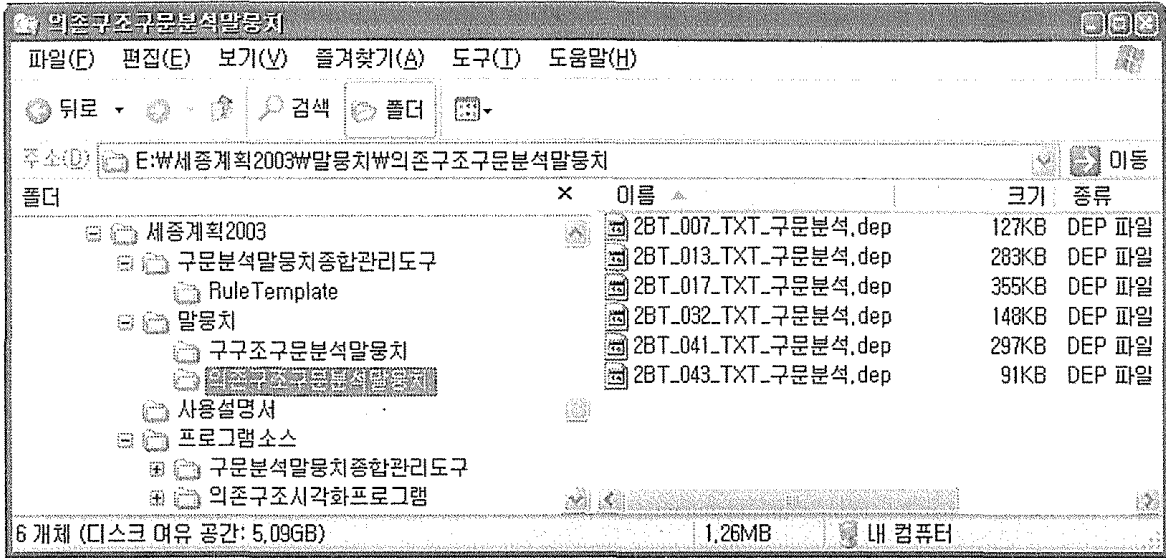
※ 만일, 여러 파일을 한꺼번에 변환하고 싶으면 폴더 열기 버튼을 클릭한다.



⑤ 의존구조 변환도구에서 [의존구조로 변환하기] 버튼을 클릭한다.

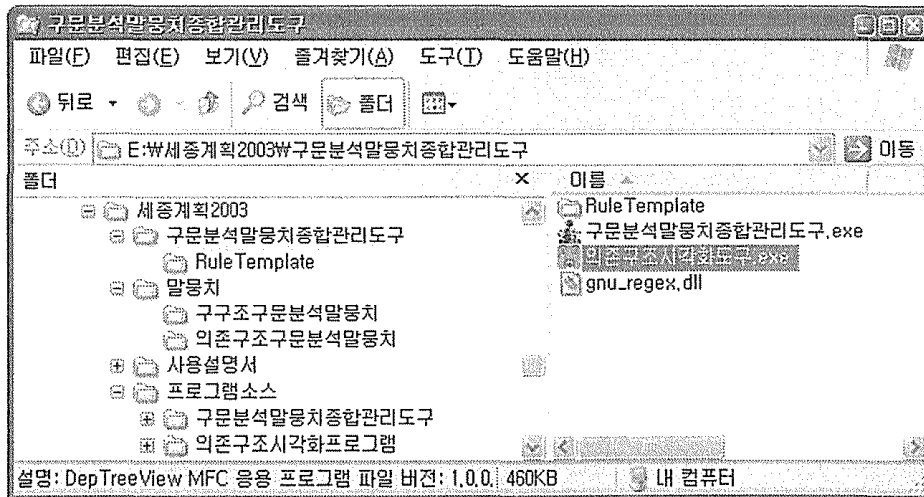


⑥ 의존구조 변환 작업이 종료되면, '구구조구문분석말뭉치' 폴더에 의존구조로 변환된 파일(텍스트 파일 형식의 *.dep 파일)이 생성되었는지 확인한다. (아래 그림은 변환된 의존구조 파일을 '의존구조구문분석말뭉치' 폴더로 복사한 것이다.)

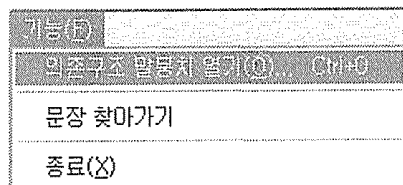


(2) 의존구조 말뭉치의 시각화

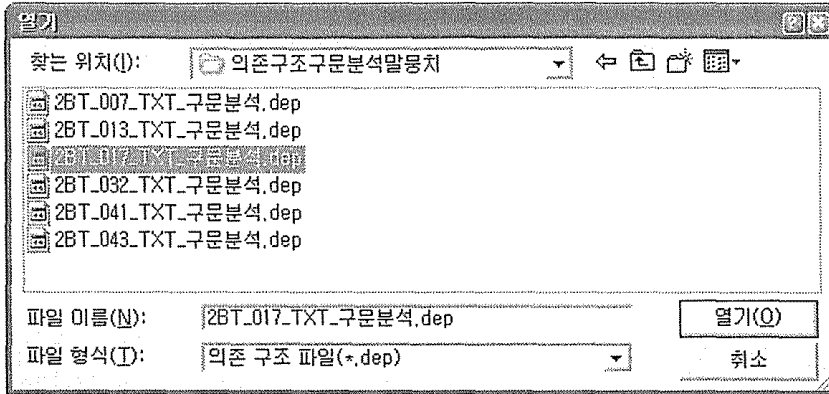
① '구문분석말뭉치종합관리도구' 폴더에 있는 "의존구조시각화도구.exe"를 더블 클릭한다.



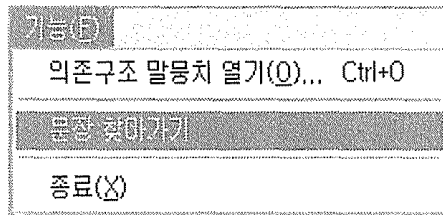
② 메뉴에서 기능을 클릭하고, '의존구조 말뭉치 열기' 항목을 클릭한다.



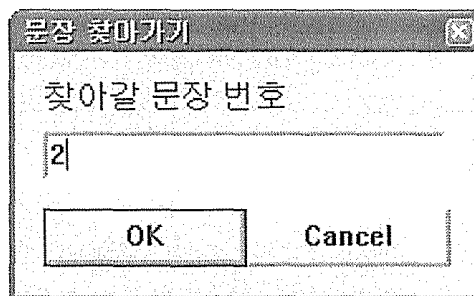
③ '파일 열기' 대화상자가 나타나면, 원하는 의존구조 말뭉치 파일을 클릭한다.



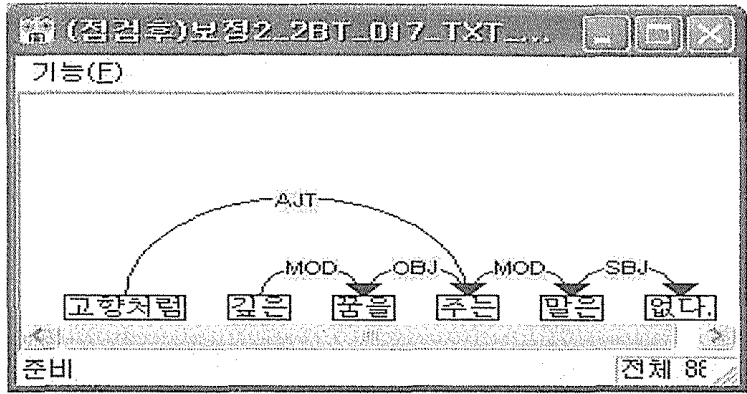
④ 메뉴에서 기능을 클릭하고, '문장 찾아가기' 항목을 선택한다.



⑤ 파일 열기 대화상자가 나타나면, 찾아갈 문장 번호를 입력하고 [OK] 버튼을 클릭한다.

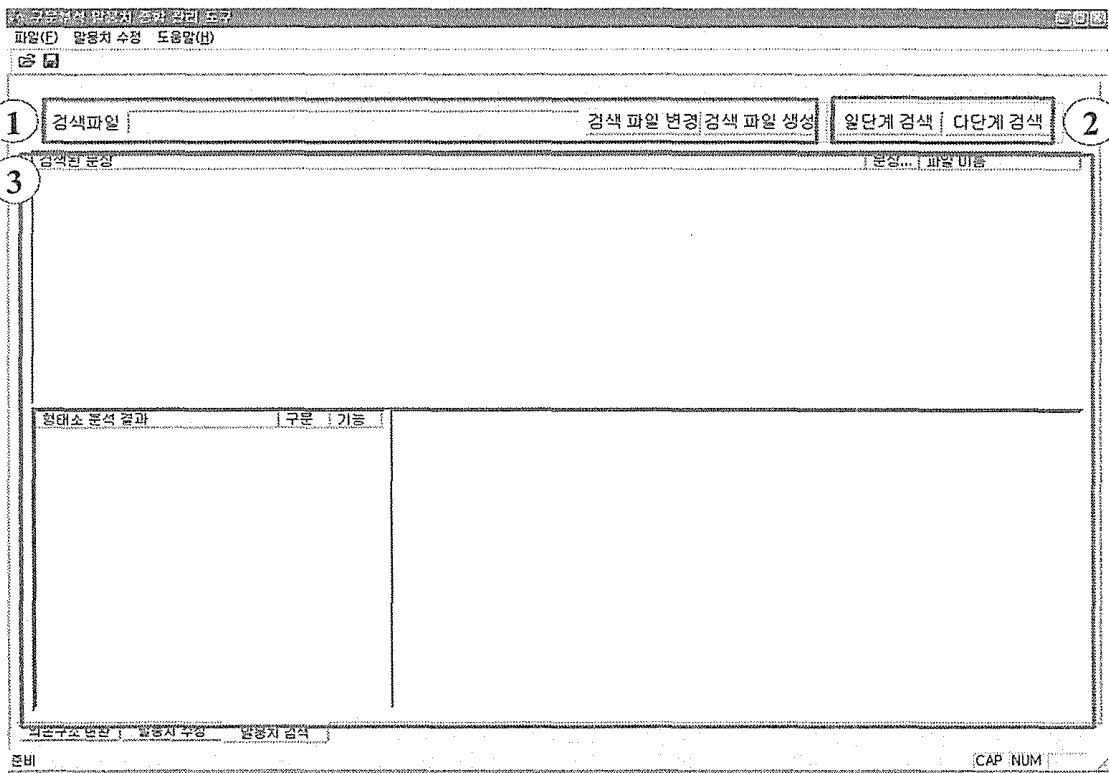


⑥ 찾은 문장의 의존구조를 확인한다. 의존구조 시각화도구는 구축된 의존구조 말뭉치를 눈으로 쉽게 확인하는 것만 가능하고, 구조 수정 기능 등은 제공하지 않는다.



다. 구문분석 말뭉치 검색도구

이진 구구조 말뭉치에서 원하는 조건에 맞는 구구조 (구문노드)를 검색하는 도구이다. 빠른 처리를 위해 색인 탐색만을 지원한다. 즉, 사용자는 이진 구구조 말뭉치(.bnk)를 지정하여 검색 파일(.idx)를 구축한 후에 이 파일을 통해 검색을 수행한다. 검색 조건으로는 일단계 유도(simple/one-level derivation)식을 기본으로 제공하며, 다단계 유도의 검색까지 지정할 수 있게 되어 있다. 본 도구의 화면은 다음과 같고, 각 구역의 기능을 간단하게 설명하면 다음과 같다.



❶ 검색파일변경/검색파일생성 버튼 : 이진 구구조 말뭉치 파일에 대한 기존의 검색 파일을 변경하거나 새로 작성할 때에 사용하는 버튼이다.

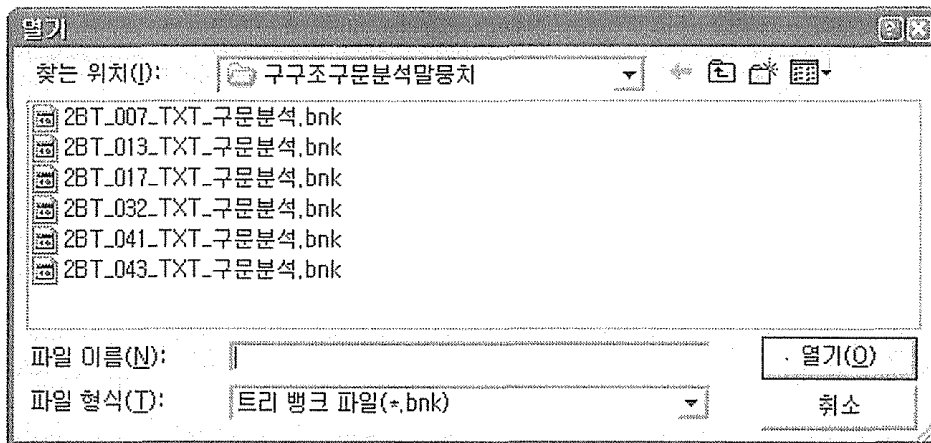
❷ 일단계검색/다단계검색 버튼 : 검색 조건을 입력해 넣을 때에 사용하는 버튼이다. '일단계검색'은 “부모 > 왼쪽자식 오른쪽자식” 형태의 유도형식을 GUI를 통해 입력할 때에 사용하며, '다단계검색'은 좀더 복잡한 검색식을 입력할 때에 사용한다.

❸ 검색결과창: 사용자가 입력한 검색조건에 맞는 결과를 출력해주는 창이다.

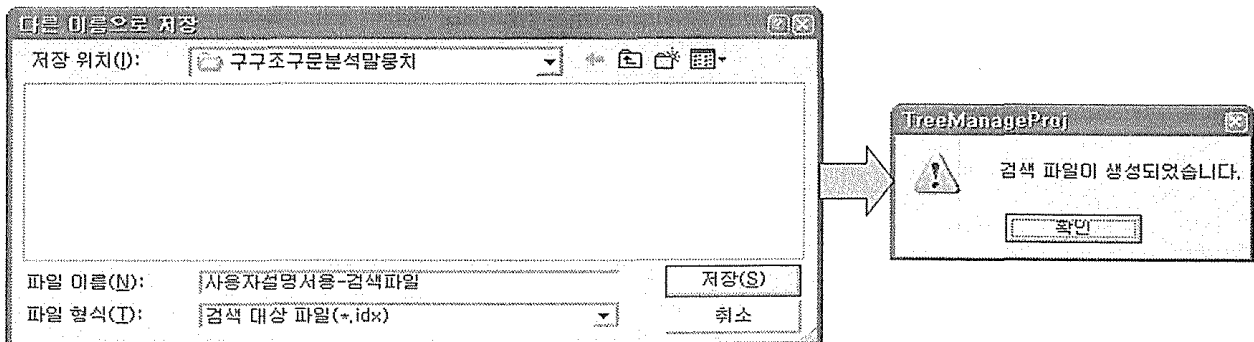
크게 윗부분과 아랫부분으로 나뉘는데, 윗부분은 검색한 조건에 맞는 구구조를 포함한 문장의 리스트가 출력되는 창이며, 아랫부분은 사용자가 검색문장 리스트에 있는 문장 들 중 하나를 선택했을 때 그 문장에 대한 세부 구구조가 출력되는 창이다.

(1) 검색 파일 생성

① [검색파일생성] 버튼을 클릭하면, 다음과 같은 대화상자가 나타나며, 검색 파일을 만들고자 하는 이진구구조 말뭉치 파일을 선택하고 [열기] 버튼을 클릭한다. CTRL 키를 누르고 여러 개의 파일을 동시에 선택할 수도 있다.

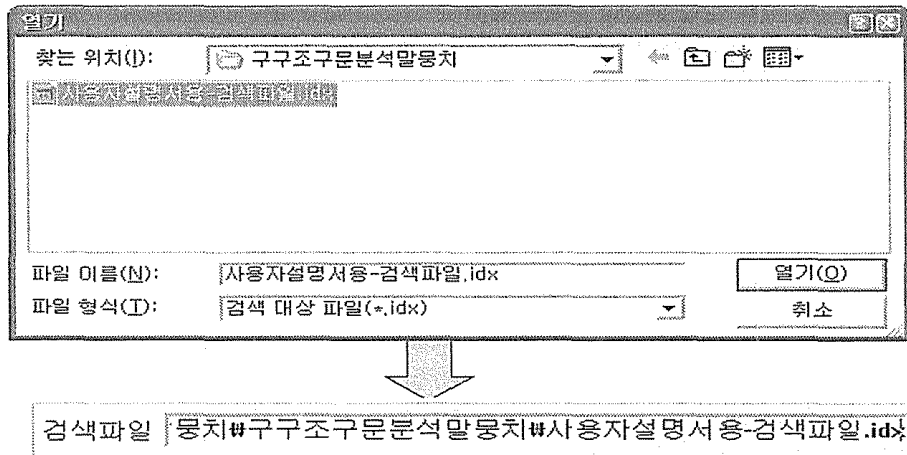


② 작성할 검색파일의 이름을 물어보는 대화상자가 나타나는데, 적당한 이름을 지정하고(기본적으로 *.idx 확장자가 만들어진다) 다시 [저장] 버튼을 클릭하면, 약간의 시간이 흐른 뒤에 검색 파일이 작성되며, 완성을 알리는 메시지 상자가 나타난다.



(2) 검색 파일 지정

① [검색파일변경] 버튼을 클릭하면, 다음과 같은 대화상자가 나타난다. 검색에 사용할 검색파일을 선택하고 [열기] 버튼을 클릭하면, 선택된 검색 파일의 경로명이 왼쪽 편집 창에 나타난다.

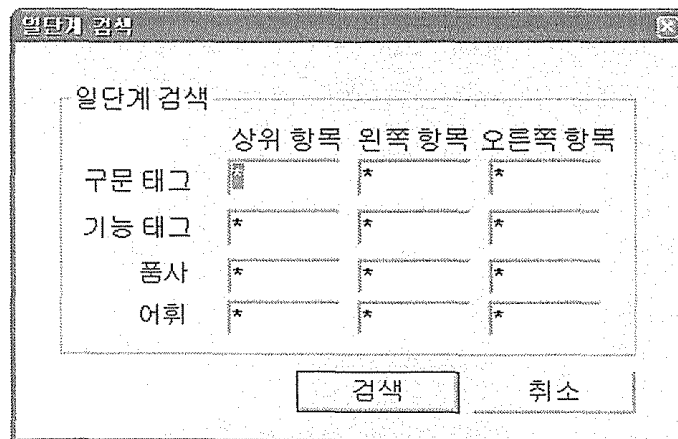


② 참고: 왼쪽 편집 창에는 가장 최근에 만든 검색파일의 경로명이 나타나 있는데, 이 파일을 그대로 사용하려면 굳이 검색파일 지정을 해줄 필요가 없다.

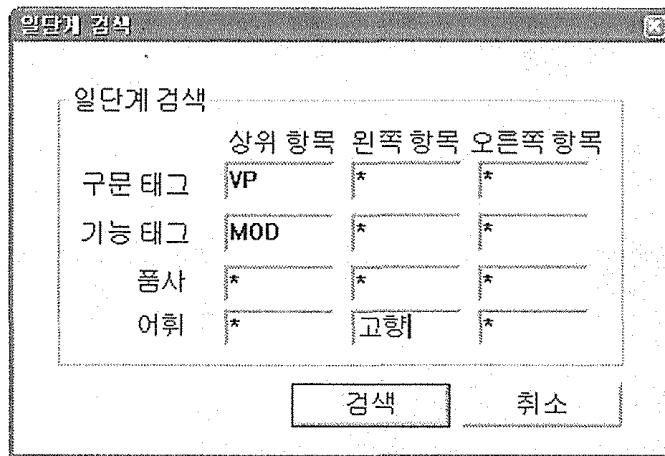
(3) 검색 조건 지정 / 검색 실행

① 일단계 검색

㉠ [일단계검색] 버튼을 클릭한다. 다음과 같은 대화상자가 나타난다



㉔ 기본적으로 모든 항목의 모든 조건 정보에는 와일드카드(*)가 붙어 있는데, “해당 조건에 상관하지 않는다”라는 뜻이다. 조건을 지정해 주고 싶은 부분을 선택하여, 직접 키보드를 사용하여 해당 정보를 입력해 넣는다. 다음 그림은 부모노드의 구문태그와 기능태그가 VP와 MOD인 동시에, 왼쪽 자식 노드가 “고향”이라는 어휘를 포함한 구조를 지정한 예이다.



㉕ 검색된 결과가 검색결과창에 나타난다. 앞에서 설명했듯이 윗부분에는 검색된 문장의 엔트리 리스트(문장, 문장번호, 해당문장을 포함한 파일의 경로명)가 나열되며, 아랫부분에는 검색된 문장들 중에 사용자가 선택한 한 문장의 구문 구조가 표시된다. 검색된 구문 구조에서는 해당 문장에서 조건을 만족하는 구구조의 최상위 부모 노드가 강조되어 표시된다. 또한 도구의 상태표시줄에는 검색된 문장의 개수와 검색된 구문 구조관련된 간단한 통계정보가 명시된다. 아래의 그림은 앞에서 지정한 검색 조건을 입력하여 검색을 실행한 예이다.

구구조 분석 엔진의 구문 트리 도구

파일(F) | 말뭉치 수정 | 도움말(H)

검색파일 | 동치#구구조구문분석말뭉치#사용지설명서용-검색파일.idx | 검색 파일 변경 | 검색 파일 생성 | **일단계 검색** | 다단계 검색

검색된 문장

문장...	파일 이름
광경남도 예산군이 고향인 김철중 씨(60사건관 운영 김현도 춘천시 효자동)는 요즘 고향에 살아 계산 어머니(전희옥 씨 60)를 만나...	28T_007.TXT-구...
어머니가 책꽂이에 생분해 있다는 소식을 들은 김씨는 작년 여름 고향 동네와 안쪽길 상류를 사이에 두고 잡해 있는 중국 음식점 장...	28T_007.TXT-구...
김씨는 다음에는 김씨가 아니라 고향마을과 동국학을 잇는 다리(배정교) 중간에서 어머니를 포옹하는 환상을 자꾸만 떠올리고 있...	28T_007.TXT-구...
고향을 그리워하는 데에는 할던 사람도 본던 사람도 구별이 없고, 젊은 사람도 젊은 사람도 차이가 없다.	28T_017.TXT-구...
떠난 자가 오랜 사람이면 오랜 사람일수록, 떠나 있는 곳이 먼 사람이면 먼 사람일수록 고향에 대한 그리움은 깊어지게 된다.	28T_017.TXT-구...
우리 나라 국토 8만대정맥이 있는 곳이라면, 으름 활천 해인사? 하고 그제서야 겨우 고개를 끄덕일까말까 한 그런 시골이지만, 그러...	28T_017.TXT-구...

형태소 분석 결과

형태소	구문	기능
고향/NING + 처럼/KB	NP	AJT
갔다/VA + 은/ETM	VP	MOD
꿈/NING + 을/IKO	NP	OBJ
주/VX + 는/ETM	VP	MOD
말/NING + 은/IX	NP	SBJ
있/VA + CI/EF + /SE	VP	

구문 트리 다이어그램

```

NP_A |
  |
  +--- VP_M --- NP_O --- VP_M --- VP_M --- NP_S --- S
  |
  +--- NP_O --- VP_M --- VP_M --- NP_S --- S
  |
  +--- VP_M --- NP_S --- S
  |
  +--- NP_S --- S
  |
  +--- VP --- S
  
```

의문구조 변환 | 말뭉치 수정 | 말뭉치 검색

색인된 전체 문장 수 : 2559, 검색된 문장 수 : 7, 검색된 표현 수 : 6

CAP_NUM

② 다단계 검색

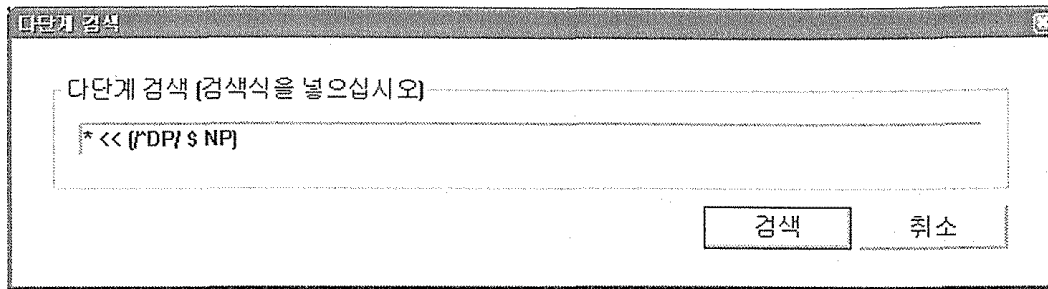
㉠ 다단계 검색 버튼을 클릭한다. 다음과 같은 대화상자가 나타난다.

다단계 검색

다단계 검색 (검색식을 넣으십시오)

검색 취소

㉡ 편집 상자에 검색식을 입력한다. 검색식의 형식은 Pennsylvania 대학에서 만든 표준 구구조 탐색 도구인 Tgrep에서 정의한 형식과 완전히 호환된다(검색식 사용법에 관련된 내용은 4장 '부록'에서 참고한다). 아래 그림에 입력한 검색식은 "DP"라는 문자열이 들어있는 왼쪽 노드와 자매관계를 가지는 오른쪽 노드 "NP"를 동시에 지배하고 있는 모든 구문 구조를 가리킨다.



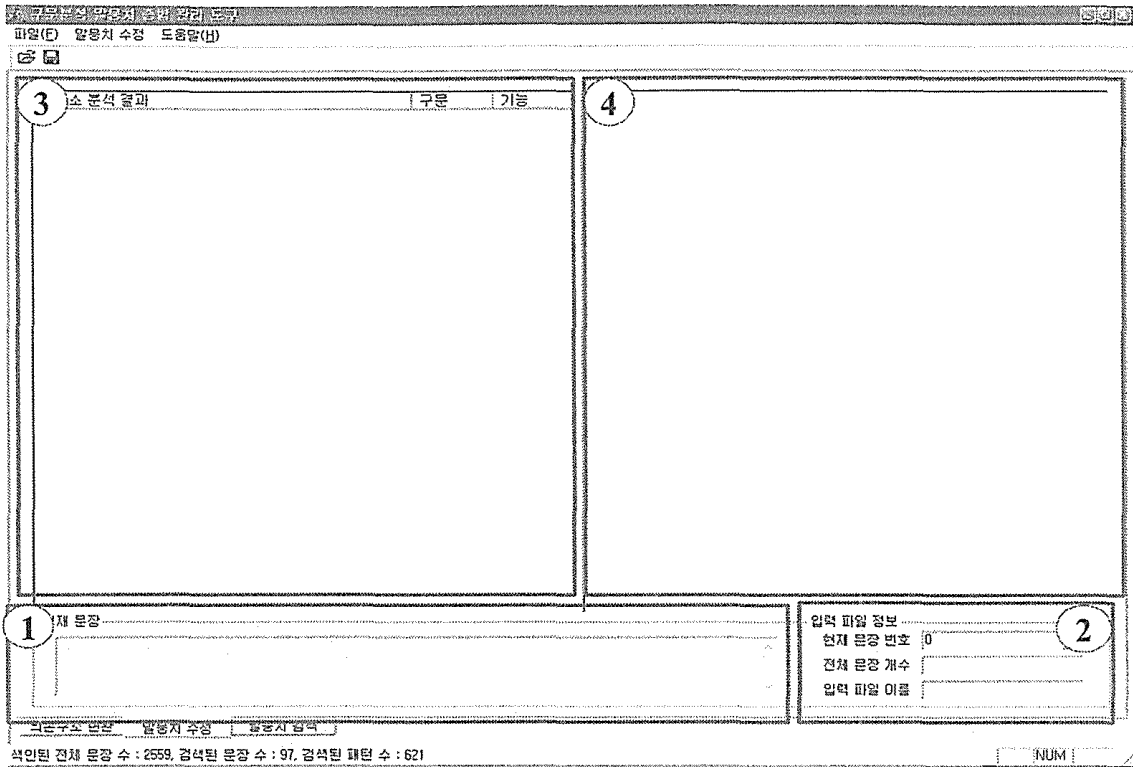
㉔ 검색된 결과가 검색결과창에 나타난다. 일단계검색을 수행할 때와 동일하다.

(4) 말뭉치 수정으로의 전환

검색결과가 출력된 후, 검색결과창의 문장 리스트 중 하나를 마우스로 더블클릭하면, 클릭된 문장이 말뭉치 수정도구로 보내어짐과 동시에 작업도구가 말뭉치 수정도구로 바뀐다. 이 기능은 말뭉치 오류 수정을 작업하고 있을 때에 대단히 편리하게 사용할 수 있다. (말뭉치 수정 도구 및 제공 기능에 대해서는 3.4를 참고한다)

라. 구문분석 말뭉치 수정도구

기존에 만든 구문구조를 수정하는 기능은 트리 बैं크 파일을 직접 읽거나, 검색도구에서 검색된 문장을 더블 클릭하는 경우 들어오게 된다. 구문분석 말뭉치의 수정기능은 크게 1)구문 구조 수정, 2)구문태그·기능태그 수정, 3)원어절 수정으로 나뉘어져 있다. 본 도구의 화면은 다음과 같고, 각 구역의 기능을 간단하게 설명하면 다음과 같다.



1 문장 디스플레이 패널 : 현재 작업 중인 문장을 보여주는 패널로서, 형태소 분석 결과가 붙어 있지 않은 원시 문장이 표시된다.

2 파일 정보 디스플레이 패널 : 현재 작업 중인 파일(.BNK)의 정보를 보여주는 패널로서, 작업 대상 문장의 순서(번호), 해당 파일에 들어 있는 문장의 전체 개수, 작업 파일의 이름이 순서대로 표시된다.

3 형태소 분석결과 디스플레이 패널 : 현재 작업 중인 문장의 형태소 분석 결과를 보여주는 패널이다. 한 줄에 한 어절 단위로 형태소 분석 결과, 구문 표시, 기능 표시가 표시된다.

4 구문 구조 디스플레이 패널 : 작업 중인 문장의 구문 구조를 시각적으로 보여주는 패널이다.

구문 구조를 수정하기 위해서는 메뉴의 '말뭉치 수정' 메뉴를 사용한다. 각각의 메뉴 항목에 대한 설명은 아래와 같다.

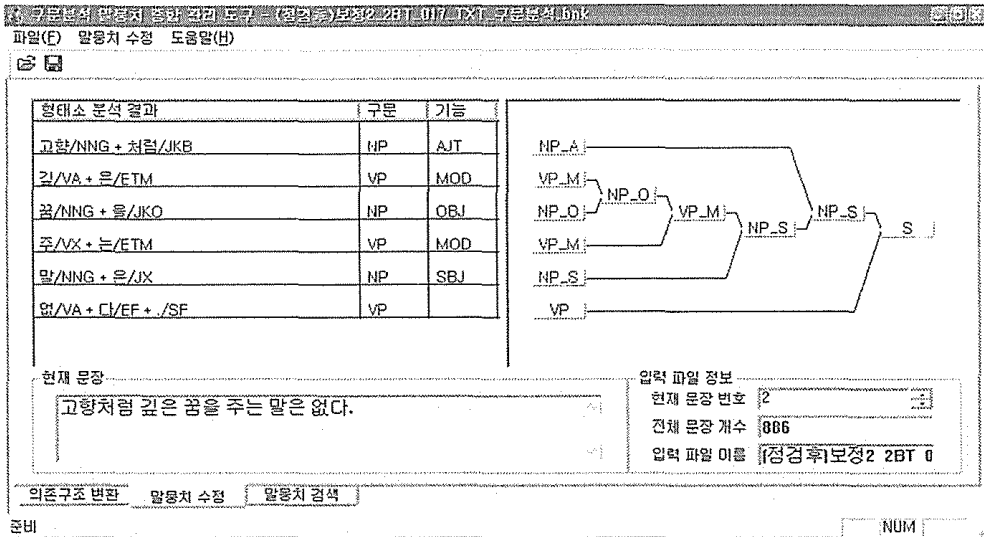
말뭉치 수정	도움말(H)
어절 쪼개기	
어절 합치기	
뒤로 가기(B)	
이동 하기(N)	
묶기(오른쪽 투사D)	
묶기(왼쪽 투사L)	
구문 구조 해체(C)	

- ❶ '어절 쪼개기' : 구구조 구문분석 트리에 있는 1개의 원어절(단말 노드)을 2개의 원어절(단말 노드)로 분리한다.
- ❷ '어절 합치기' : 구구조 구문분석 트리에 있는 2개의 원어절(단말 노드)를 1개의 원어절(단말 노드)로 합친다. 단, 합쳐지는 2개의 노드는 같은 부모 노드를 공유해야만 한다.
- ❸ '뒤로 가기' : 구문 구조를 수정할 때 사용하는 메뉴로, 묶기/이동 재수행시 잘못된 연산을 취소하는 역할을 담당한다. (단축키 : 'z')
- ❹ '이동 하기' : 구문 구조를 수정할 때 사용하는 메뉴로, 두 노드를 묶지 않고 다음 단계로 진행하는 역할을 담당한다. (단축키 : 'x')
- ❺ '묶기(오른쪽 투사)' : 구문 구조를 수정할 때 사용하는 메뉴로, 두 노드를 묶는 역할을 담당한다. 오른쪽 투사 항목은 부모노드의 구문태그·기능태그를 오른쪽 자식 항목에서 투사된 것으로 설정한다. (단축키 : 'c')
- ❻ '묶기(왼쪽 투사)' : 구문 구조를 수정할 때 사용하는 메뉴로, 두 노드를 묶는 역할을 담당한다. 오른쪽 투사 항목은 부모노드의 구문태그·기능태그를 왼쪽 자식 항목에서 투사된 것으로 설정한다. (단축키 : 'd')
- ❼ '구문 구조 해체' : 구문 구조를 수정할 때 사용하는 메뉴로, 트리 내의 부분적인 구문 구조를 없애고 재구축 할 수 있게 하는 역할을 담당한다. (단축키 : 'v')

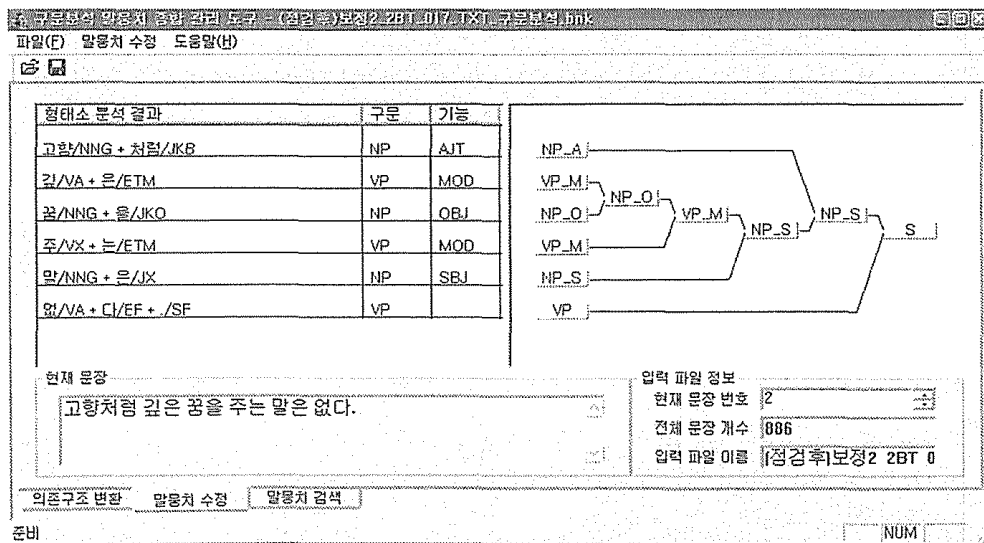
(1) 구문 구조 수정

① 수정할 구조 선택 : 수정이 가해질 구문 구조의 최상위 부분을 선택한다. 이때 해당 범주와 이에 속하는 모든 최하위 범주의 색깔이 바뀌어 표시된다. 아래의 그림의 경우 “깊은 꿈을 주는”과 “고향처럼”이 먼저 묶이고, 다음에 “말은”과 묶여야 올바른 구문구조이다. 그러나, “깊은 꿈을 주는”과 “말은”이 먼저 묶이고 다음에 “고향처럼”과 묶였기 때문에 잘못된 구조이다. 따라서 최

상위범주인 "고향처럼 깊은 꿈을 주는 말은"에 해당하는 NP_SBJ 노드를 클릭한다.

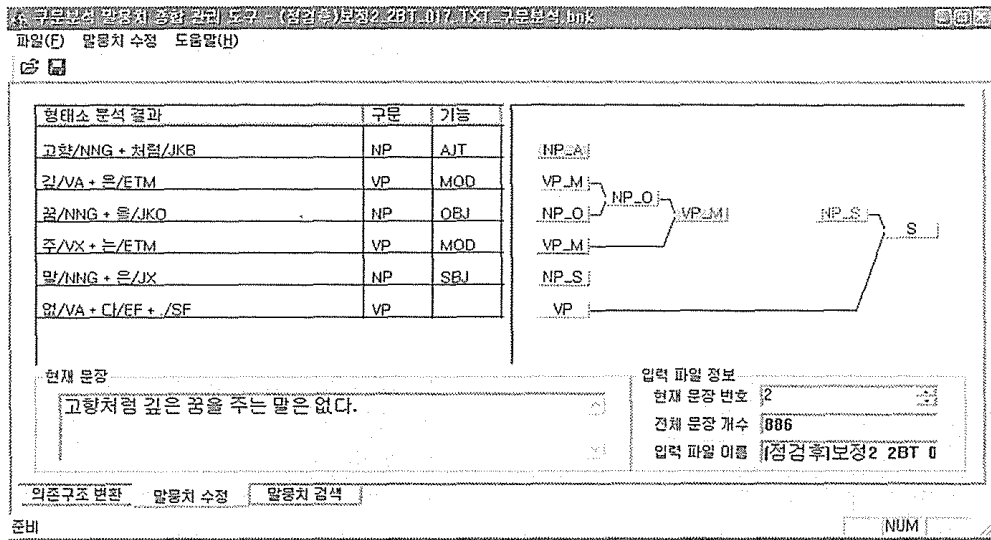


② 재표시 : 선택된 범주의 내부에 있는 구문 구조 중에, 옳은 구조를 가진 상위 범주(node)를 마우스의 왼쪽 버튼을 클릭하여 선택하여 재표시한다. 선택된 노드는 하이라이팅되어 표시되고, 하얀색으로 표시된 부분은 나중에 구문 구조 해체 시 해체되지 않는다. 위의 예제의 경우, "깊은 꿈을 주는"에 해당하는 구문구조는 올바른 구문구조이기 때문에, 이에 해당하는 VP_MOD의 노드를 클릭하여 재표시를 수행한다.

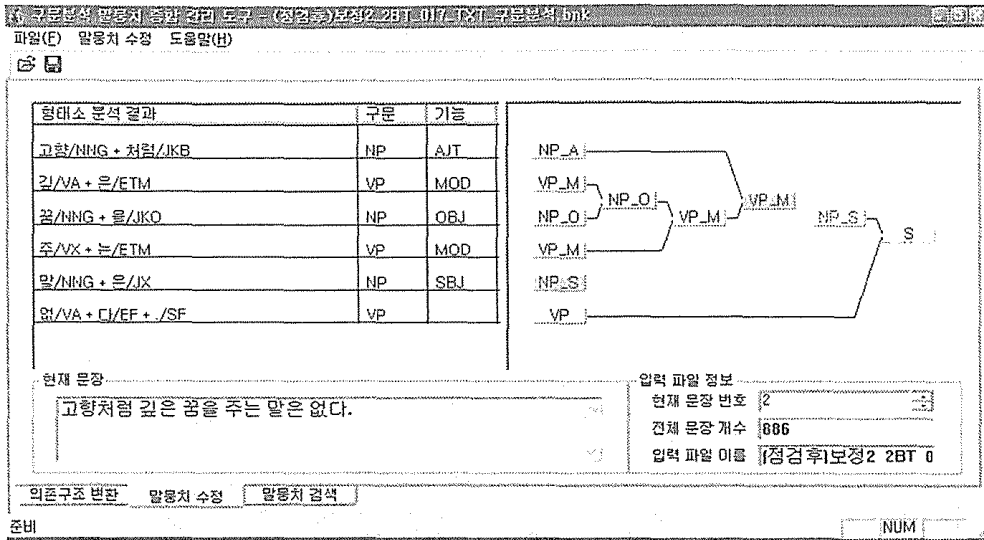


③ 구문 구조 해체 : 재표시되지 않은 모든 구문 구조를 해체하여 초기화하

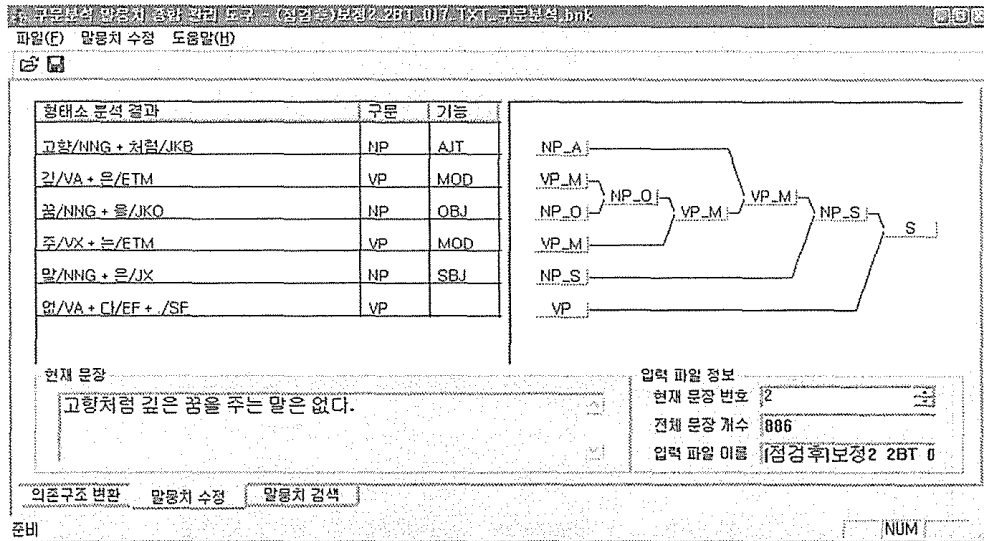
는 것으로, 이때 메뉴의 ‘말뭉치 수정’ -> ‘구문 구조 해체’ 항목을 선택한다. 이때 선택된 노드 사이의 노드들은 삭제되고, 선택된 노드 상·하위에 있는 노드들은 삭제되지 않는다. 그리고 묶일 후보들이 파란색으로 표시되어 사용자의 결정을 기다린다.



④ 이동/묶기 제수행 : 아래 그림과 같이 이동/묶기를 결정해야 하는 두 개의 노드는 파란색으로 표시된다. 두 개의 파란색의 노드가 묶이는 것이 옳바르다면 “말뭉치 수정”내의 ‘묶기(오른쪽 투사)’ 항목을 선택하거나 ‘c’키를 눌러준다. (왼쪽 하위 항목이 투사되어야 하는 경우는 ‘묶기(왼쪽 투사)’를 선택하거나 ‘d’키를 눌러준다.) 그리고, 두 개의 파란색 노드가 묶이는 것이 옳바르지 못하다면, ‘말뭉치 수정’내의 ‘이동 하기’ 항목을 선택하거나 ‘x’키를 눌러준다. 이와 같은 묶기/이동의 연산을 통하여 해체된 부분 구문 구조를 재구축하여 준다. 위의 예제에서는 “고향처럼”과 “깊은 꿈을 주는”이 묶이고, 그 후에 “말은”이 묶여야 하기 때문에, ‘묶기(오른쪽 투사)’ 항목을 두 번 연속해서 수행한다.

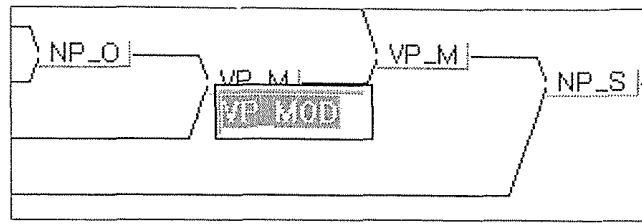


⑤ 구문 구조 수정 완료 : 위와 같은 단계를 거쳐 구문 구조를 수정하게 되면 아래 그림과 같이 “깊은 꿈을 주는”과 “고향처럼”이 먼저 묶이고, 다음에 “말은”과 묶이는 올바른 구문구조를 만들 수 있다.



(2) 구문태그 · 기능태그 수정

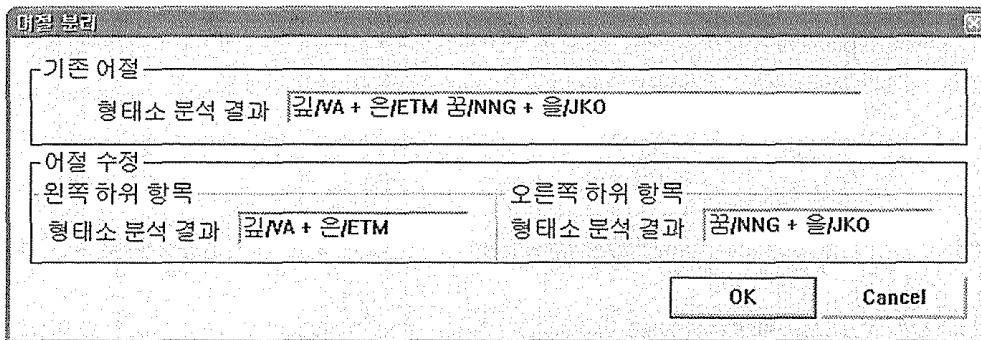
기존에 작성된 구문구조 중 잘못 태그가 할당된 노드를 수정하기 위해서는, 해당 노드를 마우스 오른쪽 버튼으로 클릭한다. 그러면, 아래 그림과 같이 구문태그 · 기능태그를 수정할 수 있는 조그마한 창이 뜨고, 이 창에 새로운 구문태그 · 기능태그를 입력하여 태그를 수정한다.



(3) 원어절 수정

기존에 작성된 구문구조 중에서 원(raw) 어절을 수정할 필요가 있을 경우, 원어절 수정기능을 사용한다. 원어절 수정기능에는 하나의 어절을 두개로 쪼개는 ‘어절 쪼개기’와 두 개의 어절을 하나의 어절로 합치는 ‘어절 합치기’의 두 가지 기능이 있다. 단, ‘어절 합치기’에서 주의할 점은 합쳐질 두 개의 어절은 반드시 같은 부모 노드를 공유해야 한다는 것이다. 따라서, 서로 부모 노드가 다른 두 개의 인접한 어절을 합치기 위해서는 구문구조 수정을 먼저 수행하여 두 노드의 부모를 동일하게 수정한 후에, ‘어절 합치기’ 기능을 수행하여야 한다.

① 어절 쪼개기 : “깊은 꿈을”은 두 어절로 구문분석 되는 것이 올바르지만, 하나의 어절로 분석되었을 경우, 아래와 같은 ‘어절 쪼개기’의 기능을 사용하여 “깊은”과 “꿈을”의 두 개의 원어절로 분리하여 준다. 어절 쪼개기를 수행하기 위해서는 쪼개기를 수행할 어절을 선택한 후, 메뉴의 ‘말뭉치 수정’에서 ‘어절 쪼개기’ 항목을 선택한다.



② 어절 합치기 : “고향처럼”은 하나의 어절로 구문분석 되는 것이 올바르지만, “고향”과 “처럼”의 두 어절로 분석되었을 경우, 아래와 같은 ‘어절 합치기’의 기능을 사용하여 “고향처럼”으로 하나의 원어절로 합쳐준다. 어절 합치기를 수행하기 위해서는 합치기를 수행할 2개의 어절을 선택한 후, 메뉴의 ‘말뭉치 수정’에서 ‘어절 합치기’ 항목을 선택한다.

어절 합치기

기존 어절

왼쪽 하위 항목	오른쪽 하위 항목
형태소 분석 결과 고향/NNG	형태소 분석 결과 + 처럼/JKB

어절 수정

형태소 분석 결과 고향/NNG + 처럼/JKB

OK Cancel

4. 부록

가. 구문분석 말뭉치 검색도구에 사용되는 다단계 검색식 사용법

본 절에서는 말뭉치 검색도구의 '다단계검색' 기능을 사용할 때에 입력하는 검색식의 사용법을 설명한다. 본 검색식은 PennTreeBank를 검색하도록 만들어진 도구인 Tgrep이 제시하는 검색식과 호환되므로, Tgrep을 잘 알고 있거나 검색식을 사용할 수 있는 사람은 읽을 필요가 없다. Tgrep 설명서의 원전은 책 한권으로도 모자랄 정도로 방대한 분량을 자랑하지만, 본 절에서 설명하는 내용은 이분지(binary-branching) 구문구조를 기술하는 데에 자주 사용되는 일부만을 중심으로 간추린 것이다.

나. 기본 원칙

- 검색식은 공백문자(space)를 포함하여 총 2048byte를 넘을 수 없다.
- 별개의 노드 이름과 연산자는 구분을 위하여 공백문자로 띄어쓴다. 단 괄호([,],(,))는 띄어쓸 필요가 없다.

다. 단일 노드(Single Node) : 노드(node) 이름

하나의 노드를 나타낼 때 사용한다. 노드 이름은 구문구조 내의 단일 노드의 이름이면 어느 것이든 상관없다. 구문범주의 이름(예: NP_SBJ)이어도 상관없고, 단 말 노드에 있는 어휘(예: 고향)이어도 상관없다. 다음의 사항에 주의한다.

- 하나의 노드 이름은 공백문자로 구분되어선 안 된다.
- *(와일드카드)를 써주면, 해당 노드는 어떤 문자열이든 상관없다는 의미를 나타낸다.
- “/”과 “/” 사이에 임의의 문자열을 써넣은 형식의 노드는 그 문자열을 포함한 모든 문자열을 만족한다.
예1) /^NP/ : NP_OBJ, NP_SBJ, NP 등을 모두 만족한다.
- “/.”과 “/” 사이에 임의의 문자열을 써넣은 형식의 노드는 그 문자열 앞에

하나이상의 임의의문자를 포함한 모든 문자열을 만족한다.

예1) /._MOD/ : NP_MOD, X_MOD, VP_MOD 등을 모두 만족한다.

- 특수문자(;:,.&|<>()[]\$!@%`?)를 써줄 때에는 반드시 큰따옴표(“”)로 묶는다.
- 절대 주의사항: 세종계획 구문분석 말뭉치를 사용하면서 한글어휘를 검색식에 사용할 경우, 말뭉치에 있는 형태소 형식을 입력해야 한다(원래의 어절은 불가능하다)

예1) “자라나/W + ㄴ/ETM”의 원어절인 “자라난”은 입력할 수 없다.

라. 노드 사이의 연결관계(link) 표시하기

구문구조를 구성하는 노드의 연결 관계는 특수 기호를 사용해서 표시한다. 기본적으로 “노드A 연결관계기호 노드B”의 형식을 가지고 지정하며, 모든 관계 기호는 노드A를 중심으로 이루어진다. 4.1.3 절에서 설명할 불린 표현식을 사용하면 복잡한 관계도 표현할 수 있다. 여기서는 두 노드 사이의 관계를 설명하며, 편의상 ‘노드A’는 A로, ‘노드B’는 B로 나타낸다. A와 B 사이의 관계를 나타내는 기호를 나열해 보았다.

- A < B A는 B를 직접 지배한다(바로 위에 있다)
- A <N B B는 A의 N번째 자식이다(N이 1이면 왼쪽 자식)
- A <-N B B는 A의 마지막에서 N번째 자식이다 N이 1이면 오른쪽자식).
- A <: B B는 A의 유일한 자식이다
- A << BA는 B를 지배한다(직/간접 모두 포함함. A는 B의 조상이다).
- A <<, B B는 A의 가장 왼쪽 후손이다.
- A <<' B B는 A의 가장 오른쪽 후손이다.
- A <<: B A의 후손들은 자매를 가지고 있지 않고, B는 그 후손중 하나다.
- A . B A는 B 바로 앞에 있다(A가 B를 바로 지배하거나 바로 왼쪽자매)
- A .. BA는 B의 앞에 있다(A가 B의 조상이거나 왼쪽 자매이다)
- A \$ BA는 B의 자매이다(그리고 A와 B는 다르다).
- A \$. BA는 B의 바로 왼쪽자매이다

마. 복합 노드(Complex Node) : 괄호(()) 사용

어떤 관계를 이루고 있는 두 개 이상의 노드를 묶어서 하나의 대상으로 나타낼 때에 사용한다. 즉 한 쌍의 괄호로 묶인 노드는 하나로 취급한다. (예: “S < (A < B)”는 B를 자식으로 가지고 있는 A를 자식으로 가진 S를 지칭한다)

바. 불린 표현식(boolean expression)을 사용한 복잡한 관계 표시

이 기능은 Tgrep에는 없는 기능으로서, ‘그리고’나 ‘또는’ 등을 뜻하는 불린(boolean) 연산자를 써서 노드 사이의 복잡한 관계를 표현하는 것이다. 여기서 표기하는 노드의 이름은 역시 편의상 A, B, C, D, E... 식으로 나타낸다. 불린 표현식 사용을 위해서는 다음의 사항을 명심한다.

- & 연산자는 ‘그리고(AND)’를 나타낸다. 즉 연산자의 양쪽에 있는 식이 모두 만족해야 한다는 뜻을 나타낸다. 연산자를 쓰지 않으면 기본적으로 AND로 취급된다

예1) A < B & . C : A는 B를 자식으로 가지고 C의 바로 앞에 있다

예2) A < B . C : 예1과 같은 뜻임.

- | 연산자는 ‘또는(OR)’을 나타낸다. 즉 연산자의 양쪽에 있는 식 중 하나를 만족하면 된다는 뜻을 나타낸다.

예1) A < B | . C : A는 B를 자식으로 가지거나 C의 바로 앞에 있다

- !는 부정(negation)을 나타낸다. 즉 주어진 관계를 부정하는 효과를 낸다.

예1) A ! < B : A는 B를 지배하고 있지 않다

- &는 |보다 먼저 처리된다.

예1) A < B . C | < D . E : (A는 B를 자식으로 가지고 C의 바로 앞에 있다) 또는 (A는 D를 자식으로 가지고 E의 바로 앞에 있다)

- 대괄호([])는 항(term)을 그룹화하여 알기 쉽게 하고 싶을 때에 사용한다.

예1) A < [B | . C] [< D | . E] : (A는 B를 자식으로 가지거나 C의 바로 앞에 있다) 그리고 (A는 D를 자식으로 가지거나 E의 바로 앞에 있다)

사. 예제 패턴들

① VP를 직접 지배하고 있는 NP_SBJ:

NP_SBJ < VP

- ② NP, NP_MOD, NP_SBJ 등의 이름을 가지고 있는 모든 노드:
/^NP/
- ③ 기능 태그가 PRD인 모든 노드:
/._PRD/1)
- ④ PP를 지배하지 않는 NP:
NP !<< PP
- ⑤ 위의 조건을 만족하는 NP이며 동시에 다른 NP를 직접 지배하거나 VP를 지배하는 것:
NP !<< PP [< NP | << VP]
- ⑥ NP를 지배하고 있으며 VP의 바로 앞에 있는 NP_MOD:
NP_MOD << NP . VP
- ⑦ “고향”을 포함한 모든 단말 노드 어휘를 지배하고 있는 NP_MOD:
NP_MOD << (* < /^고향/)
- ⑧ 품사 VX를 포함한 모든 단말 노드 어휘를 지배하는 노드이자 “VP”를 포함한 모든 것:
/^VP/ << (* < VX)
- ⑨ NP_SBJ를 왼쪽 자식으로, VP를 오른쪽 자식으로 가지는 직계 부모 노드 S:
S < (NP_SBJ \$. VP)
- ⑩ 바로 위의 조건을 만족하며 오른쪽 자식의 단말노드 어휘가 “중국”인 것
S < (NP_SBJ \$. (VP << (* < 중국)))

1) 세종계획 이진구구조말뭉치에서 기능태그는 ‘_XXX’ 형태를 가지고 구문태그 뒤에 붙기 때문에 ‘_’를 삽입해 준 것임.

이 사용자 설명서는 문화관광부에서 시행한 21세기 세종계획
프로그램 사용자 설명서입니다.